Politechnika Wrocławska

FIELD OF SCIENCE: Engineering and Technology

DISCIPLINE OF SCIENCE: Biomedical Engineeering

DOCTORAL DISSERTATION

Computational studies of bacterial functional amyloids and their human protein interactors

Alicja Wojciechowska

Supervisor/Supervisors:
Prof. Dr. Hab. Eng. Małgorzata Kotulska

Keywords: amyloid, proteins, interactions, neurodegeneration, microbiome

WROCŁAW 2025

Abstract

Neurodegenerative diseases, such as Parkinson's and Alzheimer's disease, affect 50 million people worldwide, reducing the quality of life of patients and their families. Many neurodegenerative diseases are associated with amyloid proteins, which can form stable and insoluble fibrillar structures and are the main focus of research on these pathologies. The ongoing transformation of the age structure in developed countries will triple the population of patients with neurodegeneration by 2050. Despite the high importance and prevalence of neurodegeneration, no effective treatment is available, raising the need for novel ways to study these pathologies.

Finding novel drug targets and uncovering disease mechanisms is possible with protein-protein interaction networks (PPINs). PPINs may be interpreted as disease roadmaps, whose detailed analysis can provide new high-level information on the pathology. This work discusses the application of available PPINs in neurodegeneration, showing that current PPIN datasets are biased by our scientific interests, which harness their biological interpretation. The published research regarding Parkinson's and Alzheimer's disease discusses mostly amyloid proteins and their interactions, hindering the generalisation of protein-protein interactions in these disorders on a proteome-wide scale. Therefore, to better understand neurodegeneration, studies on the understudied groups of proteins are needed.

Microbial amyloids, including bacterial functional amyloids, are a great example of understudied topics in neurodegenerative diseases with the potential to shed light on the onset and progression of neurodegeneration. Such proteins are purposefully produced by an organism, e.g. to serve as biofilm scaffolding. Previous studies have shown that bacterial functional amyloid proteins may be present in the human microbiome and affect the rates of amyloid deposition in the brains of patients with neurodegeneration. In this thesis, bacterial functional amyloids are analysed in detail. Examination of sequences of these proteins reveals that their aggregation propensity might be regulated by characteristic sequence repeats. Structural analysis of bacterial functional amyloids is not yet possible, as AlphaFold is generally shown to struggle with amyloid proteins. This is the result of the low abundance of amyloid structures in the AlphaFold training dataset, which leads to frequent prediction of high-quality globular models instead of fibrillar structures for amyloid proteins. The presence of bacterial functional amyloids in the human microbiome is analyzed to give grounds for discussion about their potential clinical importance. Through a designed pipeline, 805 such proteins, potentially produced by a broad spectrum of bacterial species, are identified in the microbiome proteome. Predictions of interactions between human proteins and functional bacterial amyloids suggest that bacterial functional amyloids could affect multiple molecular pathways, including inflammatory response, cell transport and signalling, and even harness the functioning of cell junctions responsible for intestinal permeability.

This thesis demonstrates that current research on neurodegeneration is biased by scientific interests in this topic. Studying different protein groups, such as bacterial functional amyloids, can shed new light on the pathology and extend our biological knowledge. The

quality and quantity of the experimental data are always a limit for the computational analysis. Hence, future research on neurodegeneration may greatly benefit from general high-throughput experiments that provide information on a proteome-wide scale, enabling big data analysis approaches to uncover biological patterns in these disorders. Until such high-throughput experiments on amyloid proteins become a very common practice, reaching beyond the most studied group of proteins can make an impact.

Streszczenie

Około 50 mln ludzi na całym świecie cierpi z powodu chorób neurodegeneracyjnych, takich jak choroba Parkinsona, czy Alzheimera. Oba schorzenia istotnie obniżają jakość życia pacjentów, czyniąc ich w pełni zależnymi od opiekunów. Wiele chorób neurodegeneracyjnych jest związanych z białkami amyloidowymi, które mogą tworzyć stabilne i nierozpuszczalne struktury fibrylarne. Takie amyloidy patologiczne są głównym przedmiotem badań nad chorobami neurodegeneracyjnymi. W wyniku zmian socjoekonomicznych, które powodują obserwowany wzrost populacji seniorów, do 2050 roku liczba pacjentów z neurodegeneracją ulegnie potrojeniu. Mimo tego, że choroby neurodegeneracyjne są powszechnie spotykane i uniemożliwiają pacjentom normalne funkcjonowanie, do dziś nie dysponujemy terapiami w pełni leczącymi te patologie. Dlatego nowe podejścia do badania tych schorzeń są wciąż bardzo potrzebne.

Identyfikacja nowych celów terapeutycznych i analiza mechanizmów molekularnych stojących za obserowowanym fentoypem choroby jest możliwa przy użyciu sieci interakcji białko-białko (PPIN). PPIN mogą być interpretowane jako molekularna mapa jednostki chorobowej, której szczegółowa analiza może dostarczyć nowych oraz bardziej ogólnych informacji o patologii. Niniejsza praca omawia zastosowanie dostępnych PPIN w neurodegeneracji, pokazując, że obecne zbiory danych PPIN są efektem zaintersowań naukowych, a nie jednorodną próbką wiedzy biologicznej, co powoduje, że interpretacja biologiczna PPIN jest znacząco utrudniona. Dostępne i opublikowane badania dotyczące choroby Parkinsona i choroby Alzhei-mera omawiają głównie wspomniane patologiczne amyloidy i ich interakcje, co utrudnia uogólnienie interakcji białko-białko w tych zaburzeniach na skalę całego proteomu. Dlatego, żeby lepiej zrozumieć choroby neurodegeneracyjne potrzebujemy wyjść poza zbiór popularnie badanych białek w tych patologiach.

Amyloidy mikrobiologiczne, w tym bakteryjne amyloidy funkcjonalne, są doskonałym przykładem niedostatecznie zbadanych tematów w chorobach neurodegeneracyjnych. Różne organizmy celowo produkują takie białka, np. do budowy biofilmu. Poprzednie badania wykazały, że bakteryjne amyloidy funkcjonalne mogą być obecne w ludzkim mikrobiomie i wpływać na tempo odkładania się złogów amyloidowych w mózgach pacjentów. W niniejszej pracy bateryjne amyloidy funkcjonalne są szczegółowo analizowane. Badanie sekwencji tych białek ujawnia, że ich skłonność do agregacji może być regulowana przez charakterystyczne powtórzenia sekwencyjne. Analiza strukturalna bakteryjnych amyloidów funkcjonalnych nie jest jeszcze możliwa ponieważ, jak zostało wykazane w tej pracy, AlphaFold ma ogólne trudności z białkami amyloidowymi. Wynika to z małej liczby rozwiązanych struktur amyloidowych w zbiorze treningowym AlphaFolda, która prowadzi do częstego przewidywania wysokiej jakości modeli globularnych zamiast oczekiwanych struktur fibrylarnych dla białek amyloidowych. Skala występowania bakteryjnych amyloidów funkcjonalnych w ludzkim mikrobiomie jest szacowana w celu omówienia ich potencjalnego znaczenia klinicznego. Zaprojektowany protokół identyfikacji bakteryjnych amyloidów funkjconalnych wykazał 805 takich białek w proteomie ludzkiego mikrobiomu, które są potencjalnie produkowane przez szerokie spektrum gatunków bakterii. Przewidywania interakcji między białkami ludzkimi, a bakteryjnymi amyloidami funkcjonalnymi

sugerują, że białka te mogą wpływać na wiele ścieżek molekularnych, w tym odpowiedź zapalną, transport i sygnalizację komórkową, a nawet wpływać na funkcjonowanie połączeń komórkowych odpowiedzialnych za przepuszczalność jelitową.

Niniejsza praca pokazuje, że dotychczas dostępne dane na temat neurodegeneracji są zaburzone zainteresowaniami naukowców w tej dziedzinie. Badanie innych grup białek, takich jak bakteryjne amyloidy funkcjonalne, może rzucić nowe światło na nasz sposób postrzegania tych schorzeń. Każda analiza obliczeniowa jest niestety ograniczona danymi eksperymentalnymi. Dlatego, wysokoprzepustowe eksperymenty na skalę proteomu, które w sposób jednorodny próbkują naszą wiedzę biologiczną są rozwiązaniem, które umożliwia dokładną analizę obliczeniową z potencjałem do odkrywania nowych wzorców rządzących neurodegeneracją. Zanim to się stanie bardzo powszechnie stosowaną praktyką, badania wykraczające poza najczęściej analizowane grupy białek, takie jak amyloidy patologiczne, mogą rozszerzyć naszą ogólną wiedzę na temat neurodegeneracji.

Acknowledgments

It is all about the journey, not the destination. And no journey is taken alone. I owe a very big thank you to many people. Starting from my parents, who patiently answered more questions than one can imagine and made me curious and brave to start this path. Going to my beloved husband, Jakub, who also listened to many questions, asked his own, shared with me all the good and bad moments and made me look on the bright side. You will always be my biggest discovery. I am very grateful to my supervisor, Prof. Małgorzata Kotulska, who supported me throughout all these years. Thank you very much for all the effort, trust, help, time and more than I can write in this section. It was you who showed the way, and I am happy I followed it. I would like to thank all my internship supervisors and mentors: Dr. Johannes Soeding, Dr. Isabel Marcelino, Dr. Tomasz Kościółek, and Prof. Gert Vriend. Because of you, I could learn different ways of doing science and looking at problems; these exciting adventures were priceless and largely enriched me as a person and scientist. I would like to thank all of my lab members and other collaborators. Working together was not only a great pleasure but also taught me that science is a "people thing" that requires collaboration and discussion.

Finally, I greatly acknowledge the OPUS grant from the National Science Centre titled 'Wyznaczenie wzorów interakcji krzyżowych między białkami amyloidowymi' (2019/35/B/NZ2/03997), which provided funding for this work, including doctoral scholarship.

Table of Contents

Introduction

Around 300,000 years ago, somewhere in a seemingly boundless, vivid savanna, history changed as we left the safe trees and started exploring the world beyond them. Despite the primitive life-threatening dangers that made us more prey than predator, surprisingly, we survived and prospered. All because of a risky bet of evolution, when everything was put on one card, intelligence.

It is undeniable that our brains distinguish us from other species. 100 billion neurons, more than stars in our galaxy, connected by 100 trillion connections, build the universe of our consciousness [1]. Thanks to this complexity, as no other animal on the planet, we can contemplate abstract concepts like 'science' or 'PhD dissertations', solve complicated puzzles left by nature, and finally use the idea of language to communicate about them. But the risky bet of evolution has a very high price that must be paid as we age.

More than 1 in 10 people aged 65 and above suffer from Alzheimer's disease [2], and 1 in 100 from Parkinson's disease [3]. The prevalence of both diseases increases dramatically in older populations. At the age of 85 years or older, 1 in 4 people live with neurodegeneration [4, 3, 2]. Although we share multiple characteristics with other primates, neurodegenerative disorders seem to be the plagues that reap the harvest only in our population, as no other animal has been found to have the full clinical phenotype of Alzheimer's or Parkinson's disease [5, 6]. This remarkable vulnerability of the human brain to neurodegeneration has been hypothesized to be related to subtle changes that increase our cognitive abilities but at the cost of a higher risk of brain-related disorders [7, 8]. One of our species-specific traits is a structural, functional, and vascular reorganization of the parietal lobe, a brain region responsible for abstract thinking. This relatively novel idea of evolution posed a significant increase in metabolic demands in this region, which could be difficult to meet as our cells age and experience declines in their metabolic performance [9]. As a result, a bottleneck appeared in the parietal lobe, the common starting point for Alzheimer's disease. A similar story can be told about Parkinson's disease. Humans are characterized by multiple connections in the cortical-subcortical regions that are highly demanding at the bioenergetic level but give us a cognitive advantage. As in the case of the hypothesis on the origin of Alzheimer's disease, the metabolic needs in this region are difficult to meet as we age and, consequently, the weak point in the structure of the brain emerges [10, 11]. To our detriment, there is no evolutionary pressure to eliminate these disorders as they appear in the post-reproductive age.

In some cases, neurodegeneration may be heritable and directly related to genetics. Approximately 5% of all patients with Alzheimer's disease suffer from a heritable Early Onset Familial Alzheimer's Disease. The occurrence of Early Onset Familial Alzheimer's Disease is directly related to pathogenic mutations in at least one of three important genes: PS1, PS2 or APP [12]. These patients experience symptoms of Alzheimer's disease earlier in life, and the course of the disease is often more aggressive. Parkinson's disease can also have a familial character, although it occurs only in around 10% of the cases. Until now, 20 different genes have been associated with familial Parkinson's disease [13]. The rest of the Parkinson's and Alzheimer's disease cases have an idiopathic character.

Today, it is estimated that 50 million people worldwide suffer from neurodegeneration, and the future is not filled with optimism. By 2050, this number will triple and neurodegenerative diseases will become the second most common disease at the time of death, overtaking cancer [14]. At first sight, this seems an inevitable consequence of the socioeconomic changes that result in longer life expectancy and an increase in the human population. However, this is only part of the story.

Cases of severe cognitive impairment were rare in ancient times, even though these disorders were already known [15]. Even more interestingly, lifestyle changes, which appeared between ancient Greece and the more industrialized Roman era, were correlated with an increase in the incidence of neurodegeneration, suggesting the role of environmental factors in the prevalence of these disorders. Studies of the Tsimane population in Bolivia, who live a primitive life with a high dose of physical activity, support these hypotheses. Among Tsimane people, mild cognitive impairment occurs but incomparably rarely to the epidemiological statistics observed in modern societies [15].

The contemporary lifestyle and environment seem to be important ingredients in the recipe for neurodegenerative plagues. We start our day breathing the 'fresh' city air, full of harmful particles that form air pollution. These particles can cross our blood-brain barrier, altering the fragile homeostasis of our brain, increasing inflammation and oxidative stress and finally resulting in the overall neurotoxic effect [16]. Next, we eat breakfast rich in simple sugars and saturated fatty acids, building a good foundation for metabolic syndrome that is another risk factor for neurodegeneration [17]. We go to work and sit for hours, although physical activity, even in low doses, decreases the chances of getting Alzheimer's and Parkinson's disease [18, 19]. The western diet and low physical activity increase the risk of cardiovascular diseases, which in turn increase the risk of Alzheimer's [20]. A history of strokes is more common in patients with Alzheimer's disease or Parkinson's disease than in healthy groups [21, 22]. Furthermore, genetic factors can interact with environmental ones, further modulating the risk, e.g. genetically determined high systolic blood pressure increases the risk of Alzheimer's disease [23]. Our lifestyle is often characterized by high levels of stress, typical of the modern fast-paced world. The pro-inflammatory properties of stress also affect our brains, and hence, stress-related disorders are risk factors for neurodegeneration [24, 25]. Sleep disorders, which even one in ten people experience [26], pose another heavy burden on our health. Patients suffering from them are more likely to also suffer from Parkinson's and Alzheimer's disease, and other types of dementia [27]. The full list of lifestyle factors influencing the risk of neurodegeneration is still not exhaustive, as our mental health, microbiome, smoking, alcohol consumption, pesticide exposures, and others also matter.

Behind the numbers raising our awareness of the prevalence and risk of Alzheimer's and Parkinson's diseases, patients experience the disease symptoms. Both diseases start imperceptibly. The first signs of Alzheimer's disease include a minor memory loss that progresses as days pass. The person loses initiative, can get lost in the areas they know, struggles to solve tasks that used to be easy, and can experience personality changes, e.g. more aggressive behavior. Once the disease reaches its full potential, people are incapable of communicating, have little or no awareness of their surroundings, suffer physical deterioration, and lose control over basic physiological activities [28]. Patients with Parkinson's disease share many of these symptoms. This disorder is also characterized by memory loss, mental health deterioration, and general cognitive decline. However, Parkinson's disease usually starts with tremors and shaking. The other motor symptoms include slowed movement, muscle stiffness, loss of balance, and struggle to perform automatic motions

[29].

At the molecular level, the hallmarks of neurodegeneration are amyloid plaques that are visible in the human brain scans of patients with Alzheimer's and Parkinson's diseases. Amyloids, by definition, are proteins or peptides that can aggregate to form stable fibrillar structures with a characteristic cross-beta pattern detectable by X-ray diffraction. Fibrils are made up of protofilaments that are formed of in-pairs associated beta sheets. The side chains of the mating beta sheets are tightly packed and form a so-called steric zipper. The core of the amyloid fibril is often hydrophobic; meanwhile, its surface is rich in polar residues [30]. The amyloid fibrils emerge via an aggregation mechanism that starts with the formation of a nucleus. As other monomers join the nucleus, it becomes an oligomer and then a fibril. Other mechanisms contributing to the aggregation may include, e.g. fibril fragmentation. The amyloid formation is characterized by a slow nucleation phase and rapid fibril elongation that ends in the stationary phase [31]. The aggregation is highly dependent on environmental conditions, including concentration or pH.

The amyloid deposits in the case of Alzheimer's disease consist of aggregated forms of the Aβ peptide and phosphorylated tau protein [32]. Aβ is the product of a proteolytic cleavage of the Amyloid-beta precursor protein (APP) - a cell-surface receptor. APP promotes synaptogenesis and is involved in cell mobility. APP may be cleaved in two modes: non-amyloidogenic (cleavage occurs within Aβ) and amyloidogenic ones (cleavage occurs on the N- and C-terminal of Aβ and results in secretion). Heritable mutations in APP and presenilins, which occur in Early Onset Familial Alzheimer's Disease, negatively alter the production of Aβ [33]. The microtubule-associated protein tau is a large neuronal protein that promotes microtubule assembly and stability. When phosphorylated, tau forms cytotoxic tangles built of amyloid fibrils [34]. In Parkinson's disease, the deposits, the so-called Lewy bodies, are made of α-synuclein, a neuronal protein associated with synaptic activity. Some mutations in α-synuclein are known to cause familial Parkinson's disease [35]. Tau and α-synuclein are two-sided proteins [32]. In their native state, they perform their functions that contribute to cellular well-being. However, under certain environmental conditions or mutations, they can change their form and create fibrils.

The current consensus in the field argues that the formation of amyloid fibrils present in Alzheimer's and Parkinson's disease is harmful to humans. Hence, Aβ, tau, α-synuclein and similar amyloid proteins associated with other diseases (e.g. amylin in type II diabetes), shall be called 'pathological amyloids'. Several factors support the hypotheses about the negative impact of pathological amyloids. In the first place, once pathological amyloids start to aggregate, they stop executing their primary functions in the cell due to loss of the intrinsic structure. Secondly, aggregates of pathological amyloids, and especially oligomers that precede the emergence of fibrils, can disrupt cellular membranes [36, 37]. This property is one of the most fundamental aspects that support the view on the cytotoxicity of pathological amyloids. Pathological amyloid formation is also associated with the immune response and induction of oxidative stress. The Aβ and α-synuclein oligomers can activate certain cell surface receptors, such as toll-like receptors, which in turn release pro-inflammatory cytokines and chemokines that promote neuroinflammation and cell damage [38, 39, 40]. Finally, pathological amyloids can negatively affect endothelial cells, consequently leading to blood-brain barrier permeability [41, 42, 43, 44].

We are tempted to see neurodegeneration through amyloid lenses in light of these observations. One of the common views places the "amyloid seeding" at the centre of these disorders. In this mechanism, once one protein misfolds and recruits others to form a seeding oligomer, the structural template propagates and causes the misfolding cascade. This

process lies at the foundation of prion diseases, like Creutzfeldt-Jakob disease (CJD). In CJD, misfolded proteins act as 'infectious agents' - they induce normal proteins to misfold. One can contract CJD, e.g. through the consumption of a sick animal with misfolded proteins. In the prion-like hypothesis of neurodegeneration, amyloid propagation occurs between cells, leading to the expansion of misfolding to further brain areas [45]. Several studies support this hypothesis. Amyloid fibrils, e.g. of tau or α-synuclein can appear once an amyloid seed is introduced to the cell [46, 47, 48] . Certain amyloid pathologies, like Aβ plaque pathology, were observed to be even transmissible between humans via neurological surgery [47]. Even more convincing is the fact that the spatiotemporal distribution of amyloid plaques in neurodegeneration is related to spatial connectivity (in the case of α-synuclein [49]) or proximity (in the case of Aβ [50]) in the brain.

Pathological amyloids alone cannot tell the whole story of neurodegeneration. The recent trials of anti-amyloid drugs in Alzheimer's disease provided controversial results that did not meet the great hopes associated with them. Although they remove the aggregates of Aβ, patients do not experience the expected significant cognitive improvement [51, 52]. It can be argued that the drug was administered too late and did not target the most toxic amyloid oligomers that appear first; hence, it simply did not work spectacularly. On the other hand, tangles of tau proteins appear to be more correlated with the disease symptoms than Aβ aggregates, thus, drugs targeting tau proteins could be more effective. Similarly, antibodies designed to clear aggregates of α-synuclein in Parkinson's disease do not seem promising. Although clinical trials are ongoing, so far, no effect on disease progression has been demonstrated [53]. Current treatment of Parkinson's disease only attempts to alleviate some symptoms, e.g., by administering drugs that mimic dopamine. Unfortunately, with time, the organism's response to therapy decreases significantly [54, 55]. The struggles of the mentioned clinical trials clearly illustrate that our understanding of neurodegeneration is far from complete. This marks the new line for scientific exploration.

Recent developments in a relatively new field of network science can shed new light on how we conceptualize human diseases. Network science, in its essence, aims to capture complex relations occurring between different entities. This framework can easily be applied to modelling a neurodegenerative disorder. We can identify all proteins known to be involved in a pathological process, such as Parkinson's disease, and the interactions that exist between them. In such a way, we build a protein-protein interaction network - a map for the molecular process of choice. Then, research is limited only by our imagination. We can attempt to identify the most important proteins in the network (potential drug targets), model the spread of the information (e.g., cell signalling), or examine if a network disintegrates after the removal of certain proteins (effect of the protein inhibitors). The network approach proved useful in gene prioritization in Parkinson's and Alzheimer's diseases, revealing new drug targets [56, 57, 58, 59], supported the construction of a dataset of gene risk factors dataset in both diseases [60], highlighted the molecular pathways disturbed in these disorders [61, 62], helped to identify proteins responsible for neurotoxicity in neurodegeneration [63], and provided us with a map of interactions between amyloid proteins [64, 65], to name a few applications. Despite these successes, we must keep in mind that understanding molecular crosstalk in neurodegeneration requires reliable maps of the disease.

How authentic are available data regarding protein-protein interactions, and can they provide faithful information regarding neurodegeneration? These questions are a starting point for this thesis. Before heading out into the wide waters of the Ph.D. journey, we

discuss the utility of the relatively novel and promising network approach. The topological structures of various protein-protein interaction networks representing human disorders are assessed. A set of measures that can help a researcher in diagnosing the network utility is provided. Finally, an online tool for the diagnosis of the topology is available to everyone. The first chapter demonstrates that the protein-protein interaction data, especially in the context of neurodegeneration, are not always reliable and require individual examination. Otherwise, false conclusions about the disease mechanism can be made. Interests of the scientific community frequently result in data biases that hamper proper in-silico modelling of human disorders. This, in turn, makes protein-protein interaction networks a method with limited applicability. The analysis of human protein-protein interaction networks also has one more drawback - it ignores 20 million nonhuman proteins that are present in our gut.

The human microbiome is often called a vital meta-organ. It produces substances of critical importance for our health, such as folate, riboflavin (vitamin B12), or short-chain fatty acids [66, 67], regulates our immune response by influencing the production of cytokines and the activation of lymphocytes [68, 69], and participates in maintaining metabolic homeostasis [70]. Although our microbes constitute only 2-3% of our mass, they dictate our well-being. The microbiome changes throughout our lifetime. We are born with a restricted community that rapidly expands and plays a critical role in shaping our immune system in those early years. As years pass and we enter the fall of our lives, the microbiome transforms to support us in healthy ageing. Some genera become less abundant in our intestine, such as Prevotella or Bifidobacterium; meanwhile, colonies of Akkermansia or Butyricimonas grow [71]. In unhealthy ageing, changes also occur, but are quite distinct.

Multiple studies revealed a vast array of microbiome-related factors associated with the presence and/or progression of neurodegeneration. Different microbiome composition, increased lipopolysaccharide secretion, decreased presence of bacteria producing short-chain fatty acids, increased presence of pro-inflammatory bacteria, such as Proteobacteria, and increased gut permeability linked to inflammation were observed in diseased patients [72, 73, 74]. Despite these efforts to understand microbiome alterations in neurodegeneration, multiple issues remain unsolved. How do changes in the microbiome affect the gut-brain axis in these disorders, and what are the detailed molecular mechanisms of this crosstalk?

It could be that metabolites, frequently capable of crossing the blood-brain barrier, are crucial in this interplay. Short-chain fatty acids received particular attention in this context. Parkinson's disease patients have lower levels of butyrate, and mice with induced Alzheimer's disease, when given butyrate, experience cognitive improvement and a decrease in amyloid plaques in the brains [75]. Other studies point to lipopolysaccharides that are potent microbiome-derived neurotoxins capable of activating multiple immune-related pathways. Lipopolysaccharides are abundant in the brains of Alzheimer's patients, with a prominent presence in the most impaired regions [76]. Another hypothesis focuses on the set of interesting proteins, so-called bacterial functional amyloids, suggesting that these microbial entities could be triggers for neurodegeneration.

Bacterial functional amyloids belong to a wider group of so-called functional amyloids discovered more than 20 years ago. Functional amyloids share crucial structural characteristics with previously mentioned pathological amyloids: they form regular, stable, and insoluble fibrils with a cross-beta structure via the aggregation process. However, they fundamentally differ in their cellular role. Functional amyloids, as the name suggests,

are purposely produced by an organism to perform certain functions. Such proteins exist across different branches of life, including humans, other mammals, insects, fungi, bacteria, and even viruses. A prominent example of functional amyloid in humans is Pmel17. Its fibril formation has an impact on the condensation of melanin pigment [77]. Some hormones also take advantage of the stable amyloid form, as it provides the most dense packing possible, and can serve as a material for hormone storage [78]. The protective character of amyloid fibrils, arising from their stability and insolubility, also makes them a great material for eggshells. Silkmoth oocytes are protected by a shield built of chorion protein amyloid fibrils [79]. In fungi, amyloids are used in signalling processes that control non-self-recognition [80]. Viral amyloids are widespread and can regulate viral gene expression, inhibit necroptosis and participate in virion budding [81]. In bacteria, functional amyloids are frequently involved in biofilm formation [78]. The stable amyloid structures formed outside of bacterial cells provide a scaffold for a biofilm matrix that protects the bacterial community from environmental dangers. The most well-known example of such proteins is curli (CsgA, produced, e.g. by E. coli), which forms degradation-resistant fibrils on the cell surface [82]. Although beneficial for bacteria, CsgA-mediated biofilm formation can be an important virulence factor that increases antibiotic resistance [83, 84]. Other biofilm-related amyloids include Bap and Esp, which are long proteins from Staphylococcus aureus and Enterococcus faecalis, respectively, FapC protein from Pseudomonas or TasA from Bacillus subtilis [85, 86, 87, 88]. The formation and stabilization of the biofilm is not the only possible function of bacterial functional amyloids. They can also be involved in the binding of DNA and RNA. For example, the Hfq protein, found in around half of the bacterial species, has multiple functions related to its amyloid formation, including the pairing of sRNAs with mRNAs, regulating mRNA stability and DNA compaction [89, 90].

Bacterial functional amyloids are produced by bacteria that inhabit our intestinal tract, and their prominent structural similarity to pathological amyloids inspired an interesting hypothesis about the emergence of Parkinson's disease. In the first place, we must view Parkinson's disease in light of Braak's hypothesis, which postulates, based on observation of the pathology expansion pattern, that this disorder starts in the enteric nervous system and then propagates to the brain via the vagus nerve [91]. The aggregates of α-synuclein in the enteric nervous system were shown to appear in the early stages of the disease and are correlated with its severity, supporting the view that the intestine is the starting point [92, 93]. These considerations may now go further. If Parkinson's disease launches outside of the central nervous system, what could be the trigger in the intestine?

There is notable evidence supporting the hypothesis that bacterial functional amyloids could directly or indirectly influence the onset and progression of Parkinson's disease [94, 95]. In the first place, it was widely demonstrated that the presence of bacterial functional amyloids can affect the aggregation process of pathological amyloids due to mechanisms of molecular mimicry [65]. The so-called cross-seeding implies that the aggregation of one protein is catalyzed by the addition of preformed fibrils of the same protein or another. Cross-seeding is specific, to some extent, as not all proteins can impact the aggregation of another [96]. Curli protein, although different in sequence from α-synuclein, can accelerate its aggregation in biochemical assays. Furthermore, colonization of the intestinal tract of mice with CsgA-producing bacteria promotes the pathology of α-synuclein in the intestine and brain, and its inhibition decreases aggregation of α-synuclein in the brain [97]. CsgA also interacts with Aβ and promotes Alzheimer's pathology in C. elegans models [98].

FapC has similar properties and can accelerate the aggregation of both α-synuclein and Aβ [99, 100]. Very recent work revealed that biofilm-related bacterial functional amyloids (Bap) are more abundant in patients with Parkinson's disease [101]. Importantly, the same study showed that Bap proteins can co-localize with α-synuclein in neurons and increase its aggregation.

Bacterial functional amyloids could have a broader effect on neurodegeneration than only affecting the rates of aggregation of pathological amyloids. Patients suffering from Parkinson's disease and Alzheimer's disease often have a dysbiotic gut that experiences chronic inflammation [73, 102]. Pathological amyloids, as mentioned previously, have cytotoxic effects and affect multiple immune-related pathways. Reasoning based on molecular mimicry leads to the hypothesis that bacterial functional amyloids, similarly to pathological ones, could have pro-inflammatory properties, especially during dysbiosis, activating the same immune-related pathways and hence worsening the disease [103, 104]. Phenol-soluble modulins (PSM), microbial amyloids secreted by Staphylococcus, activate the formyl-peptide receptor 2, which leads to the attraction of neutrophils. Interestingly, the secretion of these bacterial functional amyloids was even correlated with bacterial pathogenicity, suggesting that differentiation between pathogenic and nonpathogenic Staphylococcus could be related to PSM identification by the immune system [105]. Furthermore, both microbial and human amyloids have been shown to stimulate a range of toll-like receptors that are part of human innate immunity [106]. Curli fibrils can activate the NLRP3 inflammasome pathway that plays an important role in inflammatory signaling [107].

These observations highlight the potential importance of bacterial functional amyloids in neurodegeneration and lead us to the next topic considered in this thesis. When this work started, no full dataset of all known bacterial functional amyloids was available. Hence, the second chapter starts by answering this need via literature mining. Then, I move to sequence analysis of a created dataset, in order to better understand the role of bacterial functional amyloids in neurodegeneration, we must get to know these proteins better.

Current methods for computational studies of amyloids are often focused on identification of aggregation-prone regions that are drivers of nucleation. In normally folded proteins, aggregation-prone regions are frequently buried inside the protein, but when exposed to solvent, they can trigger the amyloid formation process [108]. Many algorithms predicting aggregation-prone regions search for characteristic motifs in the sequence. A classic example is Waltz [109]. The Waltz authors experimentally screened a wide array of hexapeptides to identify the amyloidogenic ones. Then, a tool was built based on a calculated position-specific scoring matrix. Aggrescan has a similar philosophy [110]. Firstly, a range of mutational experiments on Aβ were performed. Then, they developed a method that screens the sequence and scores its fragments for their amyloidogenicity. To account for the fact that aggregation-prone regions can be buried and never activated in the globular protein, Zambrano and colleagues developed a newer version of the algorithm - Aggrescan 3D [111]. It combines information about the accessibility of the residue to the solvent and its aggregation propensity. Archcandy, instead of identifying the aggregation-prone regions, searches for potential information about formation of beta-arches in the sequence, as this motif can be found in the majority of disease-related amyloids [112]. Many recent software take advantage of machine learning. Amylogram combines n-gram sequence analysis with a random forest model, [113]. PATH uses template-based modelling to derive energy terms that are used as an input to a machine learning model, [114].

FishAmyloid relies on correlated occurrences of sequence elements and machine learning classification [115]. AggBERT uses embeddings from a large language model adjusted to proteins, ProtBERT, to predict aggregation-driving hexapeptides [116].

Many of the presented methods, and similar ones, are based on sliding windows that screen the sequence for aggregation-prone regions. The false positive ratio is often low, only 5-10% per test. However, we should keep in mind that, in the case of longer proteins, the tool will perform many tests and almost always identify some false aggregation-prone regions in the sequence. Hence, even if the sequence is entirely not amyloidogenic but long enough, by statistics, the approach with the sliding window will almost always yield at least one false positive result. For example, if a sequence is 100 amino acids long, and the sliding window is of size 6, we have 94 tests. If 5-10% of the tests are incorrect, around 4-9 false aggregation-prone regions can be found on average. This is, of course, a rough simplification, as the tests are correlated, but it still gives an idea of possible problems with the application of these methods to the large-scale identification of novel amyloids.

Unfortunately, few predictors perform a single test to examine the aggregation propensity of the entire protein sequence. One of them is ECAmyloid. It extracts sequence features such as evolutionary and secondary structure information, amino acid composition, solvent accessibility, and others to train an ensemble learning classifier [117]. Another is AMYPred-FRL, which uses a similar approach. It extracts sequence features and performs feature representation learning to produce a meta-predictor [118].

The vast majority of these tools are trained on the sequences derived from pathological amyloids, as functional amyloids are vastly underrepresented in amyloid databases. Functional amyloids differ in amino acid composition from pathological ones [119]. Therefore, the usage of these tools should be treated with caution in the case of bacterial functional amyloids. This was confirmed by a recent study by Szulc et al. Correctly predicting all amyloid fragments in curli (CsgA) was impossible with the common amyloid predictors [82]. Based on this, it can be concluded that to analyse how bacterial functional amyloids aggregate, other methods than amyloid predictors are welcome.

The protein alphabet is formed of twenty amino acids, which are protein building blocks. With this alphabet, words and entire sentences, which relate to sequence motifs and domains, are built. This analogy to natural language gives room for the application of complex language models to study the language of proteins. One of the important families of algorithms in this area is protein embeddings that aim to encode the sequence in the form of a fixed-size vector. Although the idea of 'protein vectorization' is not new, as a protein-describing vector can be built, e.g. by calculating physicochemical properties, modern natural language models bring it to the next level. By training the language models on the protein sequence data sets, one can provide a descriptive representation of a protein that provides more information than the mere protein sequence [120]. From a technical point of view, two paradigms are used to train a protein language model: transfer learning and self-supervision. Transfer learning accounts for the applicability of successful neural network architectures, developed, e.g. for natural language. With self-supervision, one can use the vast amount of unlabeled data to train the model. This latest idea of protein language models makes them particularly useful, as unannotated proteins make up a significant portion of the reference databases.

Protein embeddings proved to be useful in a variety of tasks. The shallow word2vec embeddings for the proteins were highly precise in classifying antibacterial peptides [121]. SeqVec embeddings, based on the ELMo language model, captured the information on secondary structure, disordered regions, and subcellular localization, succeeding models

based on biophysical properties and evolutionary information, and at the same time providing fast calculations [122]. Similarly, ProtBERT embeddings performed at the level of state-of-the-art in tasks related to structure, post-translational modifications, and function prediction [123]. Embeddings from the ESM2 model were successfully applied to estimate per-residue sequence conservation [124]. Finally, protocols using protein embeddings to visualize protein datasets were developed [125].

In the second chapter, protein embeddings are used to better understand the protein space of the bacterial functional amyloids. As these proteins are poorly studied, this unsupervised approach seems particularly valuable. Hence, the embeddings for bacterial functional amyloids are predicted and analyzed to observe if these proteins form a homogeneous cluster that separates from the entire protein space. After this initial step, which provides the map of the bacterial functional amyloids and relations between them and other proteins, we discuss the results from a more classical sequence analysis approach that is based on the evolutionary perspective.

Evolution purposefully designed functional amyloids. As an example, we can consider the Csg and Fap amyloid systems which take part in biofilm formation and consist of various genes that regulate transcription and aggregation of CsgA and FapC proteins. In the case of the Csg system, two operons are present: CsgBAC and CsgDEFG. The first encodes fibrillar CsgA (major curli subunit) and CsgB (minor curli subunit), as well as the chaperone CsgC, while the second encodes accessory proteins responsible for fibril translation, transcription and secretion [126]. Fap system is constructed similarly. FapC is the major component of the fibril, and FapB is the minor one. FapA is a chaperone that regulates fibril assembly and morphology [127]. Fap and Csg amyloid systems have been shown to evolve to have amyloid agents and phylogenetically conserved mechanisms that regulate their rates of aggregation [128, 129]. Similar observations were made for yeast adhesins with amyloid properties - their aggregation-prone regions turned out to be conserved in this protein family, [130]. In addition, the formation of A-body functional amyloids, which is related to the physiological response to environmental stressors, was conserved between different species, including fish, human, and chicken [131].

Functional amyloids, although similar to their pathological counterparts in many ways, have their own characteristics. Most importantly, their fibrils have a wider range of stability and lifespan than pathological ones, even with the potential to disassemble under certain conditions once the cell needs it, [119, 132, 131]. This fascinating flexibility requires that aggregation control mechanisms are incorporated into the sequence of these proteins [133].

It seems reasonable to think that tandem repeats, sequence regions with a repeating pattern of amino acids, could regulate protein aggregation propensity in the case of bacterial functional amyloids. This view is supported by statistics that reveal the wide presence of tandem repeats across different proteomes and their supporting role in the binding properties of their proteins. Repetitions are under pressure of the evolution that carves them to perfectly execute their well-designed roles in the cell [134]. The presence of repeats introduces multiple symmetric interactions that could contribute to the precise steering of amyloid fibril formation in the case of these proteins. The CsgA and FapC proteins, which are part of phylogenetically conserved amyloid systems, contain repetitive sequence motifs that relate to the formation of beta-sheets in amyloid fibrils. The presence of repetitive motifs is conserved, although the parameters of tandem repeats differ between bacterial species [128, 129]. Long biofilm-related amyloids also have repeats in their sequences, but their role is not yet fully understood.

The second chapter also discusses the role of tandem repeats in bacterial functional amyloids. It is hypothesised that the similarity between repeat units, their number and their lengths could be important factors regulating the aggregation process in the case of bacterial functional amyloids. To evaluate this, the number of bacterial functional amyloids with repeated regions is examined. It is also demonstrated that repeats in bacterial functional amyloids consist of a small number of units that are only a rough repetition of each other. Finally, the results are compared with those of other bacterial proteins, revealing the particular aggregation-regulating character of the repeats in functional bacterial amyloids. Nevertheless, sequence analysis is just a first step that slightly removes the veil of secrecy of these proteins. Structure, which is so closely related to protein function, could tell more.

Unfortunately, few structures of bacterial functional amyloids are experimentally resolved. The scarce examples include phenol-soluble modulin and the CsgA protein. The full fibril structure of these proteins were studied with electron microscopy [135, 136]. Structures for pathological amyloids are also not widely abundant, especially in the case of longer sequences. This low abundance of the amyloid structural data is a consequence of experimental limitations, such as the sensitivity of amyloid fibrils to environmental conditions and their high molecular weight, which make experiments difficult and expensive. Little data on amyloid structures and issues associated with solving them experimentally make computational approaches to structure prediction highly welcome.

Few attempts to predict amyloid structure have been made. Fibrepredictor was released in 2016. It is similar to template-based modelling. It estimates which of the known fibrillar architectures is the most suitable for a provided sequence [137]. Another algorithm is BetaSerpentine [138], which tries to predict how beta-arches are placed in a structure of an amyliod fibril. A very recent tool from 2025 is RibbonFold [139]. It is a neural network with built-in constraints that allow for the prediction of amyloid structures. Although all these solutions are very promising and remarkable, they do not allow for a generalized modelling of amyloid proteins. That means that modelling different sizes of the fibrils, predicting structures for monomers and multimers, adding ions, which may affect amyloid fibrils and discovering unknown architectures of amyloid proteins is not possible with them.

Recent breakthroughs in protein structure prediction methods are highly promising in this regard. The current state-of-the-art model, AlphaFold, not only opened but also broke the door to accurate computational modeling of proteins, reaching very high accuracy in the CASP (Critical Assessment of Structure Prediction) competition [140]. AlphaFold is a neural network trained on a large amount of structural data available in the Protein Data Bank (PDB). AlphaFold proved useful in a variety of tasks. It provided predictions for millions of proteins, covering 98.5% of the known proteome, that made up the AlphaFold Database [141]. It can aid drug discovery [142, 143]. Structural information produced by AlphaFold boosts reliable protein annotation [144]. Different variants of the AlphaFold were used to predict multiple protein-protein interactions, for example, to discover the interactomes of whole organisms [145]. These successful applications of AlphaFold and its wide usage by the community naturally raise the question of it can deal with bacterial functional amyloids.

In the third chapter, we investigate how the latest version of AlphaFold, AlphaFold 3, deals with amyloid structure. The performance of AlphaFold on the cases of well-studied amyloids, which were part of the training, is examined based on multiple monomeric and multimeric predictions with different parameters. Models of amyloid proteins with an un-

solved experimental structure of a fibril are also evaluated. The discovered trends are then compared to the general AlphaFold predictions available for the human proteome. The detailed analysis reveals that AlphaFold, regardless of its version, generally struggles with amyloid proteins, as these constitute the minority of its training dataset. In consequence, it prefers to produce high-quality globular models for amyloid proteins instead of the fibrillar ones. The latest occur occasionally, but are more common for shorter sequences and are often associated with poor quality scores. This result underscores that, to better understand amyloid proteins, including bacterial functional amyloids, we need more experimental data on other amyloid proteins than on the handful of pathological amyloids most commonly studied. Otherwise, the powerful machine learning approaches will not work as we expect them to.

The variety of analyses produced for bacterial functional amyloids expands our understanding of them. However, until Chapter 5, I do not consider these proteins from the perspective of the gut-brain axis, which is particularly important for medical applications. The structural similarity of bacterial functional amyloids to pathological ones gives room for speculation about their impact on the onset and progression of neurodegeneration.

In the final part of this thesis, we merge the gathered knowledge on bacterial functional amyloids and protein-protein interactions to investigate this topic. An atlas of predicted bacterial functional amyloids in the human gut proteome is provided to estimate to what extent such proteins are present. For this goal, a computational pipeline is developed. The screening method is based on the homology search and two amyloid predictors, which together aim to minimise false positive results when searching the large dataset of human gut proteins. The predicted dataset of bacterial functional amyloids in the human gut microbiome is analyzed from multiple perspectives, including the taxonomic origin of identified bacterial functional amyloids, their cellular localization and interactions with human proteins. It is shown that the bacterial functional amyloids in the human microbiome may be expressed by a wide array of bacterial species. Many of the potential bacterial functional amyloids seem to be extracellular or membrane proteins capable of interacting with human proteins. The clinical significance of bacterial functional amyloids is discussed. The metagenomic analysis of the abundance of bacterial functional amyloids in stool samples from patients with Parkinson's and Alzheimer's disease and their respective controls was performed, revealing that the amount of intestinal amyloids may be different during neurodegeneration. The predicted protein-protein interactions between human proteins and bacterial functional amyloids from the human gut microbiome point toward a couple of important molecular pathways that these proteins may affect. Specifically, many of the human proteins related to e.g. chemokine and cytokine signalling, leukocyte migration, and cell junctions, which are responsible for gut impermeability, may be affected by bacterial functional amyloids. It seems that bacterial functional amyloids could share some parts of the human interactome with pathological amyloids, triggering the same molecular pathways. Based on this analysis and data from the available literature, we discuss a conceptual framework for the possible role of bacterial functional amyloids in neurodegeneration. This work is the first large-scale analysis of intestinal amyloidogenic proteins and their relation to human proteins. Chapter 5 closes the full story about bacterial functional amyloids, giving a solid basis for the discussion of their role in neurodegeneration.

The upcoming chapters reveal that our scientific interests lead to a nonuniform sampling of biological data. This means that our level of understanding of different biological phenomena varies and depends on how popular a certain topic is in the scientific community. The resulting knowledge gaps and interest biases may be detected by studying

protein-protein interactions. Once found, to fill them, we may study proteins that are rarely the main focus of researchers, such as bacterial functional amyloids. To better understand this understudied group of proteins, a dataset of bacterial functional amyloids must be prepared. Analysis of their sequences and structures may give new insight into the aggregation and interaction mechanisms of functional amyloids. The clinical importance of this protein group may be assessed by studying their presence in the human microbiome and predicting how they interact with human proteins. In the following chapters, all these topics are discussed in detail, and all the results and methodological pipelines supporting this work are shown.

Goals and hypotheses of this work:

Hypothesis 1: Exploratory topological analysis of available protein-protein interaction datasets can provide a new perspective on neurodegeneration.

Goal 1: To assess the quality of the real-world protein interaction networks through topological analysis and identify if topological analysis can shed new light on disease mechanisms.

Hypothesis 2: Sequence tandem repeats influence the aggregation of bacterial functional amyloids.

Goal 2: To examine the role of tandem repeats in bacterial functional amyloids.

Hypothesis 3: Structures of amyloid fibrils can be predicted with AlphaFold 3.

Goal 3: To investigate AlphaFold 3 performance on amyloid proteins.

Hypothesis 4: Bacterial functional amyloids in the human intestine may influence neurodegeneration.

Goal 4: To computationally identify bacterial functional amyloids in the human gut proteome and their potential interactions with human proteins.

Current protein-protein interaction data cannot fully describe neurodegeneration

2.1 Introduction

All multiomics studies that drive the modern progress of life sciences require advanced data analysis techniques. One of the methods that leads to a high data load is protein-protein interaction experiments. Protein-protein interactions are highly specific, following the idea of a lock and key. Their large-scale discovery provides an overview of the different disease mechanisms that occur at the molecular level. This is particularly useful in the case of proteinopathies, such as neurodegenerative diseases. Research in Alzheimer's and Parkinson's diseases often focuses on the interactions between pathological amyloids and other proteins, revealing altered molecular pathways and drug targets [146].

Large-scale analysis of protein-protein interactions results in protein-protein interaction networks (PPIN). PPINs are mathematical objects consisting of nodes, proteins, and links between them that are experimentally confirmed interactions. PPINs are prominent examples of complex systems that aim to represent molecular mechanisms, e.g. during the disease. They can provide us with a new perspective on the complicated nature of neurodegeneration due to their intrinsic goal of capturing complexity. At the same time, they require a sophisticated methodology for their analysis.

The investigation of PPINs is based on mathematical workflows for graph studies with a frequent focus on network topology and dynamics [147]. Real networks often demonstrate certain structural features. They are 'small worlds', which means that any two nodes are relatively close to each other. Consequently, they contain clustered regions where nodes have high clustering coefficient values. Real networks are also 'scale-free' - the majority of the nodes have a low degree, and nodes with a high degree (hubs) appear rarely. This property results in a heavy-tailed degree distribution, which can be described by the power law distribution: $p_k \sim k^{-\gamma}$ [148]. The power law parameter γ for PPINs is expected to follow: $2 < \gamma < 3$ [148]. The impact of high-degree nodes may also be defined by examining the number of links they bring. PPINs have a disassortative structure, which means that proteins with a high number of interactors interact mostly with proteins with a low number of interactors [149, 150]. The assortativity may be evaluated with the degree correlation coefficient and Average Nearest Neighbour Degree (ANND) plots. The described structure of the PPINs implies that they are robust to the random failure of one protein in the network but vulnerable to the targeted attack when the most important proteins, for example hubs, fail. Such nodes may also be fundamental in the case of failure cascade simulation when the size of the error propagation in the network is measured. These topological characteristics lead to the important question: Which proteins are the

most important for the network?

The centrality metrics attempt to answer this issue. The most basic one is the *degree centrality*, which is the number of interactors per protein. *Betweenness centrality* captures nodes that lie between different clusters and serve as bridges between parts of the network. *Eigenvector centrality* defines the node's importance as a function of its neighbours' importance. Finally, *closeness centrality* describes the node's role in the network by checking its distance from other nodes. The application of these metrics, in the perfect scenario, can lead to the discovery of novel drug targets.

PPINs data suffer from several problems that may limit their applicability. PPINs, in a standard approach, are represented as static systems, while proteins form a dynamic one, with temporal interactions. More importantly, the experimental data resources, which are the building blocks for PPINs, are highly heterogeneous because different experimental methods are used. Unfortunately, all the experimental protocols used to study protein-protein interactions are error-prone and result in some false positives [151, 152]. Consequently, the data structure of the PPINs is not only heterogeneous, but also contains artifacts. Finally, the choice of the proteins used to build PPINs is also important and may depend on the individual research interests, leading to biased structures of PPINs.

Examination of the protein-protein interaction data can give a bird's-eye perspective on our knowledge about neurodegeneration and other diseases, uncovering not only molecular mechanisms but also artifacts and unexplored ideas. In this chapter, available human interactomes, including the Parkinson's disease one, are studied to better understand the structure of these systems. It is speculated that different data-gathering procedures and scientific interests pose doubts about the validity and reliability of the available interactomes, calling for research in less-studied directions. To make the study more applicable, also in other contexts, a mathematical evaluation of how different topological characteristics may guide the researcher in the diagnosis of their network biases and their implications is provided. To further support this, an easy-to-use Colab notebook with applied methods is made available.

The first hypothesis that ***available protein-protein interaction data differ in quality, which can affect their applicability*** is discussed. The quality of the real-world protein interaction networks is assessed by studying their topological characteristics. As a result of this analysis, knowledge gaps in neurodegeneration research are detected.

Some preliminary results for this chapter were part of my Master Thesis. Full development of this project took part during my PhD research.

2.2 Methods

PPIN definition

PPIN is a protein-protein interaction network $G(V,E)$. V represents the set of nodes (proteins) and E the set of links between them (interactions). The size of the V set is denoted as N and the size of the E set as L.

Data retrieval

Three PPINs were recovered from the IntAct website https://www.ebi.ac.uk/intact/. First, *Parkinson*, regarded protein-protein interactions related to Parkinson's disease. It

was downloaded twice: 4 November 2020 and 8 November 2021. Second, *Cancer* regarded protein-protein interactions related to cancer. It was also downloaded twice on 1 February 2021 (*Cancer I*) and on 8 November 2021 (*Cancer II*). Third, *HuRI*, Human Reference Interactome, covered the human protein interactome.

Parkinson and *Cancer* datasets are available in the section *Download/Curated datasets* on the IntAct website. *HuRI* can be found using the identifier IM-25472 in IntAct search. Topological analysis was performed on the downloaded datasets after the removal of parallel links and the extraction of the largest component.

Degree

The degree of a node k is the number of its neighbours, equal to the number of protein interactors. The high-degree nodes are termed *hubs*.

Powerlaw fitting

In scale-free networks, the degree distribution follows the powerlaw distribution $p_k \sim k^{-\gamma}$. In practice, it is often impossible to fit the powerlaw to the entire sequence of degree values. Hence, the cutoff k_{min}, from which the fitting is possible, can be used. Then, only the nodes with a degree greater than k_{min} are considered for fitting. If k_{min} it is not high and the powerlaw fit is successful, the network is still considered scale-free.

To fit the powerlaw distribution to the degree distribution, a maximum likelihood (ML) estimator implemented in the poweRlaw package was used [153]. For different numbers k_{min}, the best γ value was estimated. The hypothesis H_0: *degree distribution does follow the powerlaw* was not rejected when the p-value of the fit was greater than 0.1. With this approach, it was possible to balance between the best fit possible with the lowest k_{min}. The detailed algorithm is presented in Figure 2.1.

Please note that the estimation of the γ value by fitting a linear line on the log-log plot is not accurate [154]; hence, here, the mentioned maximum-likelihood estimator was used for approximation of the γ value instead.

1: $k_{\min} = 1$, p-value $= 0$
2: **while** p-value < 0.1 **do**
3: find γ using ML method
4: calculate p-value of the test
5: **if** p-value> 0.1 **then**
6: BREAK the loop
7: **else**
$k_{\min} = k_{\min} + 1$
8: **end if**
9: **end while**
10: return $k_{\min}$ and γ

Figure 2.1: Algorithm for the powerlaw adjustment to the PPIN degree distribution.

Centrality metrics and clustering coefficient

For the detailed definitions of centrality metrics and clustering coefficient used in this work, see the documentation of the graph-tool library available at: https://graph-tool.skewed.de/.

Degree correlation coefficient

Degree correlation coefficient of the network is defined as:

$$r = \frac{\sum_x e_{xx} - \sum_x a_x b_x}{1 - \sum_x a_x b_x}.$$

Here, e_{xy} denotes a fraction of links in the network that link nodes with degree x with nodes of degree y. Moreover, $\sum_{x,y} e_{xy} = 1$, $\sum_y e_{xy} = a_x$, $\sum_x e_{xy} = b_x$. When $r < 0$, the network is called disassortative, when $r > 0$ assortative, and r value equal to 0 corresponds to a random distribution of links between nodes in the network.

Average Nearest Neighbour Degree (ANND)

Average Nearest Neighbour Degree (ANND) is defined for each node i as:

$$a_{nn}(i) = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} k_j.$$

A_{ij} equals 1, when node i and j are connected, and 0 otherwise. ANND plot on y-axis has the average $< a_{nn}(i) >$ for all nodes i with degree k, and on x-axis degree values k. Normalization of the ANND plot is performed with respect to the number of nodes N.

s1 and s2 plots

To study how high-degree, middle-degree and low-degree nodes contribute to the total number of links in the network, subnetworks $G_K(V_K, E_K)$, where $V_K = \{v : k_v \leqslant K, \text{ and } v \in V\}$ and $E_K = \{E_{ij} : i,j \in V_K \text{ and } E_{ij} \in E\}$, were defined. N_K is the number of nodes in G_K and L_K, the number of links of G_K. Then, the contribution of the nodes with degree K to the total number of links L in the network can be measured with the following scores:

$$s_1(K) = \frac{L_K}{L}$$

and

$$s_2(\frac{N_K}{N}) = \frac{L_K}{L}.$$

Measuring PPIN robustness

The algorithms for measuring the network robustness are presented in Figure 2.2 and 2.3. The decomposition fractions with respect to the degree d_k and betweenness centrality d_B are defined as points where the largest connected component contains less than 1% of all nodes in the original network.

$fraction = 0$, MC is the number of Monte Carlo repetitions performed, $step$
while $fraction < 1.0$ **do**
  for i=1:MC **do**
    Initialize the network, N is the number of nodes it has, shuffle the nodes order
    Remove $fraction \cdot N$ of the top nodes
    Calculate the size (number of nodes) of the largest connected component of the network
  end for
  Average the largest connected component size for the fraction over MC repetitions
  $fraction = fraction + step$
end while
Return array of pairs (fraction:average largest connected component size)

Figure 2.2: Algorithm for measuring PPIN robustness with respect to random failures.

$fraction = 0$, sort nodes by the centrality measure, $step$
while $fraction < 1.0$ **do**
  Initialize the network, N is the number of nodes it has
  Remove $fraction \cdot N$ of the top nodes
  Calculate the size (number of nodes) of the largest connected component of the network
  $fraction = fraction + step$
end while
Return array of pairs (fraction:largest connected component size)

Figure 2.3: Algorithm for measuring PPIN robustness with respect to targeted attacks.

Failure cascade simulation

To examine the potential of each node to propagate the error in the network, the failure cascade simulations were performed. The detailed algorithm is provided in Figure 2.4. The failure fraction F, which is the initial parameter of the simulation, is the criterion of error propagation from one node to another. The simulation returns P - the percentage of nodes affected by the error propagation.

failure fraction F, GO = TRUE, provide the initial node n
set the status of n to *failed*, the rest of the nodes have the status *not failed*
while GO = TRUE **do**
  GO = False
  for each node v from the network with the status *not failed* **do**
    extract statutes of the neighbours of the node v
    if more than F·number of neighbours have a status of *failed* **then**
      change status of v to *failed*
      GO = TRUE
    end if
  end for
end while
Return P - a final percentage of nodes with the status *failed*

Figure 2.4: Algorithm for the simulation of the failure cascade.

Data analysis and visualization

All data analyses, simulations and visualizations were performed in Python 3 with the following packages: NumPy [155], Pandas [156], SciPy [157], Matplotlib [158], seaborn [159], graph-tool [160] and in R with ggplot2 [161].

2.3 Results

Three protein-protein interaction datasets were downloaded from the protein-protein interaction database IntAct [162]. The first data set, *Parkinson*, referred to protein-protein interactions related to Parkinson's disease, with particular attention given to the LRRK2 protein (Leucine-rich repeat kinase 2). It was manually curated. To observe its evolution in time, *Parkinson* was downloaded on two separate dates: 4 November 2020 and 8 November 2021. On 4 November 2020, it contained 59 912 links between 5955 proteins and on 8 November 2021, 55 930 links between 5956 proteins (22 new proteins and 70 new interactions were added, and 21 proteins and 65 interactions were lost). *Parkinson* underwent minor changes within a year, and hence, the latest version (8 November 2021) is only considered for the analysis. The second dataset, *Cancer*, was also manually curated but aimed to contain protein-protein interactions that participate in cancer. *Cancer* was also downloaded twice: on 1 February 2021 (*Cancer I*) and on 8 November 2021 (*Cancer II*). *Cancer* evolved in time much more than *Parkinson*. In the first timestamp, it was formed by 20 826 links between 5380 proteins, and in the second (*Cancer II*), it consisted of 23 263 links between 6027 proteins. *Cancer II* contained 2295 new interactions, out of which 1246 were found in a study by Adhikari and Counter [163], which focused on KRAS, HRAS and NRAS interactomes. *Cancer II* lost 75 interactions and 11 proteins present in *Cancer I*. The differences in the number of proteins and interactions between *Cancer I* and *Cancer II* were significant; hence, both datasets were used in further analysis. The last dataset, Human Reference Interactome *HuRI*, had a neutral character and covered 2-11% of the entire human interactome. *HuRI* is not manually curated, instead, it is based on a single high-throughput study, [164]. *HuRI* consisted of 162 719 links between 8204 proteins.

To estimate the completeness of the mentioned datasets, two Venn diagrams were produced for *Parkinson*, *Cancer II* and *HuRI* (Figure 2.5 and 2.6). The first represents the intersections of protein sets and the second, of interactions. It can be seen that, although the overlaps between the sets of proteins are quite significant, the intersections of the interaction sets are very poor. A complete PPIN should contain protein-protein interactions found in other works, but that is not the case here.

2.3.1 Heterogeneity of PPINs

A basic data analysis of the metadata for all datasets was performed. PPIN construction could be driven by different assumptions, starting from the idea to characterize protein-protein interactions related to the disease, as in the case of *Parkinson* and *Cancer*, and moving towards a general framework of protein-protein interactions between proteins significantly expressed in humans, as in the case of *HuRI*.

Different methodological approaches lead to the emergence of PPINs. *Parkinson* data is highly dominated by one study by Haenig et al. [165], which provided 85% of the protein-

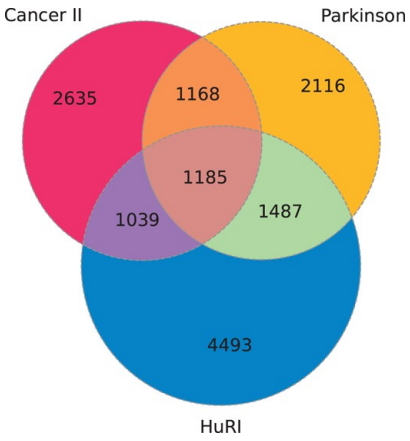

Figure 2.5: Intersection between proteins present in *Parkinson*, *Cancer II* and *HuRI* networks

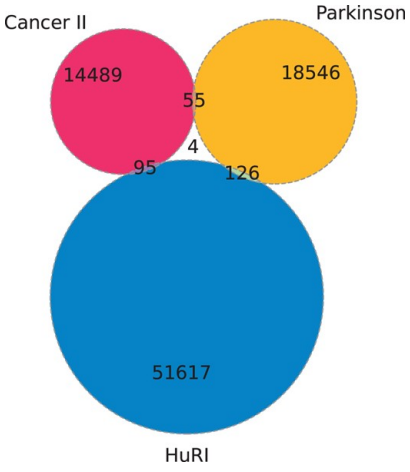

Figure 2.6: Intersection between interactions present in *Parkinson*, *Cancer II* and *HuRI* networks.

protein interactions. Haenig et al. focused on the so-called two-hybrid technique, used for protein-protein interactions identification. In consequence, this method is the main source of data in the *Parkinson* dataset. Another 10% of protein-protein interactions in *Parkinson* was provided by antitag coimmunoprecipitation experiments. *Cancer I* was less dominated by a single publication, although 30% of all recorded interactions were found by Kennedy et al. [166]. The rest was obtained from multiple studies without any other work standing out. 55% of the interactions were assigned as found by coimmunoprecipitation. Other interactions were discovered with a variety of methods including tandem affinity purification, pull-down, and protein kinase assay. *Cancer II* had similar characteristics. 52% of the records were obtained by coimmunoprecipitation and the next important group consisting of 7% of the records were identified with proximity-dependent biotin. This last group was formed by new records derived from the added study of Adhikari and Counter, which was not present in *Cancer I*. Almost all records in *HuRI* had a yeast-two-hybrid tag as an experimental method. Finally, I plotted the distribution of the IntAct MI score (Figure 2.7), which is the reliability metric of the record. The MI score values were the lowest for *Cancer* datasets, indicating their highest uncertainty.

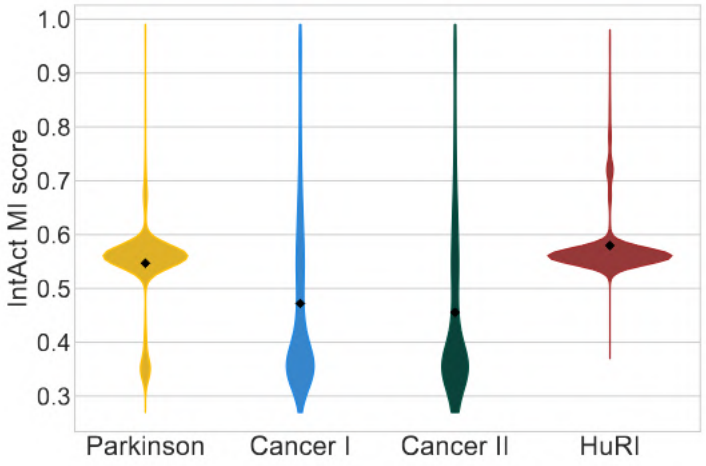

Figure 2.7: Violin plots of IntAct MI score distribution.

PPINs in IntAct may also contain non-protein entities, such as small molecule, gene, single/double-stranded deoxyribonucleic acid, molecule set and peptide. In *Parkinson* and *HuRI* such elements were rare and only 34 and 3 such elements were found, respectively. In the *Cancer* datasets, non-protein elements were more common, and 617 and 610 were found, for *Cancer I* and *Cancer II*, respectively. The non-protein items may have an important biological justification for their presence. For example, in the *Parkinson* datasets, ponatinib and imatinib, tyrosine inhibitors interacting with the LRRK2 protein, were found. Despite that, the appearance of non-protein data is unexpected when working with protein-protein interaction data.

PPIN data may contain nonhuman protein source taxids even if it concerns only the human organism,. Some of the proteins are present in a couple of versions in the database, each relating to another organism. In addition, many records are repeated in IntAct as half of the links in the analysed datasets were parallel. The detailed case studies of such repeated records led to the observation that the same protein may be studied in different variants but assigned to the database under the same identifier. For example, interactions between different oligomers of a-synuclein (SNCA in IntAct) were represented in the database as multiple self-loops for the SNCA node [167] (interaction identifiers: EBI-10690046, EBI- 10690676 and EBI-10690707). Repetitions also appear when two interacting proteins A and B were double assigned in the database, as "A interacting with B" and "B interacting with A". For further analysis, all parallel links were removed from the datasets, leading to a dramatic reduction in size. After that, *Parkinson* consisted of 18 731 records, *Cancer I* of 12 423, *Cancer II* of 14 643 and *HuRI* of 51 842.

PPINs may represent different structural connectivity. *Parkinson* consisted of 17 different components, the largest including 99.4% of all nodes and 99.8% of all links. *Cancer I* was formed by 120 components, the largest including 93.4% of all nodes and 97.6% of all links. *Cancer II* was built with 115 components, the largest including 94.3% of all nodes and 98% of all links. *HuRI* had 71 components, the largest including 98.5% of all nodes and 99.8% of all links.

The summary of all datasets after the extraction of the largest component and the removal of parallel links is provided in Table 2.1.

PPIN	Final number of nodes	Final number of links	Mean IntAct MI score	Dominating detection method	Special feature
Parkinson	5920	18 703	0.52	Yeast-two-hybrid	Exceptional focus on one protein
Cancer I	5025	12 122	0.42	Coimmunoprecipitation	Lower reliability of the links
Cancer II	5688	14 359	0.4	Coimmunoprecipitation	Lower reliability of links
HuRI	8082	51 758	0.58	Yeast-two-hybrid	General study

Table 2.1: Summary of PPINs after processing (parallel links and disconnected parts removed).

2.3.2 Structural Characterization of Available PPINs

Scale-free property

The fundamental property that describes real networks is their scale-freeness, which means that the degree distribution follows the powerlaw. As expected, all considered PPINs were characterized by a heavy-tailed degree distribution with multiple low-degree nodes and a few exceptionally high-degree ones. The relative size of the strongest hub, understood as its degree divided by the number of nodes in the network, were the following: 0.37 for *Parkinson* (LRRK2 protein), 0.064 for *Cancer I* (ESR1 protein), 0.082 for *Cancer II* (NRAS protein) and 0.064 for *HuRI* (CYSRT1 protein). The results of the powerlaw fitting are provided in Table 2.2. For *HuRI*, only a small fraction of nodes could be described by the powerlaw distribution. In addition, the calculated γ value for this network exceeded 3.

PPIN	$\hat{\gamma}$	Cutoff value	Fraction of nodes described by the power law (%)
Parkinson	2.2	10	11
Cancer I	2.16	4	26
Cancer II	2.21	5	20
HuRI	3.3	67	4

Table 2.2: Results of powerlaw fitting to the degree distributions.

To examine how subsequent groups of nodes contribute to the size of the network, I defined subnetworks that contain nodes of the original network whose degree value is lower than K. Then, the s1 plot represents how many links L of the original network, the subnetwork K contains (L_K/L). The s2 plot represents L_K/L as a function of the number of nodes in the subnetwork K divided by the total number of nodes (N_K/N), see Methods for details. With the s1 and s2 plots it is possible to investigate the contribution of high- and low-degree nodes to the total number of links in the network. The s1 and s2 plots are presented in Figure 2.8.

17% of all the links present in *Parkinson* were produced by the two biggest hubs, LRRK2 and HTT. Low-degree nodes with a degree value below 11, although covered for 90% of the nodes, produced only 3% of the links. For *Cancer*, hubs were also highly

important, but their contribution was lower, particularly in the case of *Cancer I*. The two biggest hubs produced 5% of all the links in *Cancer I* and 7% in *Cancer II*. Similarly, 91% of the nodes with degree below 9 were responsible for 9.5% of all the connections in *Cancer I*, and 8% in *Cancer II*. In *HuRI*, the hubs' impact was the lowest. Small degree nodes, with degree values below 34, accounted for 90% of the nodes and produced 20% of all connections.

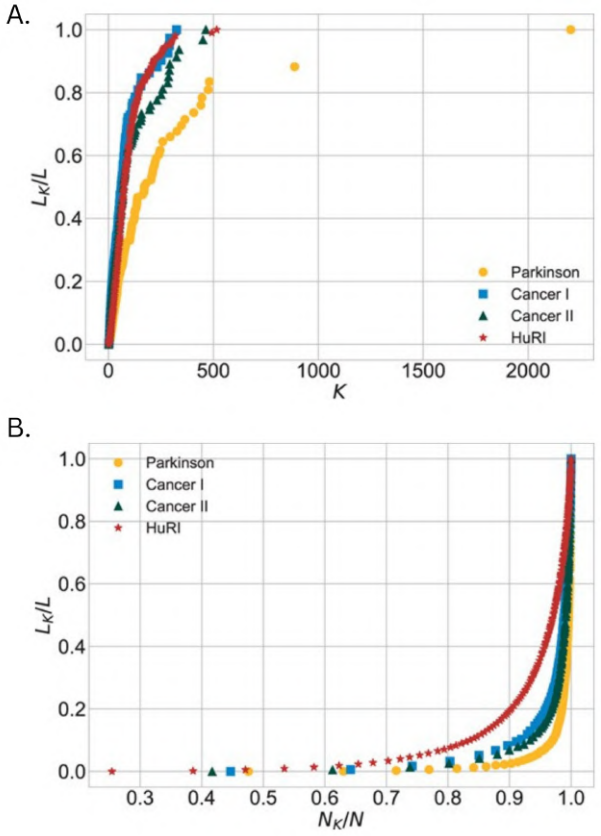

Figure 2.8: Impact of nodes with respect to their degree. A. s1 plot, B. s2 plot.

Assortatitivy

The PPINs are commonly expected to be disassortative. In consequence, their degree correlation coefficient is below 0. This was observed for *Parkinson* (r=-0.018), *Cancer I* (r=-0.001) and *Cancer II* (r=-0.011). The Average Nearest Neighbour Degre (ANND) plots (Figure 2.9) for these networks contained hyperbolic-like scatterings that are the next indicators of disassortativity. *HuRI* network broke out from this model. The degree correlation coefficient for this network was equal to 0.008 and the ANND plot had different scatterings that pointed towards assortativity.

Clusterization

The mean values of the clustering coefficient were 0.1, 0.13, 0.19 and 0.06 for *Parkinson*, *Cancer I*, *Cancer II* and *HuRI*, respectively. For *Parkinson*, *Cancer I*, and *Cancer II*, the

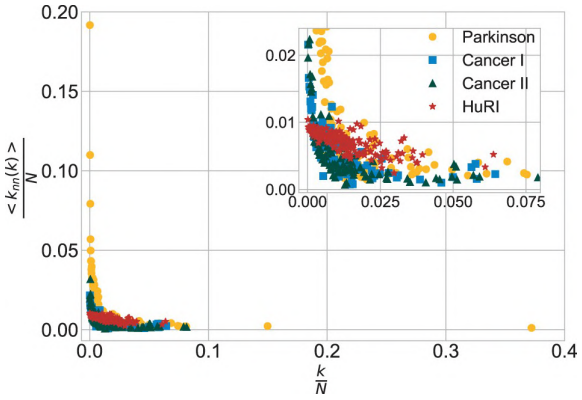

Figure 2.9: Average nearest neighbour degree (ANND) plot for PPINs of choice.

distributions of this parameter (Figure 2.10) had two regions of concentrations - one around zero and the other, smaller, around 1.0. HuRI almost lacked the second region in the distribution, indicating that highly connected regions, observed for the other networks, were not present.

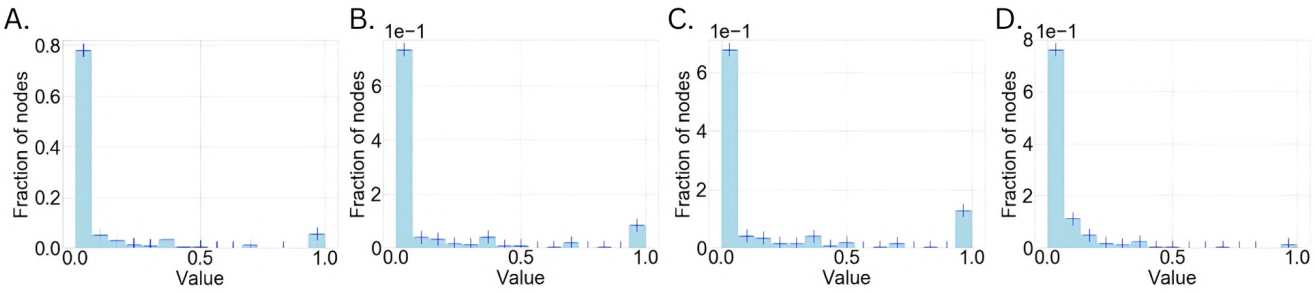

Figure 2.10: Histograms of clustering coefficient for A. *Parkinson*, B. *Cancer I*, C. *Cancer II*, D. *HuRI*.

Robustness

All networks were similarly robust to simulations of random failures. Their failure scatterings resembled a linear trend (Figure 2.11). The difference between networks was observed for the simulation of tailored attacks with respect to different centrality metrics. Degree and betweenness centrality led to a similar pace of the network decomposition, which differed from the pace observed for the eigenvector and closeness centrality, which resembled one another. The docomposition fractions with respect to the degree and betweeness centrality are provided in Table 2.3.

The exact scattering trends for the networks of choice also differed. In *HuRI*, all of them had a smooth, similar shape, in contrast to the results observed for *Parkinson*. *Cancer I* and *Cancer II* were somewhere in between the smooth scatterings found for *HuRI*, and uneven ones noticed for *Parkinson*. *Parkinson*, *Cancer I* and *Cancer II* seemed to be similarly vulnerable to the targeted attacks, meanwhile, *HuRI* was more immune to them.

PPIN	d_k	d_B
Parkinson	0.08	0.16
Cancer I	0.1	0.14
Cancer II	0.09	0.15
HuRI	0.31	0.37

Table 2.3: Decomposition fractions for the simulation of tailored attacks with respect to the degree d_k and betweeness centrality d_B.

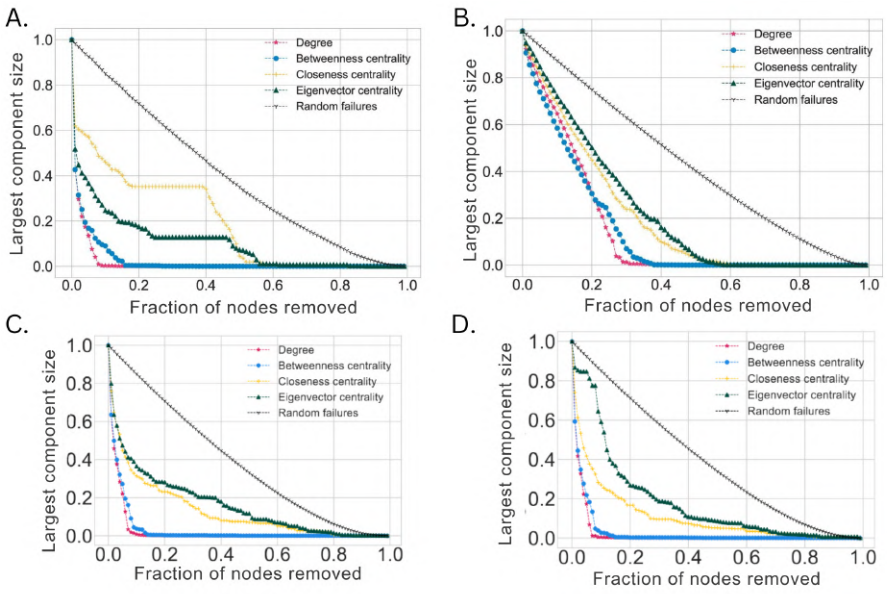

Figure 2.11: Size of the largest component as a function of the fraction of nodes removed for A. *Parkinson*, B. *HuRI*, C. *Cancer I* and D. *Cancer II*.

Error propagation

To examine how quickly the error may propagate in the networks, a failure cascade simulation with respect to the parameter F was performed; it was the first time that such a simulation was performed for a PPIN, to the best of our knowledge. In the failure cascade model, a starting node is first marked as *failed*. In the iterative process, each node is evaluated, and its status changes to failed if a fraction F of its neighbours failed. The loop is repeated as long as at least one node changes the status to *failed* in the previous iteration over the network. At the end, the algorithm returns a percentage P of nodes with the *failed* status (see Methods for details). All networks had the heavy-tailed distributions of the failure cascades. The percentages of nodes capable of propagating an error under different F values are provided in Table 2.4. The maximal P values for different networks are shown in Table 2.5. Most of the nodes capable of highly propagating an error in the network, apart from LAMP2 in Parkinson, were significant hubs and bottlenecks in their respective networks.

PPIN	F=0.25	F=0.5	F=0.75
Parkinson	10%	5.7%	3.5%
Cancer I	23%	15%	9%
Cancer II	20%	13%	8%
HuRI	3%	2%	1%

Table 2.4: Percentages of nodes capable of propagating an error.

PPIN	F=0.25	F=0.5	F=0.75
Parkinson	99% (LRRK2, LAMP2)	28% (LRRK2)	21% (LRRK2)
Cancer I	4% (NUDCD1)	3% (HSPB1)	2% (HSBP1)
Cancer II	29% (AR, RAVER1)	3% (HSPB1)	2% (HSBP1)
HuRI	2% (MEOX2)	1.5% (MEOX2)	1% (MEOX2)

Table 2.5: Maximal failure cascade sizes (P_{max}).

2.3.3 ETNA

All the discussed methods of topological analysis were encapsulated in one simple tool, ETNA, which stands for Extensive Tool for Network Analysis. ETNA was built on the graph-tool library, partially developed in C++, which guarantees high-speed performance, as other popular network libraries are too slow to deal with large protein-protein interaction datasets [160]. ETNA combines multiple Python and R packages into one single pipeline with the rpy2 library. The graphical user interface was constructed with ipywidgets [168].

The repository with ETNA source code and datasets analyzed in this chapter is available at:
https://github.com/AlicjaNowakowska/ETNA.
ETNA can be used through Google Colab using the link:
https://githubtocolab.com/AlicjaNowakowska/ETNA/blob/main/ETNAColab.ipynb.

2.4 Discussion

Protein-protein interactions are at the core of molecular systems. Recent developments in high-throughput studies have provided us with a huge increase in available data, giving hope for the acceleration of scientific discoveries. This direction seemed a promising approach in the context of neurodegenerative diseases that are characterized by multi-level complexity.

The first step in any data analysis is the diagnosis of the provided data. This requirement was fulfilled in this chapter. A qualitative and topological examination of three available protein-protein interaction datasets was conducted. These included a single experiment-based network with uniform sampling of PPIs (Human Reference Interactome - *HuRI*) and literature-based networks (*Cancer I*, *Cancer II* and *Parkinson*). In the first step, the data sources that led to the construction of the PPINs were investigated. Secondly, the topology of PPINs was examined with classical and self-developed methods. The main conclusion of this study points to the high influence of the data-gathering procedure on the PPIN topology, which may later affect the biological interpretation. It

is suggested that the discovery of topological features similar to the ones described here for different biased PPIN may help to assess the level of interest bias introduced in the dataset of choice and draw attention to its cautious biological interpretation.

The available protein-protein interaction data was unexpectedly heterogeneous and biased. Analysis of the *Parkinson* dataset revealed that 85% of the links had a single source from the experiment performed by Haenig et al. *Cancer II* differed from *Cancer I*, mostly due to the inclusion of multiple interactions from a one single experimental publication in *Cancer II*. The annotation of the all datasets was heterogeneous. Non-protein nodes, non-human taxids, temporal inclusion or removal of the interactions in the network evolution, parallel links, usage of different experimental methods and different quality of the interactions could be observed in the datasets. Single-experiment-based PPIN, such as *HuRI*, although less heterogeneous, can also be problematic, and multiple proteins and interactions may be missed. For example, a yeast-two-hybrid experiment, which was used to build *HuRI*, may result in significantly different interactomes for the same organism when repeated [169, 170, 171].

The analyzed PPINs had a generally typical topology for biological networks, confirming that they provide an overview of their respective biological phenomena. However, the detailed results revealed deviations from the expected characteristics. The scale-free property, which is a fundamental characteristic of many real-world networks, was observed for all networks. The careful adjustment of the power law distribution, expected for scale-free networks, confirmed a typical value of the degree exponent γ (between 2 and 3) for *Parkinson*, *Cancer I*, *Cancer II*, which remained in line with other studies [171]. The situation for *HuRI* was more complex, as the power law distribution could be fitted to a small fraction of nodes and the estimated γ exceeded 3, which is unusual for biological systems. Such structural disturbances present in *HuRI* may affect the detailed biological interpretation of the network.

The scale-free structure is driven by the evolutionary needs to protect the network from random perturbations. Nevertheless, makes it susceptible to targeted attacks, which may be for example, a medical intervention that inhibits certain enzymes. To understand how robust the network is towards targeted attacks, one may delete nodes one by one according to different centrality metrics and measure the size of the resulting largest connected component. Networks with a tight-knit structural pattern require that a large fraction of nodes must be removed to decompose the network. *HuRI* had such a tight-knit structure, as hubs were less impactful. In consequence, a higher fraction of the most important nodes needed to be deleted to destroy the network. On the other side of the spectrum, there were *Parkinson*, *Cancer I* and *Cancer II* networks. In these cases, the removal of only a handful of nodes with the highest degree or betweenness centrality led to the complete network decomposition. These fractions were also significantly lower than those found in another study of a yeast PPIN, where $d_k = 0.2$; $d_B = 0.25\%$ [149]. The different methods of interactome sampling also affect the smoothness of scatterings generated from the robustness examination. In the *Parkinson* study, which gave special attention to the LRRK2 protein, the scatterings were irregular. For more uniform, though incomplete *HuRI*, all points were smoothly decaying. Both behaviours were found for different PPINs, suggesting that the data-gathering procedure highly affects the structures of PPINs.

The disassortative linking pattern, which means that high-degree nodes link to low-degree ones, was confirmed for *Parkinson*, *Cancer I* and *Cancer II* (degree correlation coefficient values were below zero). Disassortativity is common for experiment-derived

and literature-based PPINs [150, 149]. *HuRI* stood out from this pattern exhibiting an assortative nature, which, although rare, sometimes can also be found [150]. The assortative structure results in a slightly different network robustness behaviour. In such a case the decomposition is initially faster for betweenness centrality than for degree, as noted for *HuRI* [149].

Although for all networks the distributions of degree, betweenness centrality and eigenvector centrality were heavy-tailed, the PPINs differed in their tail lengths. This was particularly visible for *Parkinson*. The maximal values of these centrality measures, when also compared with Ran et al. [172], were exceptionally high. Further examination of the contribution of low- and high-degree nodes to the number of links in the network with s1 and s2 plots confirmed this unusual impact of high-degree nodes. In *HuRI*, hubs were less impactful and the number of middle-degree nodes was higher reflecting the nature of the uniform sampling of the interactome. In contrast, the data bias toward LRRK2 protein in the *Parkinson* network, caused the appearance of highly significant hubs in this PPIN. *Cancer I* was between *Parkinson* and *HuRI*. *Cancer II*, which was characterized by a higher publication bias, tended towards *Parkinson*. It can be concluded that the increase in data size may come at the price of a higher bias in the dataset. Furthermore, it can also impact the biological interpretation. The inclusion of new data in *Cancer II* resulted in the topological changes that led to a new set of most influential proteins, different from that in *Cancer I* and obtained in other cancer-related studies of PPINs [173]. As hubs and bottlenecks are frequent drug targets, the biases in the data may impede the identification of the proteins most important in the biological processes, instead revealing proteins which are most interesting for the scientists.

The influence of scientific interests could also be seen in the failure cascade simulations. The failure cascade simulates the effect that one malfunctioning protein may have on the entire molecular system. As expected, most of the proteins had a low error propagation potential, indicating that their malfunctioning would not affect the larger regions of the network. Naturally, proteins with high centrality measures had the greatest capability to propagate the error. However, for *Parkinson* and *Cancer II*, exceptions from this pattern appeared, and less important nodes in the network also had high error propagation potential. This unexpected result suggests that the introduction of the interest bias in the data resources has a multi-level impact on the structure of the PPINs, posing doubt on the reliability of the detailed biological interpretation of such systems.

The modularity of the network also changes with the methodology applied to construct the network. For *Parkinson*, *Cancer I* and *Cancer II* regions with high clustering coefficient values were found. Additionally, the mean value of this parameter for these networks corresponded to the literature [149]. The results for *HuRI* were somewhat different as no nodes with high clustering values were identified. It could be that high levels of clusterization are not necessary to form functional modules. On the other hand, this could be the result of the uniform sampling of the human interactome.

Data quality analysis of the PPINs seems to be of fundamental importance for their biological interpretation. *HuRI* significantly differed from other PPINs, including ones described in the literature. The different topological characteristics of *HuRI* could suggest a different dynamic of the interactions than in other interactomes. However, the structural discrepancies result from the incompleteness of the data. The yeast-two-hybrid systems, as the ones used in *HuRI*, cannot provide the data on proteins which cannot be expressed in yeast, hence the full interactome data cannot be retrieved. According to the results, *Cancer I* aligned the most with the expected topology. Although it was not the largest, the

interpretation of the molecular pathways and analysis of the drug targets with centrality measures seemed possible. *Cancer II* followed this trend, although the data bias was more visible.

The *Parkinson* network was strongly disturbed toward one protein. For this network, the identification of drug targets with the centrality metrics, would point toward to the LRRK2 protein. As the two most influential hubs provided 20% of the links in this network, one could conclude that these two proteins are the most essential for the disease and should be the focus of further research. The failure cascade simulations would confirm these observations. In reality, other proteins can also be important but scientific interests impede this discovery. Hence, the biological interpretation of the *Parkinson* network is easily biased by the data it has.

Scientific interests are a general problem in neurodegenerative disease studies. Many researchers focus on the few prominent amyloids such as Aβ or α-synuclein. This was clearly seen in the Amylograph project, which studied publications on amyloid protein interactions [65]. Amyloids may interact with each other leading to the formation of heterofibrils or affecting the rates of aggregation. The Amylograph database provides a network representation of amyloid-amyloid interaction. 20% of the reported interactions regard Aβ, and another 15% α-synuclein, revealing a significant scientific bias toward these proteins that impedes the wide exploration of other scientific directions. It is also visible in the Google Scholar engine for publication databases. 650 thousand records can be found for 'Parkinson's disease', and 144 thousand, around 20% of the total, for 'Parkinson's disease and α-synuclein'. It is even more prominent for the "Alzheimer's disease" search, which results in 2 million publications; meanwhile, for 'Alzheimer's disease and amyloid beta', over one million hits are identified. This suggests that a significant part of the research focuses on these two proteins, potentially introducing a significant bias in our understanding of neurodegeneration.

How to detect interest and methodological biases in the data? It can be suggested that in the case of PPINs the topological analysis can give a strong suspicion about them. Assortative patterns, lower clusterization, and lower importance of the hubs are signs that the PPIN data lacks certain information, and the biological interpretation of the network structure is also troublesome at this point. On the other hand, if a couple of proteins highly dominate the network, they widely propagate the errors in the network, and their deletion leads to the very fast network decomposition; bias towards this group may be suspected. To ease the detection of such biases, I developed an easy-to-use Colab notebook that incorporates all presented methods and guarantees quick results.

This proves the first hypothesis of this thesis that *exploratory topological analysis of available protein-protein interaction datasets can provide a new perspective on neurodegeneration* and makes the first goal *To assess the quality of the real-world protein interaction networks through topological analysis and identify if topological analysis can shed new light on disease mechanisms* realized.

The analysis conducted in this chapter revealed that our knowledge about neurodegenerative diseases, and particularly Parkinson's disease, is biased toward certain groups of human proteins, and particularly human amyloid proteins. As a consequence, at this point, it is impossible to get a full picture of the molecular mechanisms that underlie these disorders. To broaden our knowledge, we must sample other scientific directions as well. Detailed studies of other proteins, and particularly nonhuman proteins expressed by organisms that inhabit us, could give valuable insight into the different biological processes involved in neurodegeneration. Hence, the following chapters will focus on an

understudied group of bacterial proteins - functional amyloids that may also have clinical significance but are not widely represented in the available databases and publications.

Results presented in this chapter were published as: ***Nowakowska, A. W.****, Kotulska, M. (2022). Topological analysis as a tool for detection of abnormalities in protein–protein interaction data. Bioinformatics, 38(16), 3968-3975.*

Less-studied amyloids: bacterial functional amyloids and their sequence analysis

3.1 Introduction

Throughout the decades, amyloid fibril formation was seen as a negative event associated with the onset and progression of the disease. The discovery of functional amyloids around 20 years ago challenged this perspective, changing our understanding of the function of amyloids in organisms. The characteristic properties of amyloid fibrils, such as stability, regularity and low solubility, make them a great material that nature uses for various purposes.

The world of functional amyloids spans different kingdoms and functions. In humans, physiological amyloid fibrils of Pmel17 can be found in melanosomes, where they promote pigmentation [174]. In higher-order mammals, amyloid structures were discovered to serve as storage reservoirs for various hormones [175]. Amyloid filaments are an inherent part of spider silk [176]. In yeast, they participate in the signalling process that regulates the immune response [177]. In bacteria, amyloid fibrils are often part of biofilm matrices. Probably the most well-studied example is curli protein (CsgA) from E. coli, which forms amyloid fibrils that protect a bacterial colony.

Our understanding of functional amyloids is still limited. Little is known about their aggregation-prone regions. In the case of CsgA and CsgB, certain repeat regions are aggregation driving. The same holds for FapC. But, in many other cases, especially in longer bacterial amyloid proteins, the mechanisms of fibril formation are elusive. Only in a few examples do we have any parts of their structures experimentally solved. Despite that, the scarce structural data of functional fibrils already highlight the differences between functional and pathological amyloids. For example, the core of the amyloid fibril of the Orb2 protein, which is related to memory recall in Drosophila, was found to be hydrophilic and not hydrophobic, as in the case of pathological amyloids [178]. Functional amyloid fibrils of phenol-soluble modulins, which are found in the biofilm of Staphylococcus aureus, can reveal a cross-alpha pattern instead of the expected cross-beta [179]. As amyloid experiments are tedious, time-consuming, and expensive, computational approaches are needed to speed up the research.

Bacterial functional amyloids are particularly interesting as they are present in the human microbiome, which is dominated by bacteria. Their structural similarity to pathological amyloids and the possibility of interactions with human proteins make them a very interesting, though understudied, case. To broaden our understanding of bacterial functional amyloids, a series of bioinformatics analyses are performed. In the first step, a dataset of the known bacterial functional amyloids is prepared based on the litera-

ture search. Next, the gathered data is analysed from the sequence (this chapter) and structure (the following chapter) perspectives. As little is still known about bacterial functional amyloids, the analysis is often enriched with knowledge about other amyloid proteins, including pathological ones. Although they differ from functional amyloids, they have been subject to extensive research that can provide value in our understanding of the aggregation of their functional counterparts.

The analysis of the sequences of bacterial functional amyloids starts with the prediction of the ProtBERT protein embeddings. They are then used to visualize the relations between proteins from the prepared dataset of bacterial functional amyloids. Then, a more standard approach with repeat detection is used to examine these sequences.

The modularity of amyloid fibrils inspires the search for regularity in the sequence. In the protein world, the emergence of regular, larger, symmetric structures is frequently associated with the appearance of smaller structural and sequence motifs. Amyloid fibrils are a perfect example of larger symmetrical structures built of repeated units. In contrast to pathological amyloids, their fibril formation is often reversible, as their presence may no longer be beneficial, for example, in the absence of the stimuli [119, 132, 131]. The power to disassemble requires that multiple structural modulators are encoded in these proteins. Examples include the incorporation of charged residues [133]. It seems reasonable that tandem repeats could be another purposefully designed regulatory mechanism as they often have a strict role in the proteome.

In this chapter, the second hypothesis that ***sequence tandem repeats influence the aggregation of bacterial functional amyloids*** is investigated. For this goal, the presence of tandem repeats in bacterial functional amyloids, their imperfection and size, is studied and discussed along with the results from simple molecular simulations. We compare the results to the available literature that considers the role of the repeats in other amyloid-forming proteins to get inspiration about their potential role in bacterial functional amyloids. This chapter is an important step forward in our understanding of molecular machinery governing bacterial functional amyloid aggregation and may prove useful in future designs of aggregating peptides.

3.2 Methods

Dataset of bacterial functional amyloids (BFA)

The dataset of the known bacterial functional amyloids, referred to as BFA, was built based on state-of-the-art publications regarding functional amyloids [78, 180]. This data was further extended through Google Scholar searches. The aggregation propensity of each protein in the BFA has been experimentally confirmed.

Bacterial proteins

Bacterial proteins were downloaded from Uniprot by applying the filters “Taxonomy:Bacteria” and “Protein Existence:protein level” in September of 2022, giving 39 526 sequences. This reference dataset was used for the statistical analysis of repeat patterns between bacterial functional amyloids and other bacterial proteins. It might contain unknown and known bacterial functional amyloids, but probably in negligible numbers.

Human proteins

Human proteins used in section "Known bacterial functional amyloids as a separate protein group" were downloaded from Uniprot by applying filters "Taxonomy:Homo Sapiens" and "Protein Existence:protein level" in March of 2023. This gave 69 176 sequences.

ProtBERT embeddings

The ProtBERT embeddings for proteins were calculated with the ProtTrans pipeline available in Python for ProtBert-BFD pretrained-model [181]. For sequences with the length above 3000 amino acids, the embeddings were not calculated due to limitations of the computational power. The dimension reduction using the Principal Component Analysis (PCA) of the embeddings matrices was performed with the sklearn package in Python [182].

BFA network visualization

The ProtBERT embeddings were produced for each of the BFA proteins. As a result, a matrix consisting of 38 rows, concerning 38 BFA proteins, and 1024 columns was produced. It was then reduced with Principal Component Analysis (PCA) and 4 first columns, which explained 70% of the variation in the data, were extracted. Hence, a 4-column long vector describing each of the BFA proteins was obtained. The distances between BFA proteins, in this new reduced space, were calulated and only 15% of the shortest ones, to limit the number of edges in the network, were extracted to build the network. Nodes in the network represented BFA proteins and edges apeared if the distance between the proteins in the reduced ProtBERT spaces was in the top 15% of the shortest distances. For network visualization, the networkx package in Python was used [183].

Repeat definition

Repeat in a protein sequence is understood as a set of similar motifs, each denoted as a repeat unit, which are close to one another in the sequence.

Repeat detection

To detect the repeats in the proteins, RADAR software in the desktop version with default parameters was used [184]. Although RADAR can provide false positive results and omit certain repeat parts, it performed the best on well-annotated, CsgA and CsgB proteins when compared to HHrepID [185] and T-Reks [186].

Imperfect repeats

Similarity between the repeat units was assessed with the Multiple Sequence Alignment provided by RADAR for each detected repeat. For each MSA column, the frequency of appearance f of the most common amino acid was calculated. The identity between the repeat units is defined as the average f over the columns - f_{avg}.

Amino acid profile of the repeats

The amino acid profile was analyzed for the following four four datasets:

- BS, complete bacterial proteins sequences (the reference dataset from the Uniprot)
- BR, all repeat units detected for all bacterial sequences
- AS, complete sequences of bacterial functional amyloids (BFA dataset),
- AR, all repeat units detected for BFA proteins

For each dataset and for each sequence, the frequency of each amino acid f_{aa} was calculated. The average frequency f_{avg} is understood as the mean f_{aa} over all sequences in the corresponding dataset.

The distribution of f_{aa} vectors were compared with the Mann-Whitney U test. Mann-Whitney U test is a nonparametric, rank-based test that is often used as an alternative to the t-test. It tests the null hypothesis that two samples have different stochastic ordering [187]. As multiple such tests were performed, Bonferroni correction was applied to the obtained p-values.

Secondary structure prediction of the repeats

The secondary structure of each complete sequence from the BFA dataset with detected repeats by RADAR was predicted using PSIPRED webserver [188]. Due to the length limitations, six proteins were excluded from the analysis: Aap, Bap, Esp, PAc, SasG and YghJ.

Data analysis and visualization

All data analyses and visualizations were performed in Python 3 with the following packages: NumPy [155], Pandas [156], SciPy [157], Matplotlib [158], Seaborn [159].

3.3 Results

The dataset of the bacterial functional amyloids consists of 38 proteins presented in Table 3.1.

Protein	Uniprot ID	Reference
Aap (Staphylococcus epidermidis)	Q5HKE8	[189]
SasG (Staphylococcus aureus)	Q2G2B2	[190]
Bap (Staphylococcus aureus)	Q79LN3	[191]
Esp (Enterococcus faecalis)	Q9Z4N7	[87]
Pac (Streptococcus mutans)	P11657	[192]
CsgA (Escherichia coli)	P28307	[191]
CsgB (Escherichia coli)	P0ABK7	[191]
FapC (Pseudomonas sp. UK4)	C4IN70	[191]
FadA (Fusobacterium nucleatum)	Q4U4F1	[193]
TasA (Bacillus subtilis)	P54507	[194]
WapA (Streptococcus mutans)	P11000	[195]
Mtp (Mycobacterium tuberculosis)	P9WI87	[196]
ChpD (Streptomyces coelicolor)	Q9L1J9	[197]
ChpE (Streptomyces coelicolor)	Q9X9Z2	[197]
ChpF (Streptomyces coelicolor)	Q9KYG7	[197]
ChpG (Streptomyces coelicolor)	Q9KYH3	[197]
ChpH (Streptomyces coelicolor)	Q9AD92	[191]
RdlB (Streptomyces coelicolor)	Q934F8	[191]
CarD (Mycobacterium tuberculosis)	P9WJG3	[191]
Tuf (EF-Tu) (Gallibacterium anatis)	A0A263HIU7	[191]
YhgJ (Escherichia coli)	P0CK95	[191]
HelD (Bacillus subtilis)	O32215	[196]
Hfq (Escherichia coli)	P0A6X3	[198]
Microcin (Klebsiella pneumoniae)	Q9Z4N4	[191]
SuhB (Staphylococcus aureus)	A0A0U1MJW7	[199]
RopA (Rhizobium leguminosarum)	Q05811	[200]
RopB (Rhizobium leguminosarum)	Q52866	[200]
Hpn (Helicobacter pylori)	P0A0V6	[201]
Smu63c (Streptococcus mutans)	Q8DWI5	[195]
Spb (Staphylococcus epidermidis)	Q5HRC3	[202]
Hpag (Xanthomonas campestris)	Q83XF9	[203]
Psma1 (Staphylococcus aureus)	A9JX05	[191]
Psma3 (Staphylococcus aureus)	A9JX07	[191]
Psmb1 (Staphylococcus aureus)	A0A0H3KCA8	[204]
Psmb2 (Staphylococcus aureus)	A0A0H3KSC6	[204]
AgrD (Staphylococcus aureus)	Q53643	[191]
TapA (Bacillus subtilis)	P40949	[85]
FapB (Pseudomonas sp. UK4)	C4IN69	[127]

Table 3.1: List of bacterial functional amyloids (BFA dataset).

3.3.1 Characterization of the dataset of bacterial functional amyloids (BFA)

The BFA dataset consists of 38 bacterial functional amyloids. Five of the identified proteins in the BFA can be characterized as biofilm-related, long and multidomain. These are: Aap, its ortholog SasG, Bap, its ortholog Esp, and PAc. All of them contain a characteristic cell wall anchor domain LPXTG and undergo cleavage. Bap and Esp additionally have an EF-hand motif responsible for calcium binding. Other biofilm-related proteins are: WapA, which is a large and a poorly studied cell-wall protein, secreted TasA, TapA and Smu63, already mentioned Csg, Fap and phenol-soluble modulins proteins and their expression regulator AgrD, extracellular Mtp that forms similar fibrils to CsgA, adhesion protein A - FadA, inositol phosphatase SuhB that regulates multiple biological functions such as expression of virulence factors, exopolysaccharide biosynthesis and biofilm formation. In the dataset, there were also proteins related to RNA and DNA binding that undergo important structural changes upon binding: cytoplasmatic Hfq involved in transcription, Tuf which binds GTP to transport aminoacylated tRNAs to the ribosome, and transcription regulator CarD that binds RNA polymerase and stabilizes transcription initiation complex. The BFA dataset also contains chaplins, which are a family of proteins involved in hyphae formation in filamentous bacteria, RdlB, which has a similar function as chaplins, RopA and RopB that take part in symbiotic interactions, a secreted metalloprotease Hpn with a compositional bias towards histidine residues, secreted metalloprotease lipoprotein YghJ and plant cytotoxic harpin Hpag.

3.3.2 Visualization of the BFA dataset

To better understand the relations between the proteins in the BFA dataset, the BFA dataset is visualized as a network (Figure 3.1). Because little is known about the proteins in the BFA dataset and the information about them in common databases, like Uniprot or Interproscan, is often scarce, each protein (node) was described using ProtBERT embeddings. Protein embeddings result from large protein language neural networks that are trained on multiple proteomes. They aim to encode structural and functional information about the proteins. This makes them particularly useful in the case of less-known proteins, such as bacterial functional amyloids. ProtBERT is one such state-of-the-art model and was successfully used e.g. for Gene Ontology predictions [123]. The space of the embeddings was reduced with Principal Component Analysis, the distances between BFA proteins were calculated, and the network with 38 nodes representing the BFA proteins, and 90 edges between them, representing only the most significant similarities between the proteins in reduced ProtBERT space, was built.

The produced visualization is the map of the known bacterial functional amyloids that allows for the easy detection of the functional clusters in the BFA dataset. The main group is formed by biofilm-related proteins, that regardless to their size are similar in the considered reduced ProtBERT space. Tuf, Hfq and CarD proteins are separated from the biofilm-related proteins, but similar to one another, as all are related to DNA and RNA binding. The last group is formed by the family of chaplins, which do not resemble other proteins. Finally, Hpn, is separated and unlike other groups.

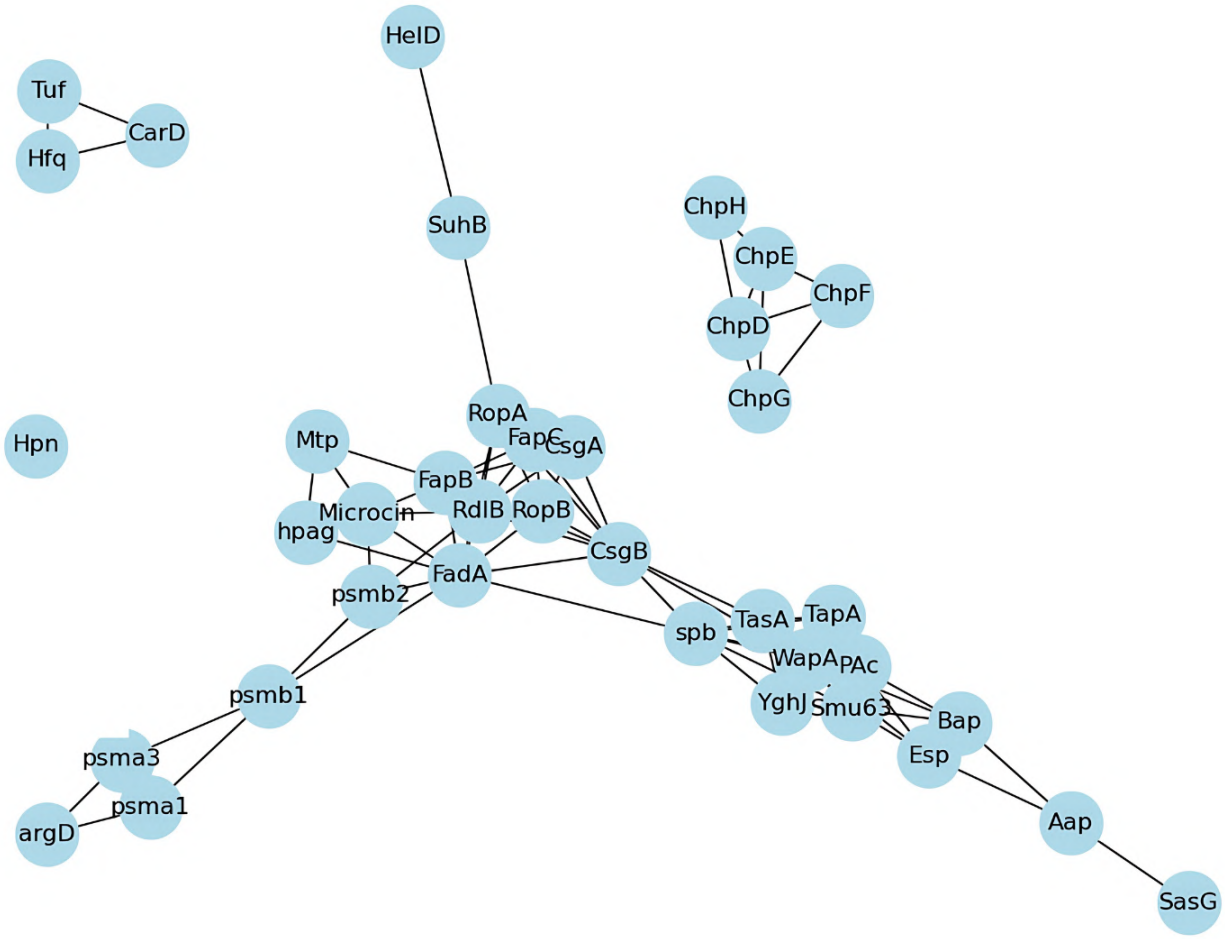

Figure 3.1: Visualization of the BFA dataset in the form of the network built on the distances in the reduced ProtBERT space.

3.3.3 Known bacterial functional amyloids as a separate protein group

To examine whether BFA proteins are a relatively homogeneous group that can be easily separated from other bacterial proteins, a new dataset that consisted of the sum of the BFA dataset and a random sample of bacterial proteins (BFA∪ Random) was extracted. Different sizes of the random proteins were tested, so that the contribution of the BFA dataset to the newly created dataset ranged from 0.1 to 0.7 (|BFA|/|BFA∪Random|). Then, the ProtBERT embeddings for BFA∪ Random were calculated and subject to the dimension reduction with Principal Component Analysis to account for 70% of the variation. For each BFA protein, its one, three, or five nearest neighbours were constructed. Finally, the fraction of the nearest neighbours belonging to the BFA dataset was calculated. This procedure was repeated 100 times (Figure 3.2).

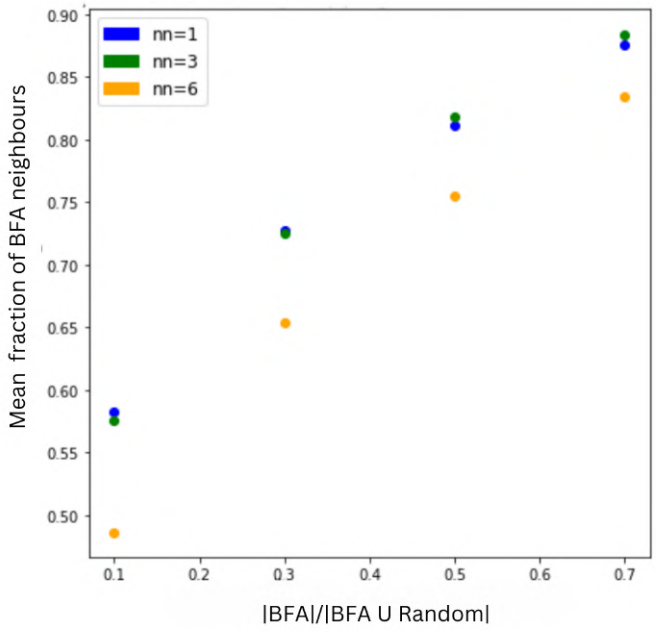

Figure 3.2: Separation of BFA proteins from all bacterial proteins. The nn denotes the number of nearest neighbours considered.

The calculations indicate that the BFA dataset is easily separable from other bacterial proteins, probably due to the high population of biofilm-related proteins, which must have characteristic embeddings in the space generated by ProtBERT.

3.3.4 Repeats in the bacterial functional amyloids

The observed relative homogeneity of the known bacterial functional amyloids, as a group, raises a question about their other possible common traits. Therefore, the following sections discuss the appearance of sequence repeats in the BFA dataset and their potential impact on aggregation.

As few bacterial functional amyloids are still known, it seems worth supporting our view on the role of the repeats in these proteins with examples from other amyloid proteins. Hence, the analysis of the repeats is in a two-fold manner. First, as an inspiration, we

Protein	Uniprot ID	Reference
CsgA	P28307	[205]
CsgB	P0ABK7	[206]
Sup35	P05453	[207]
Het-s	Q03689	[208]
Pmel17	P40967	[174]
Silkmoth chorin protein class A	-	[209]
FapC	C4IN70	[210]
PrP	P04156	[211]
Htt	P42858	[212]
Tau	P10636	[213]
α-synuclein	P37840	[214]
Apolipoprotein A-1	P02647	[215]
Apolipoprotein A-IV	P06727	[216]

Table 3.2: List of known amyloid proteins whose sequence repeats

analyse examples of how the repeats were found to work in other amyloid studies, even if the presence of such repetitions is accidental in these proteins. Then, the resulting theses are evaluated on the studied dataset of bacterial functional amyloids.

3.3.5 Abundance of the repeats

Other amyloid studies

The repeats in amyloid proteins were reported in a variety of cases. Both, pathological and functional amyloids were found to contain sequence repetitions that could modulate their amyloid formations. Amyloid proteins with sequence repeats that influence the aggregation process, according to experimental data, are provided in the Table 3.2.

Bacterial functional amyloids

The repeat detection and characterization could be performed for 33 bacterial functional amyloids, as 5 proteins were too short for the analysis (phenol-soluble modulins and N-AgrD). The detection of the repeats was performed with RADAR software and revealed positive results in 25 out of 33 studied sequences (73%). Proteins without detected repeats included chaplins related to rodlet layer formation (ChpGED), spb, mtp and Hfq. The polyH repeats present in Hpn were not found by the tool and hence, this protein was also excluded from the further analysis. In total, 69 different repeats in 25 proteins were detected. For comparison, in 38058 bacterial proteins found in Uniprot, RADAR yielded the repeats in 67% of cases. The difference was not significant with the Fisher exact test (p-value=0.3).

3.3.6 Number of units in a repeat

Other amyloid studies

It has been shown that a number of units in a repeat affects the amyloid characteristics of a protein. Such a result has been reported for the PrP protein (prion protein). It has 5

octapeptide repeats located in the N-terminus that are responsible for zinc uptake. PrP variants with a higher number of units in a repeat have been shown to be more aggregation-prone [211]. The same has been found for α-synuclein. In its native form, α-synuclein uses its 7-units repeat to create a lipid-binding alpha-helical structure [217]. A higher number of units in a tandem repeat inhibits beta-formation, affecting the aggregation propensity of this protein. However, it comes at the price of lower lipid binding affinity, suggesting the evolutionary trade-off between amyloid formation and lipid binding [214]. The impact of the repeats on the aggregation pace has also been observed for bacterial functional amyloid FapC that contains three repeat units, whose deletion slows down the aggregation [210]. The same was noted for the CsgA protein; mutants without certain units experienced a dramatic loss in their aggregation rates [218].

Bacterial functional amyloids

The number of units in each of the detected 68 repeats was calculated (Figure 3.3). 91% of the repeats in bacterial functional amyloids had 10 or fewer units. This trend was also observed in general. 99% of the bacterial proteins with the repeats had 10 repeat units or less. The analysed bacterial functional amyloids seemed to be enriched in repeats with 3-6 units, in comparison to bacterial proteins (Fisher exact p-value=0.0007). The tails of the distribution of the number of units in a repeat differed in length. The highest number of units in a repeat was 26 for bacterial functional amyloids, meanwhile for reference bacterial proteins 57. In general, the distributions of the number of units in a repeat differed between bacterial functional amyloids and reference bacterial proteins (Kolmogorov-Smirnov p-value=6e-6).

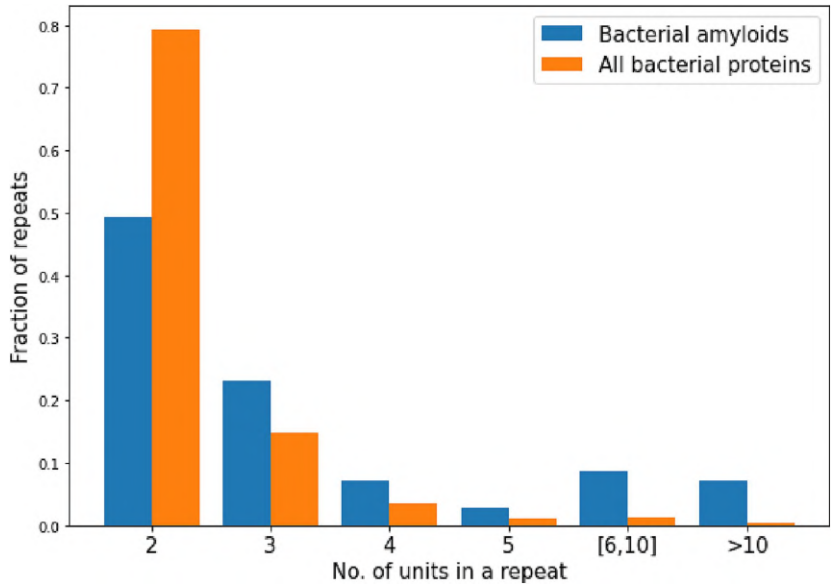

Figure 3.3: Histogram of number of units in repeats found in functional bacterial amyloids (blue bars) and all bacterial proteins (orange bars).

3.3.7 Similarity between units

Other amyloid studies

It has been shown that the similarity between units, as well as their number, affects the aggregation of the proteins. Titin protein is an interesting example in this case. The presence of more identical repeats in titin was shown to correlate with a greater aggregation propensity of the protein [219].

Bacterial functional amyloids

The similarity between the units in a repeat was different for bacterial functional amyloids that reference bacterial proteins (Fig, 3.4). Repeats in bacterial functional amyloids tended to be more imperfect than in the case of other proteins (Kolmogorov-Smirnov p-value=3e-4). The impact of units' similarity was further checked by Jakub Wojciechowski in the CsgA protein case study. CsgA contains five imperfect repeats R1, R2, R3, R4 and R5, out of which R1, R3 and R5 are aggregation-prone. The dimeric structures for three mutant models were predicted with Alpha Fold 2: 1) CsgA only with five repetitions of R1, 2) CsgA only with five repetitions of R3, and 3) CsgA only with five repetitions of R2. Models with only R4 and R5 had low quality and hence were discarded. The simulations with CABSflex software [220] enabled the investigation of the potential stability of CsgA fibrils with perfect repeats. Models with only R1 or R3 units were more stable and diverged less from the initial conformation than the original CsgA with imperfect repeats.

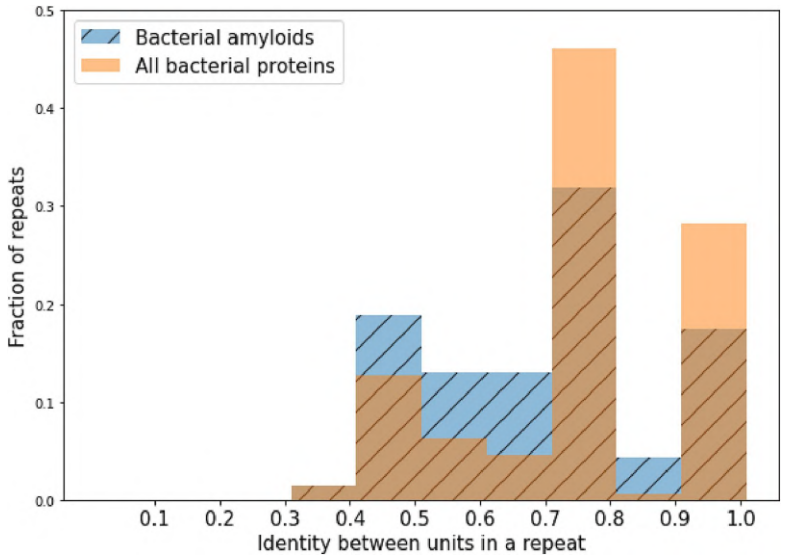

Figure 3.4: Histograms of identity between repeat units for functional bacterial amyloids (blue hatched bars) and all bacterial proteins (orange plain bars).

3.3.8 Aggregation-prone regions in the repeats

Other amyloid studies

In several studies, aggregation-prone regions have been found in protein repeats. Tau protein, although classified as pathological amyloid, contains aggregation-prone regions

within a microtubule-binding domain (MTBD). MTBD contains repeated sequence motifs that are the main aggregation-driving regions of the protein [221]. Also, the two hexapeptide motifs found on the R2 and R3 units in tau form amyloids. Examples of functional amyloids with the aggregation-prone region hidden in the repeat include: Pmel17 with 10 imperfect repeats which incorporate amyloid-prone regions, fungal HET-s with two imperfect repeats responsible for protein aggregation, and curli proteins CsgA and CsgB [222].

Bacterial functional amyloids

The identification of the aggregation-prone regions in the case of functional amyloids is problematic with the current methods. The available predictors are trained on pathological amyloids, which dominate the amyloid databases. Functional amyloids have distinct amino-acid profiles. This, in combination with their low representation in hot spot datasets, challenges the utility of the predictors in this case. Therefore, a more traditional approach, calculation of the amino acid profiles, was taken to interpret the molecular role of the found repeats.

Frequencies of amino acids f_{aa} for all found 69 repeats for the considered here bacterial functional amyloids (AR) were analysed. The results were contrasted with the frequencies determined for BS - entire bacterial protein sequences (not limited to the repeats), BR - bacterial repeats and AS - entire sequences of bacterial functional amyloids (see Table with bacterial functional amyloids). For each group, the average frequency of each of the amino acids f_{avg} in the repeats was calculated (Figure 3.5). The detailed amino acid profiles in the form of their distributions are presented in (Figure 3.6). For each amino acid, the Mann-Whitney U test with Bonferroni correction was applied to reveal if the difference was statistically significant. The cases where the mean frequency was statistically significantly different from frequencies obtained for BR, BS, and AS (p-value$\leqslant$0.05) are the most interesting. In this scenario, repeats in bacterial functional amyloids are distinct from the rest of the sequence and other bacterial proteins concerning the particular amino acid frequency.

The physicochemical interpretation of the differences in the amino acid profiles, which is discussed in this paragraph, was supported by Natalia Szulc. Firstly, it is easily noticeable that bacterial functional amyloids, as expected, are characterized by their unique frequency profile in comparison to BR, BS and AS. Repeats in bacterial functional amyloids (AR) are significantly enriched in threonine (T). Threonine, according to other works, when exposed to flat surfaces, forms a zipper interface that stabilizes the interaction [223]. This indicates that threonine abundance could mediate interactions in the repeats of bacterial functional amyloids, hence controlling the kinetics of amyloid formation. On the other hand, repeats in bacterial functional amyloids (AR) are depleted in methionine (M), histidine (H), arginine (R), cysteine (C), leucine (L), and alanine (A). The abundance of positively charged amino acids was found in pathological amyloids, where they exhibited a negative impact on amyloid toxicity, e.g. through interactions with cell membranes [119, 224, 225]. Their depletion in the repeats of bacterial functional amyloids could indicate the lower cytotoxic effect observed for the aggregation of functional amyloids [226]. The low abundance of alanine and leucine could explain the less hydrophobic character of the interactions between interfaces of beta-sheets in bacterial functional amyloids [227]. Similarly, the low prevalence of cysteine, which often stabilized amyloid fibrils via a disulfide bond, could be related to the reversible character of the functional amyloid structures [228]. The significantly higher content of asparagine (N)

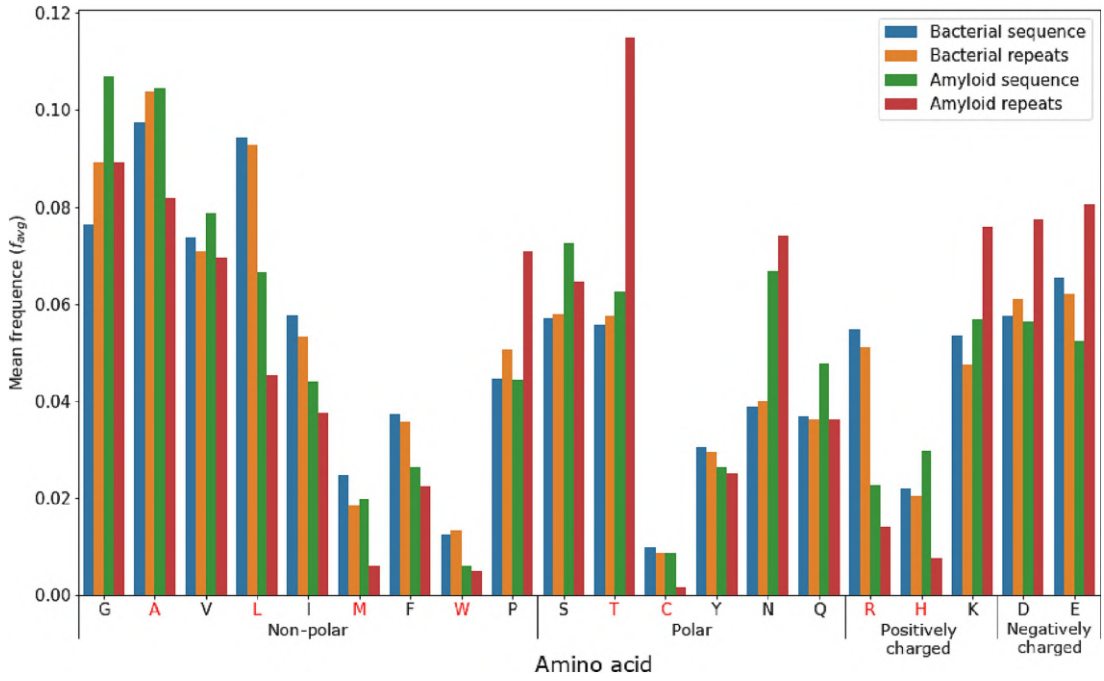

Figure 3.5: Mean amino acid frequency f_{avg} plot of entire bacterial sequences BS (blue bars), repeats in bacterial sequences BR (orange bars), entire amyloid sequences AS (green bars) and repeats in amyloids AR (red bars). For amino acids with red coloured labels, the p-values of the Mann-Whitney U test after Bonferonni correction (for amyloid repeats versus BS, BR and AS) were simultaneously below 0.05.

and aspartic acid (D) in AR than in the bacterial sequences BS and bacterial repeats BR could be a general characteristic of functional amyloids, not necessarily their repeats, keeping in line with previous studies [119]. A significantly higher content of glycine (G) in the repeats of the bacterial functional (AR) only in comparison to entire bacterial sequences (BS) could be related to their role as gatekeepers that mediate structural changes in the protein [229, 230, 231, 232].

3.3.9 Aggregation-prone regions outside of the repeat

Other amyloid studies

The aggregation-prone region may be inside or outside the repeats. Interestingly, even if the repeats are not amyloidogenic they still may affect the aggregation kinetics. An interesting example is α-synuclein. The aggregation-driving regions are present in the C-terminus, meanwhile, sequence repeats in the N-repeats. Although the repeats have a functional role not related to the aggregation, they still influence the aggregation propensity of the entire protein [214]. The appearance of the mutations in the repeats affects protein conformation and hence its fibrillation [233]. The same has been found for the prion protein PrP, which contains the aggregation-prone region in the N-termina [234]. Although the repeats in PrP have a copper-binding function and are not aggregation-driving, they still affect the amyloid formation. These observations are not limited to pathological amyloids. 5 imperfect repeats in yeast prion sup35, which are not amyloidogenic, modulate the fragmentation efficiency of the fibril [207].

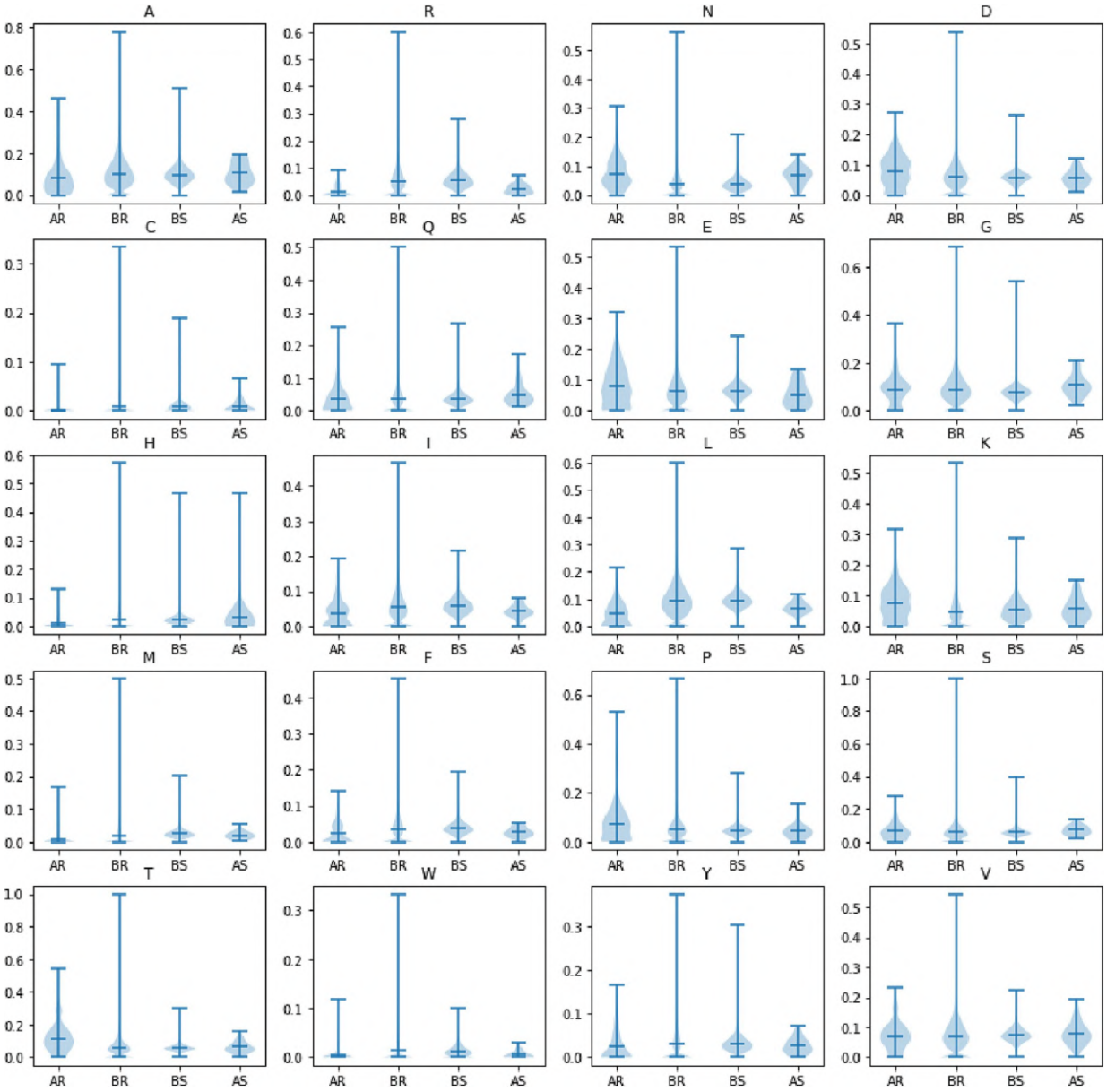

Figure 3.6: Distributions of amino acid frequencies (f_{aa}, y-axis) presented with violin plot. AR - repeats in bacterial functional amyloids, BS - entire bacterial sequences, BR - repeats in bacterial sequences), AS - entire amyloid sequences.

Bacterial functional amyloids

To further investigate the potential role of the repeats in bacterial functional amyloids, the secondary structures of entire proteins were predicted with PSIPRED [188]. Each repeat unit was then classified as either "Beta", "Coil", or "Helix". The frequencies of appearance of each of the secondary structure classes were calculated (Figure 3.7). 6 proteins were excluded from this analysis due to their length limitations.

The repeat units from HpaG, Microcin, CarD, FadA, ChpF and ChpH had no "Beta" class predictions suggesting that repeats in these proteins do not resemble beta-sheet structures. The general prevalences of "Beta" and "Helix" classes in all considered bacterial functional amyloids were not statistically significantly different for the repeats in bacterial

functional amyloids (the p-value of a two-sided Kolomogrov-Smirnov was 0.7). However, the appearance of the "Coil" class was statistically significantly more frequent (the p-value of a two-sided Kolmogorov-Smirnov was 4.5e-6 for comparison with "Beta" and 1e-4 for comparison with "Helix"). The median fraction of amino acids with "Coil" was equal to 0.5; meanwhile for "Beta", it was 0.2 and 0.26 for "Helix".

The potential abundance of coiled structures in the repeat units may be related to their appearance as disordered fragments in the proteins. Such fragments, which are often present in tandem repeats, may ease the fibril formation and facilitate amyloid interactions. Hence, even if the repeats would be not aggregation-driving in an amyloid protein, they could still be aggregation regulators.

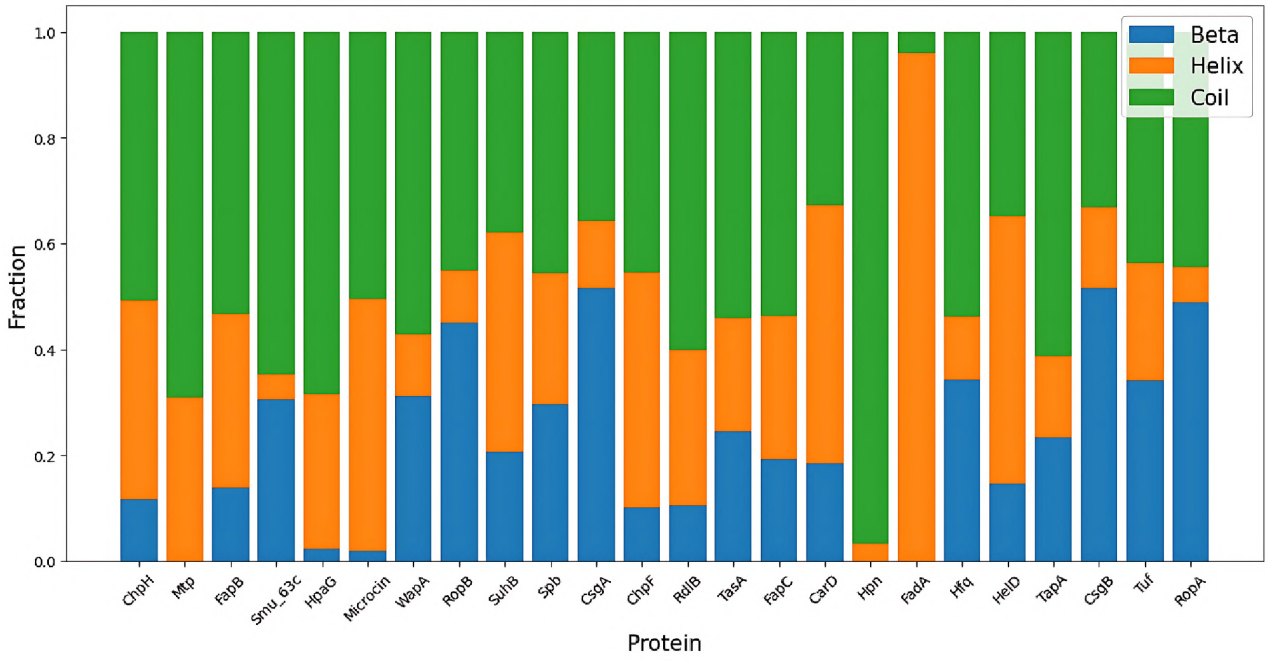

Figure 3.7: Prediction of secondary structure for the repeats in bacterial functional amyloids (AR). Blue bars correspond to 'Beta' class, orange to 'Helix' and green to 'Coil'.

3.4 Discussion

Our knowledge about bacterial functional amyloids is limited. Little such sequences are known, and even fewer structures are resolved. The state-of-the-art prediction tools for identification-prone regions that drive aggregation have limited utility for functional amyloids, which are underrepresented in the hot spots databases. This scarcity of information calls for more theoretical research on these proteins that could give us novel clues about their characteristics. This chapter focuses on the sequence analysis of the bacterial functional amyloids hypothesizing that such study could provide novel insights into their aggregation and interaction mechanisms.

The literature search highlighted the scarcity of known examples of these proteins, as only 38 instances were found. Despite the limited data size, the found sequences differed in size and function in the cell. They form a separatable group from other bacterial proteins, probably due to the frequent biofilm functions. As bacterial functional amyloids are designed by evolution to aggregate and evolution favors repetition, this study focused on a better understanding of the role of the repeats in these proteins. The modularity of the amyloid fibrils inspired the search for regularity in the sequence since the intermolecular interactions often facilitate the intramolecular ones.

The detailed analysis of the role of tandem repeats in bacterial functional amyloids was based on the three pillars. To spark ideas about the influence of the repeats on the aggregation process of bacterial functional amyloids, examples from the literature were discussed. Although these studies often focus on pathological amyloids, the presence of the repeats in such proteins, could give a hint on their impact on fibril formation. Then, the calculation of repeat characteristics was performed for bacterial functional amyloids. Finally, the results were compared to the ones found for all bacterial proteins and their repeats to assess the statistical significance of the found features.

The appearance of the repeats in the bacterial functional amyloids does not seem to be more frequent than in the case of other proteins. Nevertheless, we should keep in mind that still few examples of such proteins are known and this tendency could change as more data is gathered. However, when they do appear they have their unique character.

The repeats in bacterial functional amyloids consist of a medium number of units that are not a perfect repetition of each other. In contrast to all bacterial proteins, they have an overrepresentation of repeats with three to six units and a statistically significant lower similarity between repeat units. Their amino acid profile is quite specific, pointing towards flexibility, reversibility, and regulatory role of these regions in fibril formation, which is further confirmed by the prevalence of the predicted coil conformation. The results bring to light a purposeful design of the repeats by the evolution potentially targeted to regulate the aggregation of the protein.

The addition or deletion of units in a repeat has been previously shown to affect aggregation rates [210, 218]. The specific enrichment in a certain number of units in the repeats in bacterial functional amyloids demonstrates the evolutionary pressed best design that balances the aggregation. Further evidence supporting this hypothesis may be found in the lower similarity between units. More similar units are rather correlated with higher stability and toxicity of the fibrils, according to the presented work. The observed imperfection could also be purposeful and may facilitate the aggregation properties of the bacterial functional amyloids, which are not as rigid as pathological amyloids. On the one hand, the presence of multiple similar sequences, as within a tandem repeat, could give room for multiple symmetrical interactions broadening the number of control mechanisms

in the aggregation process. On the other hand, their lower similarity weakens potential amyloid interactions, which are rather the strongest in the case of identical sequences. Hence, the presence of imperfect repeats could give rise to a sophisticated machinery that stands behind the controlled aggregation of the bacterial functional amyloids.

The observed lower toxicity of the bacterial functional amyloids may be related to their observed amino acid profile. In pathological amyloids, the positively charged amino acids were found to contribute to the toxicity of the fibrils [119, 224, 225]. The repeats in bacterial functional amyloids are depleted in these residues. In addition, the lower abundance of hydrophobic amino acids, such as leucine and alanine, which participate in hydrophobic interactions that stabilize structures, could explain the lower stability of functional fibrils. The flexibility may also be the result of the prevalence of coiled conformations of the repeats observed from the secondary structure predictions. Such regions are often related to intrinsically disordered fragments, which could further regulate the aggregation.

The presented results do not point towards the abundance of aggregation-prone regions in the repeats. Rather, they suggest that such repeats do regulate the aggregation no matter their exact location in the sequence. To aggregate in a controlled manner, bacterial functional amyloids must contain complex molecular machinery that controls the fibril formation and dissasembly. The observed characteristics of tandem repeats in these proteins make them perfect candidates for this role.

This chapter uncovers an interesting feature of the sequence of bacterial functional amyloids related to their aggregation propensity. This may have an impact on how these proteins aggregate and interact with other proteins, including human ones, also neurodegeneration. This proves the second hypothesis of this thesis that *sequence tandem repeats influence the aggregation of bacterial functional amyloids* and makes the second goal *to examine the role of tandem repeats in bacterial functional amyloids* realized.

Sequence analysis enabled discovery of novel characteristics of bacterial functional amyloids. However, to obtain the full picture of the molecular mechanisms standing behind bacterial function amyloids aggregation and interactions with other proteins, knowledge of their structure would be very enriching. Structural information could allow for a more detailed analysis of molecular interactions between bacterial functional amyloids and human proteins, e. g. via molecular docking, leading to an in-depth discussion on their role in neurodegeneration. Hence, the next chapter covers the usage of state-of-the-art software for protein structure prediction, AlphaFold, in the context of bacterial functional amyloids and other amyloid proteins.

Results presented in this chapter were published as: ***Nowakowska, A. W.****, Wojciechowski, J. W., Szulc, N., & Kotulska, M. (2023). The role of tandem repeats in bacterial functional amyloids. Journal of Structural Biology, 215(3), 108002.*

Structure prediction of amyloid proteins with AlphaFold 3

4.1 Introduction

In 2024, with the Nobel Prize in Chemistry award, all eyes were on bioinformatics. The decades-long protein structure prediction problem has been solved, marking a new era in structural biology.

AlphaFold (AF) does not need to be introduced to anyone working in molecular biology [235]. A massive neural network model, with more than 30,000 citations in less than 3 years, can predict a protein structure in minutes with remarkable accuracy. The newest version of the algorithm, AF3, was released in 2024 [236]. The method promises to account for the environmental context, such as the presence of ligands or DNA, marking another important step forward. Both versions of the algorithm use Multiple Sequence Alignment (MSA) information and PDB (RCSB Protein Data Bank) templates, though the MSA information alone is sufficient for good predictions. The importance of MSA is lower in AF3. AF3 also has a generative character that leads to multiple different outputs, but at the price of possible hallucinations.

AF algorithms seem promising for anyone working with proteins with unsolved structures, such as bacterial functional amyloids. Amyloids, in general, pose multiple experimental challenges that limit the availability of structural data. Fibrils often have high molecular weight, dissolution follows complicated protocols, and in one sample, polymorphic species may appear. These factors exploit the usage of X-ray, NMR, and cryo-EM techniques for the determination of the amyloid structure [237, 238]. In consequence, experimental data on amyloids rise slowly, and computational approaches, such as AF, are welcome.

Modelling amyloid proteins comes with several challenges. As already mentioned, few experimental data is available, and resolved structures exist mainly for fragments of well-studied pathological amyloids. As a consequence, teaching a model to predict an amyloid structure is difficult, although an attempt to do so has been made. An example is RibbonFold, released in 2025 [139]. RibbonFold is based on the architecture of AlphaFold-Multimer. To predict an amyloid structure, the authors encoded constraints in the template module that impose fibrillar appearance. The model was then fine-tuned on a dataset of amyloid structures. RibbonFold is the first software that is designed to predict amyloid structure, and it accounts for the structural polymorphism of amyloid fibrils. Nevertheless, it has limited accuracy. By the assumption, RibbonFold is insensitive to the aggregation propensity of a sequence, and for any input, it always predicts the amyloid fibril. It is trained on a relatively small dataset of amyloid structures biased by scientific interests. The constraints incorporated into the model will also limit the creativity of the algorithm which may be needed when modelling functional amyloids, which are often different from

the pathological ones. RibbonFold does not allow for modelling different sizes of a fibril and accounting for other molecular entities, e.g. ions. The perfect software for amyloid structure prediction should be sensitive not only to the structural polymorphism of amyloid fibrils but also allow for modelling monomers and multimers. It should also take into account that many amyloids are easily affected by sequence modifications and environmental conditions. For example, although the Aβ sequences in other organisms are nearly identical, not all are aggregation-prone [239]. However, homologs of functional amyloids, e.g. CsgA, often have a conserved aggregation propensity [240]. Finally, the ideal algorithm should be able to distinguish which proteins, and not only fragments, are likely to fold into amyloid fibrils and which are not, allowing for structural proteome-wide screens of aggregation propensity and fibril structures. So far, no tool has been shown to do that. Amyloids are a difficult case that requires a structure prediction algorithm to have a high level of generalization and low training data memorization, which makes them a perfect testing dataset.

AF2 failed to correctly predict amyloid structures [241]. However, the significant changes in architecture that led to the release of AF3 give hope for performance improvement. AF3 is also the only software that attempts to generalise the structure prediction problem, allowing for the modelling of monomers, multimers, DNAs, RNAs and the addition of ions within a single tool, potentially enabling the complex amyloid modelling.

In this chapter, a third hypothesis that ***structures of amyloid fibrils can be predicted with AF3*** is investigated. Because the dataset of bacterial functional amyloids is small, the analysis is not limited to them and instead, the AF3 performance on all amyloid proteins is assessed. Monomeric and multimeric forms of amyloid proteins are predicted. To explain how models are built for amyloid proteins by AF3, the results are compared with structures available in the PDB and with predictions for the entire human proteome.

4.2 Methods

Datasets: *Positive* control

AF3 performance on the amyloid protein structure prediction problem was evaluated first for amyloid proteins without experimentally solved structures. To build such a database, amyloid proteins were extracted from: Amypro database [191], the dataset of bacterial functional amyloids BFA, introduced in the previous chapter, and AmyLoad [242], which has fragments of aggregation-prone peptides (AmyLoad also has non-aggregating sequences, see *Negative* control below). All sequences of alpha phenol-soluble modulins were not considered, as they form amyloid fibrils with a cross-alpha architecture [179]. Seven proteins, namely Aap, Bap, Esp, PAc, SasG, YghJ, and Agglutinin-like protein 3, were too long for multimeric predictions with AF3 and hence were removed. Sequences with lengths below 10 amino acids were removed. Proteins, which had the solved structure for (almost) the entire sequence (structure covered more than 80% of the sequence), with reference to the STAMP database that contains amyloid proteins and their structures [243], were excluded from the analysis. Many of the peptides from the AmyLoad database were short and constituted part of the longer proteins or mutants of such; however, it was impossible to assess whether their structure was solved. As the inclusion of the AmyLoad database in the study allowed for accounting for negative examples (see below), and the

chances that the majority of these sequences have solved structures is low, this problem was considered negligible. Finally, all sequences were clustered at a 90% level of identity with cd-hit default parameters [244]. Homologs of the *Amyloid_structure* dataset (see below) were identified with mmseqs easy-search with default parameters (minimum identity threshold of 30%) [245], and sequences appearing as hits were removed. This made the final version of the *Positive* control dataset.

Datasets: *Negative* control

Sequences confirmed not to aggregate were extracted from the AmyLoad. Peptides with lengths below 10 amino acids were removed. Homologs of the *Amyloid_structure* dataset in the *Positive-Control* dataset were identified with mmseqs easy-search with default parameters, and sequences appearing as hits were removed. This made the final version of the *Negative* control dataset.

Datasets: *Amyloid_structure*

Proteins with solved amyloid structure for (almost) the entire sequence (structure covered more than 80% of the sequence) considered were: Aβ-42, α-synuclein, glucagon, IAPP (human amylin), transthyretin, and immunoglobulin (see Table 4.1).

AF models

One AF3 structure with six protein copies and one monomer were predicted for each sequence from the *Positive* control and *Negative* control datasets. For each protein from *Amyloid_structure* dataset, 50 monomeric and 50 multimeric models of six protein copies were predicted. Additionally, for Aβ-42, 50 models for each number of protein copies from 1 to 9 were predicted. For 8 proteins (Sauvagin, Viral protease VP4, Zona pellucida sperm-binding protein 1, TasA, PMEL17, Semenogelin, Lung surfactant, p53) one model for each number of protein copies from 1 to 9 was predicted. All models were predicted with AF3 webserver available at https://AFserver.com/.

The datasets of *Positive* and *Negative* control are relatively small compared, e.g. to AF Database. Therefore, to better understand how AF predicts proteins structures, models for *Homo Sapiens* proteome were used as a reference. The monomeric structures predicted with AF2, *HSMonomers*, were downloaded from the AF database available at https://alphafold.ebi.ac.uk/download [141]. The dimeric structures for random pairs of human proteins predicted with AF-Multimer, *HSDimers*, were downloaded from the website https://predictomes.org/ [246]. At the time of this work, no larger AF3 prediction dataset was available. AF models were visualized with pymol [247].

Protein	Sequence	PDB*
Aβ-42	DAEFRHDSGYEVHHQKLVF FAEDVGSNKGAIIGLMVGG VVIA	2mxu
α-synuclein	MDVFMKGLSKAKEGVVAAA EKTKQGVAEAAGKTKEGVL YVGSKTKEGVVHGVATVAE KTKEQVTNVGGAVVTGVTA VAQKTVEGAGSIAAATGFV KKDQLGKNEEGAPQEGILE DMPVDPDNEAYEMPSEEGY QDYEPEA	2n0a
IAPP	KCNTATCATQRLANFLVHS SNNFGAILSSTNVGSNTY VVIA	6vw2
Glucagon	HSQGTFTSDYSKYLDSRRAQ DFVQWLMNT	6nzn
Immunoglobulin lambda variable 3-19 light chain	AVSVALGQTVRITCQGDSL RSYSASWYQQKPGQAPVLV IFRRFSGSSSGNTASLTIT GAQAEDEADYYCNSRDSSA NHQVFGGGTKLTV	6z1i
Transthyretin	GPTGTGESKCPLMVKVLDA VRGSPAINVAMHVFRKAAD DTWEPFASGKTSESGELHG LTTEEEFVEGIYKVEIDTK SYWKALGISPFHEHAEVVF TANDSGPRRYTIAALLSPY SYSTTAVVTNPKE	6sdz

Table 4.1: Amyloid_structure dataset. *Example of a PDB identifier of an amyloid structure that was part of the AF training dataset.

Quality of AF models

AF model quality can be described by [235]:

- pLDDT - per-residue predicted local distance difference test, takes values from 0 to 100
- pTM - describes the quality of the complex prediction, takes values from 0 to 1
- ipTM - describes the quality of the predicted relative positions of the subunits, takes values from 0 to 1.

In the case of *HSMonomers* only pLDDT scores were available.

Similarity between structures

Similarity between structures was assessed with the TM-score that takes a value from the interval (0,1]. The higher the TM-score, the more similar the structures are. Structures with TM-score>0.5 are considered to belong to the same CATH category [248].

Structure classification

There is no reliable tool for assessing whether the structure is amyloid. Therefore, each model was manually curated as resembling an amyloid fibril, similar to structures deposited in the STAMP database, or not.

Similarity of monomers, multimers, and PDBs

Similarity between monomeric and multimeric models for *Positive* control and *Negative* control sequences and *HSMonomers* and *HSDimers* was assessed with the Foldseek easy-search command (with default parameters) [249].

Foldseek compares structures to detect even distant homology relationships. Foldseek hits can be described by TM-score and homolog probability, which stands for "the probability for each match to be homologous, based on a fit of true and false matches on SCOPe" (direct quote from van Kempen, 2024). Hits discovered by Foldseek on AlphaFold models often relate to hits found by traditional sequence-based approaches like BLAST [250].

The most similar structure for each monomer and multimer in the PDB (version of March 2025) was identified with Foldseek easy-search command (with default parameters). Some of the structures in this PDB version have not been seen by AF and could not bias the predictions. However, they constitute only a minority and should not impede the observations of certain trends between AF training examples and models.

Clusterization of monomeric models predicted for each protein from the *Amyloid_structure* dataset was performed with Foldseek easy-cluster (with default parameters). Similarity between monomeric cluster representatives and 50 multimers was assessed with Foldseek easy-search (with default parameters).

Foldseek returns TM-scores and *Homolog probability*. Structural alignments were visualized with Foldseek webserver available at: https://search.foldseek.com/search.

Data availability

The AF models for *Positive* control, *Negative* control and *Amyloid_structure* datasets are available at the zenodo repository of the project: **https://zenodo.org/records/15576017**. The repository contains a metadata file, which contains sequences of the proteins, their description, human classification if a model resembles an amyloid structure, identifiers of PDB files that were similar to the AF model, and metrics of the structural alignment.

4.3 Results

The performance of AF3 in predicting the amyloid structure was evaluated on the three datasets described above: *Positive* control, *Negative* control and *Amyloid_structure*.

4.3.1 The size of the predicted fibril n influences model quality

The number of protein copies (n) is a critical AF3 parameter that defines the multimericity of the complex, and in this case, the size of the predicted fibril. Given that all predictions were made with the webserver, it was impractical to perform a comprehensive screen for optimal n for each amyloid protein. Therefore, to decide on the best value n for all amyloids in this study, AF3 was benchmarked on the well-known example of an amyloid, Aβ-42. For each number of protein copies n, from 1 to 9, 50 structures of Aβ-42 were predicted with the AF3 webserver.

The monomeric models of Aβ-42 adopted helical conformations. The dimers were either beta-sheet or helical. Higher values of protein copies n consistently yielded fibrillar architectures (Fig. 4.1A). The quality of fibril predictions increased with the number of protein copies n, peaking at a value of 5, and declined for higher multimericities (Fig. 4.1 B).

To ensure that choosing any protein copy number above 2 consistently results in the same classification with respect to the fibrillar nature of the model, a quick benchmark was conducted on other amyloid proteins. Eight random amyloid proteins were chosen (sauvagin, viral protease VP4, zona pellucida sperm-binding protein 1, TasA, PMEL17, semenogelin, lung surfactant, p53). For these examples, multimers with n values in the range from 1 to 9 were predicted. For seven proteins, classifying whether the AF3 model looks fibrillar or not was consistent regardless of the choice n. The only exception was semenogelin. In this case, dimers had a fibrillar architecture; meanwhile, higher n values led to irregular structures that were hard to classify as fibrillar due to asymmetry. Examples of results are presented in Fig. 4.2. It can be concluded that the change in the number of protein copies n above 2 should not significantly bias whether the amyloid structure is predicted or not by AF3, though it influences the model quality.

To balance between the longest fibril length possible and the model quality, 6 protein copies were considered as a reasonable multimericity for amyloid structure prediction with AF3 and used in further sections.

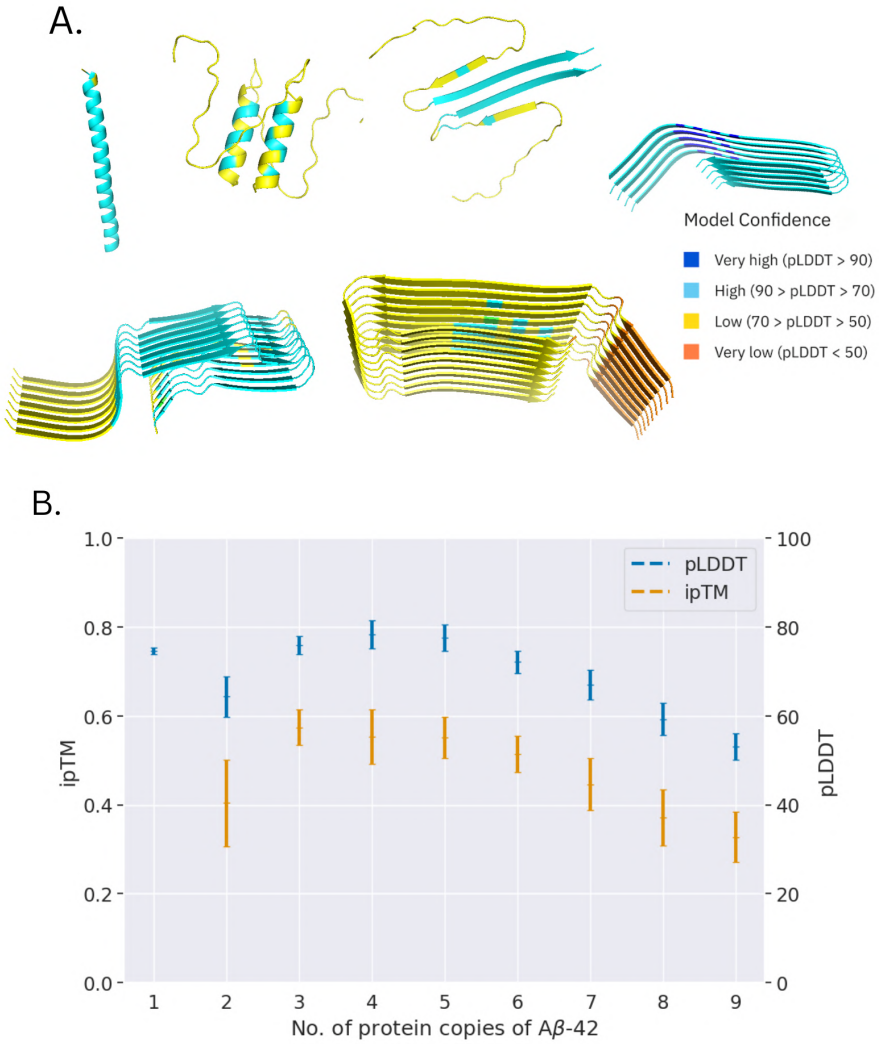

Figure 4.1: Impact of the number of protein copies n. A. Examples of AF3 models of Aβ-42 colored by pLDDT metric. From top to bottom and left to right: helical monomer, helical dimer, beta-sheet dimer, and fibrillar structures for $n = 5$, $n = 7$ and n=9 protein copies, respectively. B. pLDDT (blue) and iPTM (orange) as a function of the number of protein copies used.

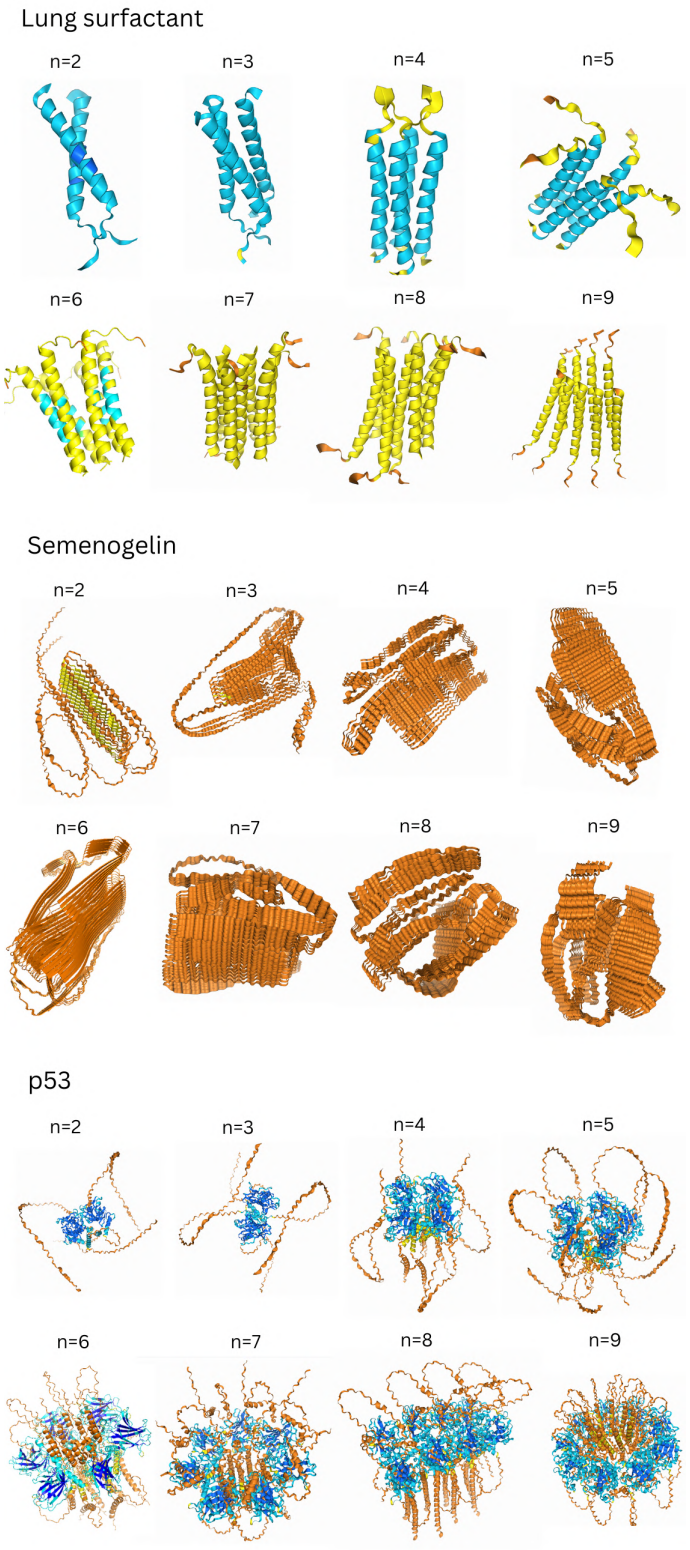

Figure 4.2: AlphaFold 3 predictions for different number of protein copies n for lung surfactant, semenogelin and p53.

4.3.2 AlphaFold 3 struggles to predict amyloid fibril

One multimeric structure with 6 protein copies for each protein *Positive* control and *Negative* control datasets were predicted. Each AF3 model was visually inspected, with the help of Jakub Wojciechowski, to determine if it resembles an amyloid structure.

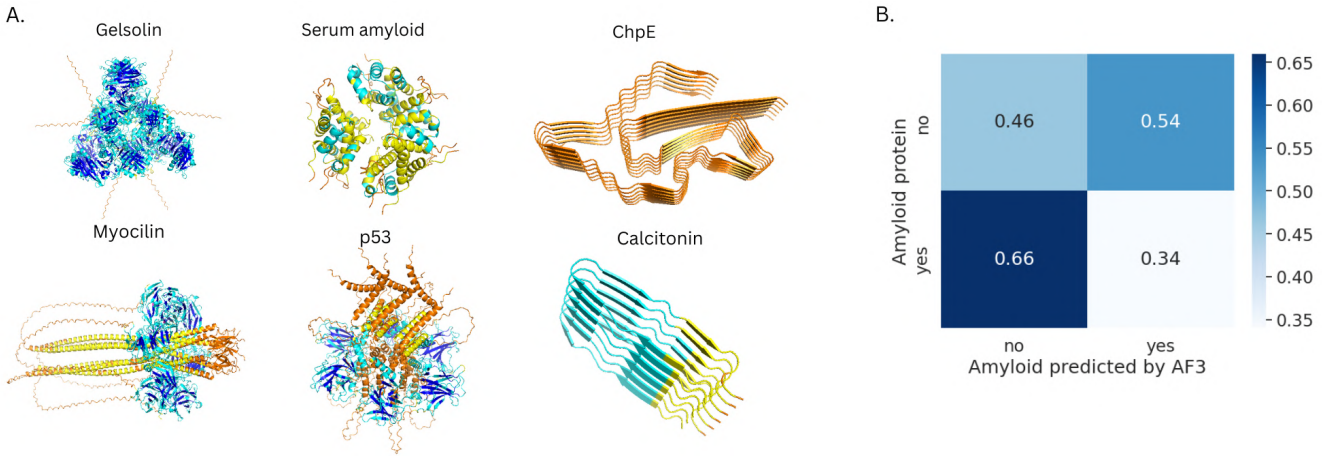

Figure 4.3: AF3 predictions of unknown amyloid structures. A. Examples of fibrillar and non-fibrillar models of amyloid proteins. From left to right, from top to bottom: Gelsolin globular model, helical model of Serum amyloid, amyloid fibrillar model of ChpE, non-amyloid model of Myocilin, non-amyloid model of p53, amyloid fibrillar model of Calcitonin. B. Confusion matrix of AF3 performance in amyloid structure prediction.

AF3 predicted highly diverse structures for amyloid proteins (Fig. 4.3A). Symmetrical, helical and fibrillar models appeared. The general performance in predicting the amyloid structure was poor (Fig. 4.3B). Only 34% of the amyloid protein predictions resembled fibrils (true positive). Similarly, only 45% non-aggregating peptides looked dissimilar to an amyloid fibril (true negative). The classification scores were the following: Accuracy=0.37, Precision=0.63, true positive rate (Sensitivity)=0.34, and F1-score=0.44. Furthermore, sequences with incorrect predictions were more likely to have higher pLDDT values (Mann-Whitney test p-value=0.02), but not pTM or ipTM, potentially misleading the user.

No statistically significant difference was found in AF3 performance for functional and pathological amyloids (Fisher's exact test p-value=0.42).

AF3 performed better for shorter fragments than for entire proteins. For sequences with a length below 36, which is the maximum length of the non-aggregating sequences, the true positive rate increased to 45% (Fig. 4.4A). The true positive and true negative cases had significantly lower sequence length values (Mann-Whitney test p-value = 3e-5, Fig. 4.4B).

4.3.3 Comparison between multimeric and monomeric predictions

Amyloid proteins are demanding cases for machine learning approaches due to their polymorphism and frequent differences between monomeric and multimeric forms. To evaluate if AF struggles with amyloid proteins because it cannot fully capture the difference between fibrillar structures and monomers, for each amyloid protein from the

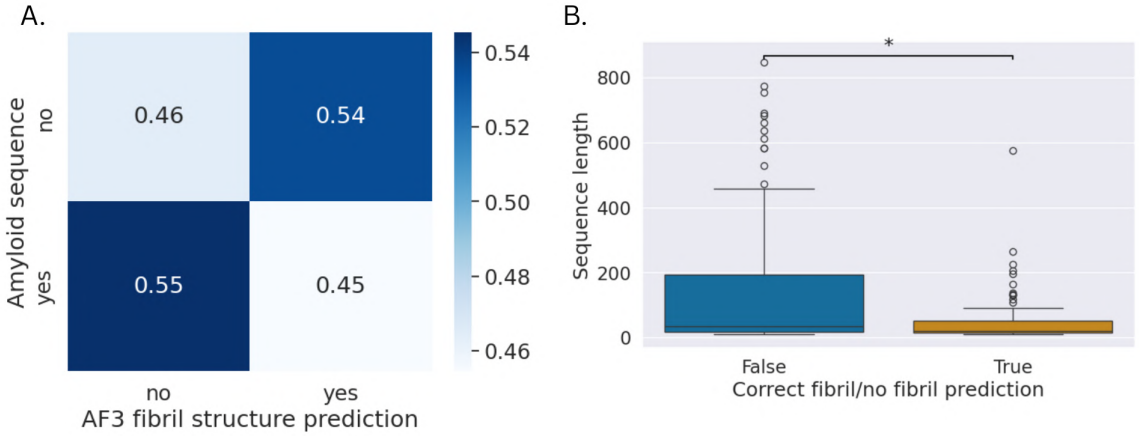

Figure 4.4: Influence of the sequence length. A. Confusion matrix of AF3 performance in amyloid structure prediction for sequences with a length below 36 amino acids. B. Distribution of sequence lengths for false positives and false negatives (blue) and true positives and true negatives (orange).

Positive dataset, the monomeric model was also predicted. Monomeric and multimeric models were then aligned with Foldseek.

In 43% of the cases, the monomer and multimer models of amyloid proteins shared significant structural similarities with a TM-score>0.5. For these proteins, the multimeric model rarely resembled an amyloid fibril (Fig. 4.5A), it was usually of higher quality, though (difference of 15 points in pLDDT, statistically significant Mann-Whitney p-value=9e-10). The similarity between monomer and multimer was more frequent in the case of longer proteins than in shorter ones, in line with previous observations on the performance of AF3 (difference in the sequence length was statistically significant with the Mann-Whitney test p-value = 2e-12).

HSDimers and *HSMonomers* were compared to evaluate how AF generally deals with multimeric models. For each chain in a dimer from the *HSDimers* dataset, corresponding monomeric models were identified in HSMonomers and compared. In 80% of the cases (group I), both chains of the dimer shared similarity with their corresponding monomers (see the TM-score distribution in Fig. 4.5B). 41% of dimers in group I, and 34% of all dimers in *HSDimers*, had TM-scores for both chains above 0.5. Cases with high TM-score (>0.8) for one of the chains and the low TM-score (<0.3) for the other were present and constituted 7% of the cases in group I. Group I models had a mean pLDDT of 62, standard deviation=14. For 8227 dimers (19%, group II), for one of the chains, no similarity with the corresponding monomer was found. Group II models were of low quality, with the mean pLDDT of 56, standard deviation=15. 99.99% of models in group I and II had the probability of structural homology, analogous to belonging to the same SCOPE family, above 0.9. For only 637 dimers (1%, group III), no similar monomer was identified for any of the chains. Group III models were of very low quality, with the mean pLDDT of

47, standard deviation=17. The difference in pLDDT scores between group I, and merged group II and III was statistically significant (Mann-Whitney p-value=0.0).

This shows that AlphaFold can do both: predict dimeric models similar to their monomers and different to them. However, when the monomers and dimers have different conformations, the pLDDT scores of a dimer are low. This stays in line with how frequently the amyloid structure was modelled for amyloid proteins and explains why the scores for amyloid fibrils were low.

4.3.4 Similarity of AF models to structures deposited in the PDB

Possible bias that affects the AF3 results may result from its training data, which is rich in globular examples and contains few amyloid structures, particularly for longer proteins. To evaluate whether the training data hinders amyloid structure prediction, for each multimeric model, a similar structure was searched for in the PDB with Foldseek.

In total, 32756 hits were found for 84 models of amyloid proteins from the *Positive* dataset and for 7 models of non-aggregating peptides from the *Negative* dataset. The number of models that resembled a PDB structure for the *Negative* dataset was small and, therefore, ignored in the further analysis.

Amyloid proteins whose AF3 predictions resembled PDB structures were significantly longer in sequence than amyloid proteins whose models were not similar to PDB structures (Mann-Whitney test p-value=2e-12). The lack of similarity between the model for an amyloid protein and the PDB structures was a rare event, as of the 61 sequences with a length greater than 100 amino acids, 7 of them did not resemble any structure of the PDB. Moreover, models matching PDB structures were more frequently predicted as non-fibrillar (Fisher's exact p-value=2e-10, Fig. 4.6), though they had statistically higher pLDDT values (Mann-Whitney test p-value=0.01).

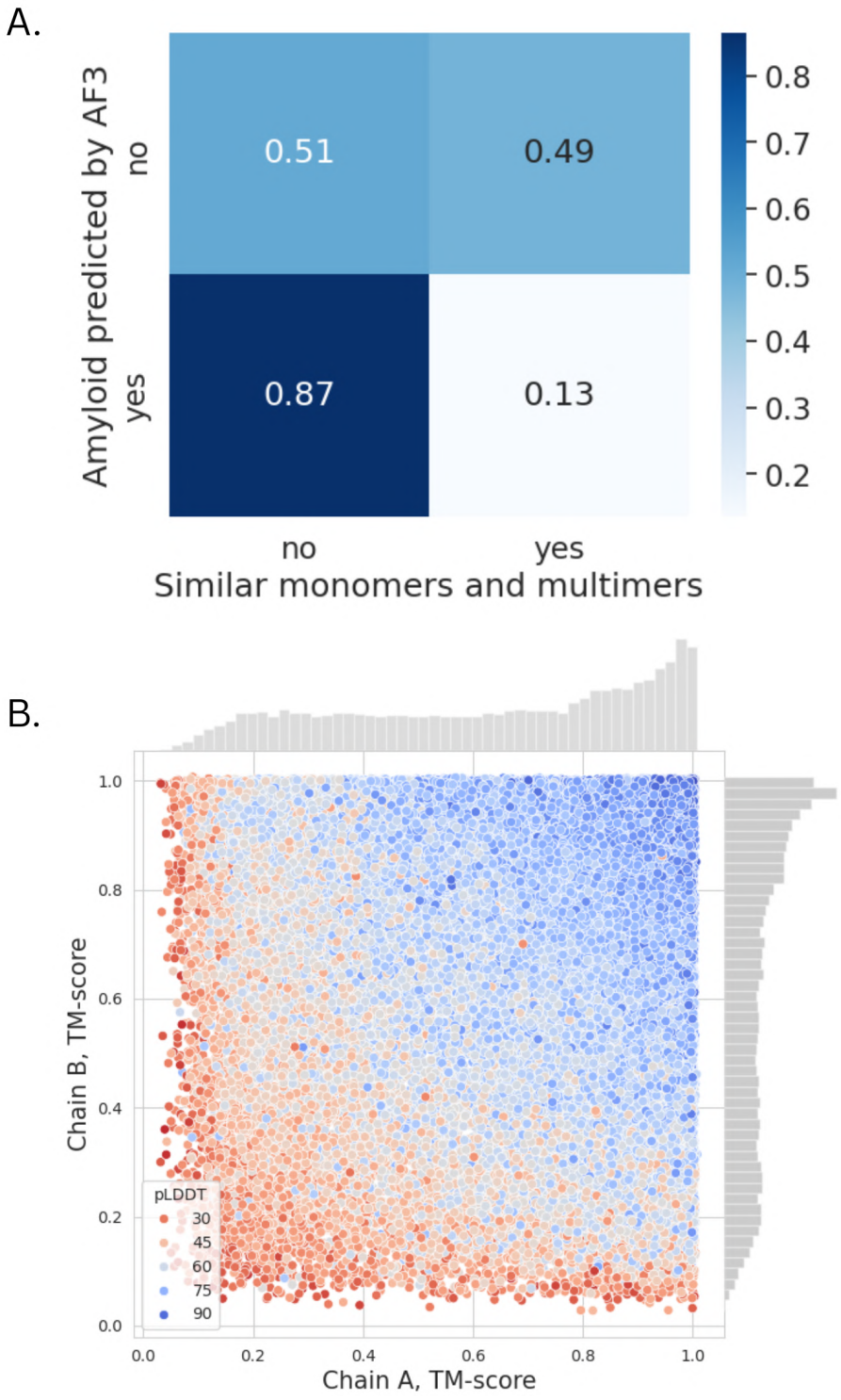

Figure 4.5: Simialrity between monomeric and multimeric models. A. Confusion matrix. B. Distribution of TM-score in group I, models are colored by pLDDT.

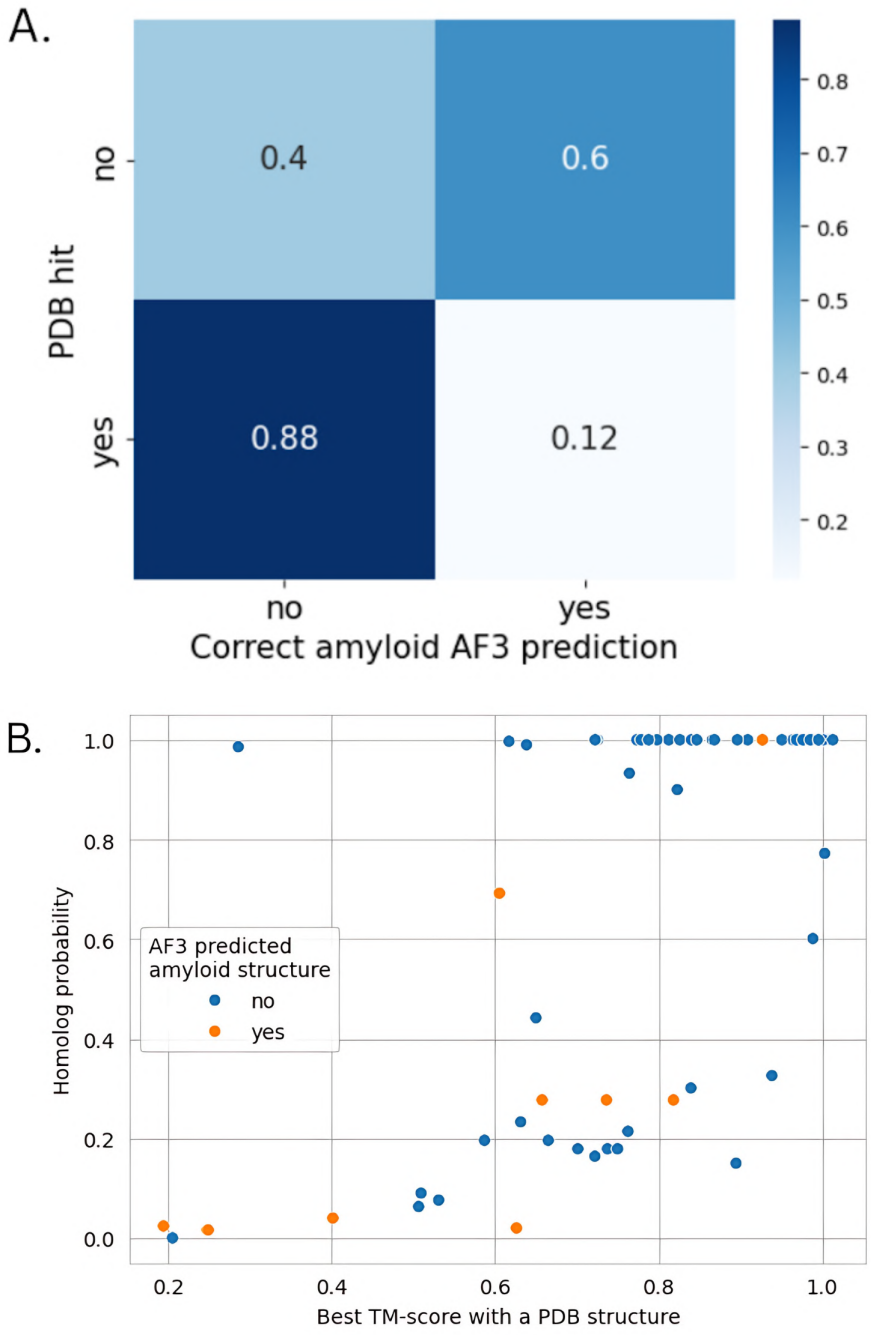

Figure 4.6: Similarity between AF3 models for amyloid proteins and PDB structures. A. Confusion matrix for amyloid proteins from the *Positive* control dataset, classification of the models with respect to the similarity to the PDB structures. B. Search for similar PDB structures to the AF3 models of amyloid proteins. The homolog probability is presented as a function of the TM-score between the AF3 model and the most similar PDB file.

The homology information in the case of amyloid proteins could be a misleading factor for AF3. Therefore, for each amyloid protein whose model resembled PDB structures, the hit with the highest TM-score was extracted along with the probability that the two proteins are homologous (Fig. 4.6). Amyloid proteins, which resembled the solved structures deposited in the PDB and were homologous, tended not to resemble amyloid fibrils, though again, they had much higher pLDDT. Specifically, the mean pLDDT of a multimeric model for an amyloid protein with a homolog probability above 0.95 was 37, meanwhile, without the homolog, it was 57; the difference was statistically significant with the Mann-Whitney test p-value=1e-4. The only protein that was modelled as a fibril and resembled the PDB structure was CsgB, the homolog of CsgA, which has a solved structure. Importantly, CsgA was probably not seen by AF3 during the training, hence, AF3 modelled CsgB well without previous bias. In Fig. 4.7 the structural alignments between globular multimeric models for amyloid proteins and their matches in the PDB are shown.

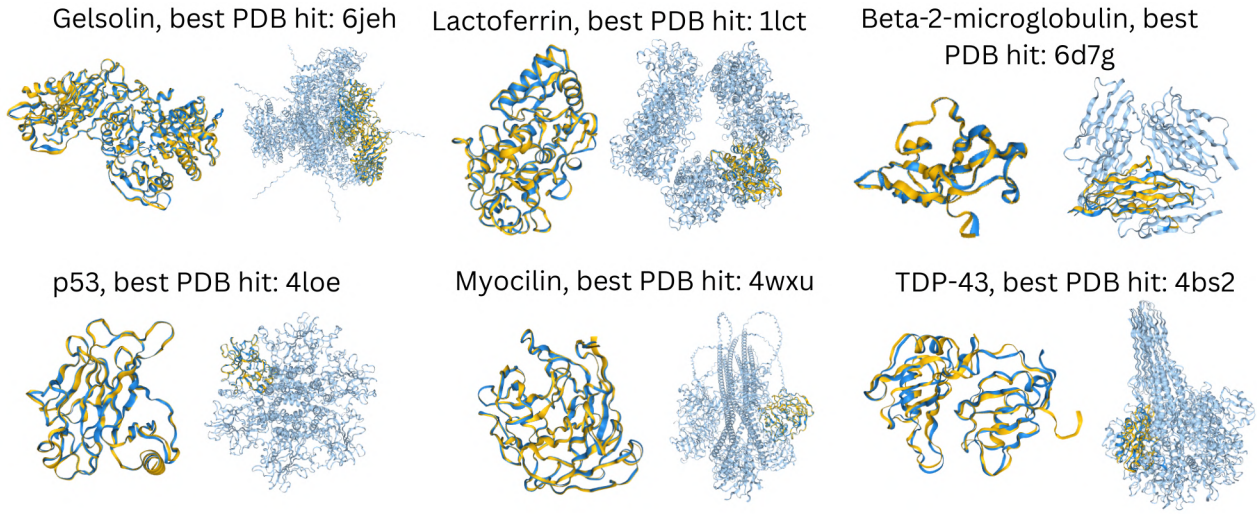

Figure 4.7: Similarity between AF3 models for amyloid proteins and PDB structures in the form of structural alignments.

To ensure that AF generally predicts models that resemble PDB structures, which may be problematic in modelling amyloid proteins, as few amyloid structures are available, monomeric predictions for the human proteome were compared with PDB structures. Specifically, for each model from *HSMonomers*, similar structures in the PDB were found with Foldseek, and the hit with the highest TM-score was extracted (Fig. 4.8 and 4.9). 92% of monomeric models resembled PDB examples, and in the vast majority of matches, proteins were likely homologous (Homolog probability above 0.9). The structural similarity was also significant, as 85% of the hits had a TM-score above 0.8. AF predictions with low pLDDT scores related to situations where the model did not resemble any of the PDB structures or the hit was present but with a low probability of homology. These observations support the previous paragraph, suggesting that homology between the modelled protein and the training data leads to a prediction of a similar non-fibrillar structure with a high score. On the other hand, as few fibrils are present in the PDB, predicting amyloid fibrils is a rare event associated with poor scores of the model quality even if the predicted structure seems more correct.

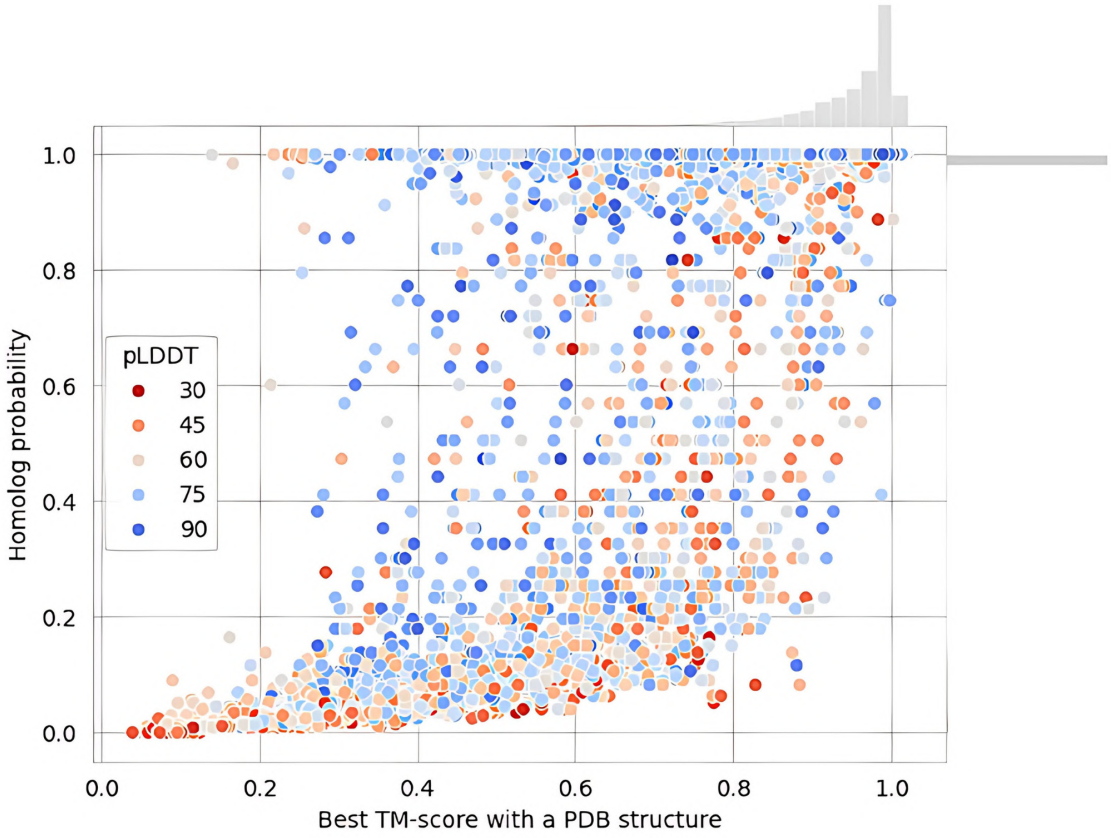

Figure 4.8: AF models for HSMonomers relate to the PDB structures. Results of the search of AF2 models in the PDB for HSMonomers. Homolog probability as a function of the TM-score between the AF model and the most similar PDB file. Outside of the X-axis, the histogram of TM-score values is provided, and outside of the Y-axis, the histogram of the homolog probability is provided.

4.3.5 Structure prediction for well-studied amyloid proteins

We investigated the performance of AF3 on amyloid proteins with unknown structure and observed problems with such predictions. Therefore, the question arises of whether AF is capable of learning the fibrillar form when it sees enough of such examples in its training dataset.

Models of well-studied amyloid proteins whose structures have been solved and seen by AF were predicted. Specifically, for each protein from the *Amyloid_structure* dataset, 50 monomeric and 50 six-unit models were generated. The summary of the results is presented in Table 4.2.

Predicted structures of 50 monomers for each protein were highly similar to each other. The clusterization procedure yielded only one representative model for transthyretin, α-synuclein, glucagon, immunoglobulin and Aβ-42. In the case of IAPP, monomeric models were grouped into two clusters. All monomeric models of Aβ-42, glucagon and IAPP were helical. For α-synuclein, 49 models were helical and 1 was disordered. This is in line with the experimental data on the secondary structure of these proteins, although often alternative conformations, such as disordered structures, are missed by AF3.

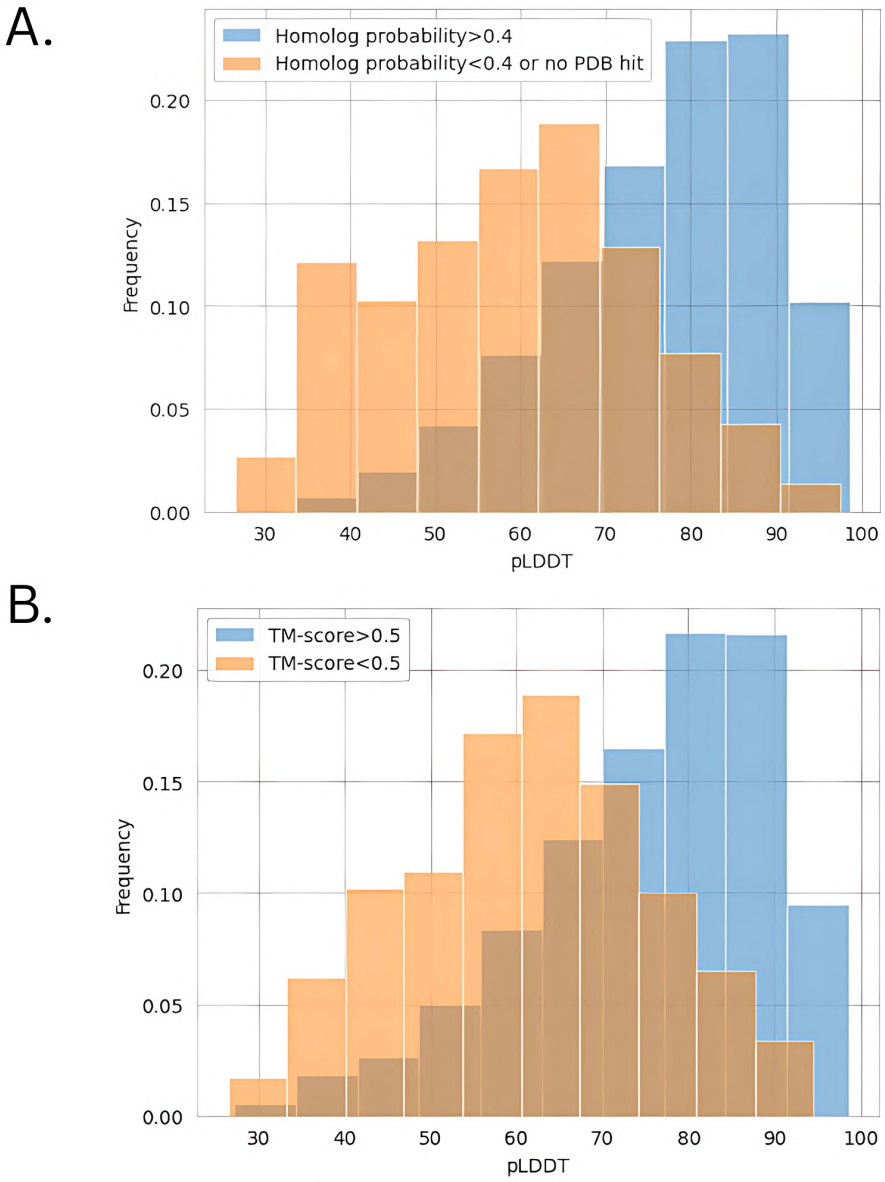

Figure 4.9: AF models for HSMonomers relate to the PDB structures. A. pLDDT distribution for HSMonomers models when the structure has a likely homologous hit in the PDB and when no similar structure was detected in the PDB or the homology was low. The difference was statistically significant; Mann Whitney test p-value=0.0. B. pLDDT distribution for HSMonomers models when the structure is similar to the hit in the PDB (TM-score>0.5) and when not (TM-score<0.5). The difference was statistically significant; Mann Whitney test p-value=1e-121.

Multimeric models of Aβ-42, α-synuclein and IAPP resembled amyloid structures (Fig. 4.10). In the case of glucagon models, 60% of them also resembled amyloid structures. None of the multimeric models of immunoglobulin and transthyretin resembled amyloid structure (Fig. 4.10).

Monomeric and multimeric models for *Amyloid_structure* were compared. Specifically, for each representative monomer identified after the clusterization, the most similar multimer was searched for with Foldseek. As expected, monomers of Aβ-42, IAPP and glucagon

Protein	Monomers' structures (**A**)	Multimers' structures (**B**)	Similarity between **A** and **B**	% of **B** similar to PDB	Hits in the PDB for **B**	*Mean TM-score (std)
Aβ-42	all helical	all amyloid	none	76%	all amyloid	0.36 (0.09)
α-synuclein	all helical apart from one disordered	all amyloid	locally in the helical part for 26 models (mean TM-score= 0.28)	100%	all amyloid	0.99 (<0.01)
Glucagon	all helical	60% amyloid	none	20%	varied	0.45 (0.14)
IAPP	all helical	all amyloid	none	89%	all amyloid	0.22 (0.03)
Immuno-globulin lambda variable 3-19 light chain	rich in beta-sheets	not amyloid	high (mean TM-score= 0.96)	100%	all globular	1.0 (<0.01)
Transthyretin	rich in beta-sheets	not amyloid	high (mean TM-score= 0.94)	100%	all globular	0.96 (0.03)

Table 4.2: Results for Amyloid_structure dataset. *Mean TM-score between the multimer (**B**) and PDB (standard deviation in parenthesis is given).

and α-synuclein (TM-score=0.28) did not resemble the multimers. Monomeric and multimeric models of immunoglobulin and transthyretin were highly similar. Multimeric models of immunoglobulin, transthyretin and α-synuclein resembled PDB structures. In the case of IAPP, glucagon and Aβ-42 this similarity to the PDB was lower.

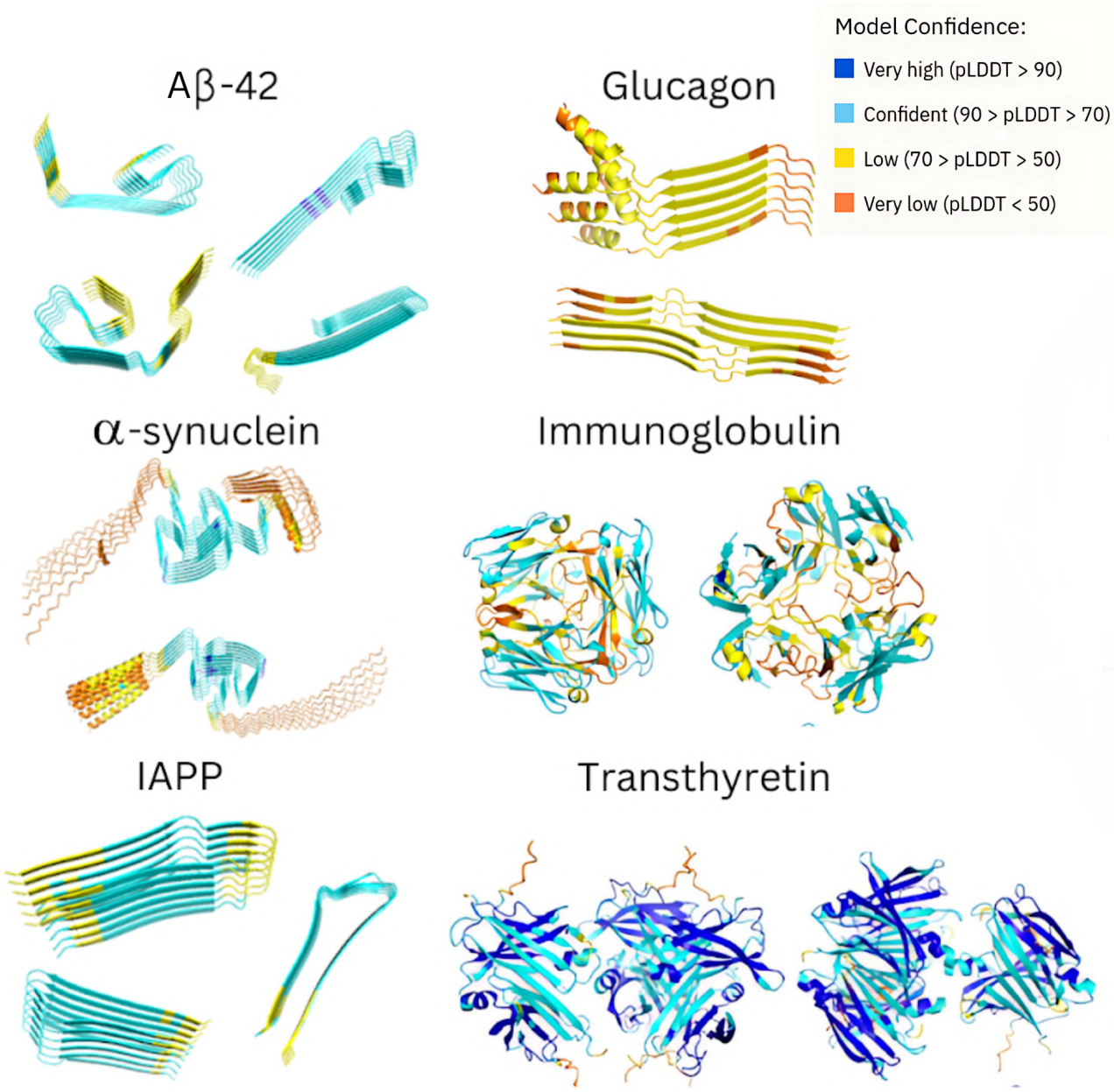

Figure 4.10: Examples of AF3 predictions for proteins from the Amyloid_structure dataset.

4.4 Discussion

AF is a state-of-the-art software that revolutionized the field of structural biology. It attempts to generalise the structure prediction problem, allowing for modelling proteins, DNA, RNA, and complexes of those, with the inclusion of ions. Accounting for such a broad context could potentially improve amyloid structure prediction that requires that multiple factors are taken into account, giving hope for proteome-wide screens of aggregation-prone proteins, and not only fragments of such. Unfortunately, AF3 struggles with amyloid proteins.

The analysis performed in this chapter revealed that only for one-third of the tested amyloid sequences AF predicted seemingly correct amyloid fibrils. Importantly, such models were often of much lower quality than the multimeric globular models, suggesting that metrics other than pLDDT may be suitable for assessing the quality of AF models for amyloid proteins. Problems with potentially misleading pLDDT scores have already been observed in other cases, such as dimers and fold-switching proteins [246, 251, 252]. Presented results suggest that AF scores a model highly when a homologous example is part of the AF training. Others have shown that it is possible to train a classifier that distinguishes correct AF models of dimers from the wrong ones [246], even when both have high pLDDT scores. Perhaps, to assess whether AF amyloid models are wrong or correct, we need to take a similar path and design new metrics specific to this case in the future.

AF also struggled with the negative dataset, predicting the amyloid fibrils in 56% of the cases. This was closely related to the fact that many peptides known not to aggregate are mutants of commonly studied amyloids, and AF is broadly known to be insensitive to mutations [253].

The problem with predicting amyloid structures with AF appears to be related to several factors. Many of the amyloid proteins require that the software manage to predict different multimeric and monomeric forms of these proteins, as these are often different. A structure prediction algorithm that does so can be suspected to generalize well to the problem and not only memorize the seen examples. Half of the multimeric models of amyloid proteins shared some similarities with monomeric ones, and one-third significantly resembled the monomers (TM-score>0.5). In 86% of the cases when any similarity appeared between a monomer and multimer, the multimeric models did not resemble amyloid fibrils. The training dataset is of fundamental importance for AF, which is natural for any machine learning method. But this approach comes at the price of struggling with unusual examples, like amyloid proteins. Few amyloid structures have been solved, particularly for longer proteins. As amyloid proteins are a vastly underrepresented group, this naturally influences the prediction of amyloid structures. However, the hope remains, as in the case of well-studied amyloid proteins with structures seen by AF3 during the training, it manages to predict the fibrillar models for them. Even though the number of solved amyloid structures must be big enough to leave a signal in the neural network, one solution may not be enough. This was observed for transthyretin and immunoglobulin, for which, although amyloid structures are available, they constitute the minority of all depositions for these proteins, leading to the bias toward globular models.

The training bias is also related to the homology information that is used by AF. In previous studies, it was shown that proteins which are homologous but can adopt two different structures are problematic for AF, and only one conformation for both is predicted [254]. This supports the results presented for amyloid proteins. Aggregation of amyloid proteins does not need to be conserved, and homologs of the amyloid protein

may not form fibrils. Therefore, AF reliance on the homology information explains the frequent presence of globular models for amyloid proteins, which resemble the training examples present in the PDB. This was particularly notable in the case of longer proteins. It seems that the presence of globular homologs in the training data biases the predictions for amyloid proteins. Hence, AF seems to be capable of predicting amyloid fibrils in some cases and undesired models resembled globular homologs present in the PDB. When modelling amyloid proteins with AF a fibrillar model of low quality with respect to the pLDDT should not be surprising, as exploration of other folds that reach beyond the PDB is often related to a poor model quality, as demonstrated for *HSMonomers*. Importantly, no difference in performance between functional and pathological amyloids was observed, suggesting that AF problems with predicting amyloid structures are a combination of the training examples and homology information.

AF's reliance on the homology information and abundance of protein structure gave it the deserved state-of-the-art place in protein structure prediction. Unfortunately, as homology information is tricky in the case of amyloids and few amyloid structures are determined, particularly for longer proteins, it is not surprising that AF struggles with these difficult proteins. Given these circumstances, it seems impressive that in some scenarios, it can predict the amyloid fibril structure, although it is often of poor quality with respect to pLDDT. We suggest that using shorter sequences and different protein copy numbers, comparing monomeric and multimeric prediction, checking if the structure of a homolog is solved in the PDB and was used in the training, predicting more than one structure, and careful interpretation of the model with metrics other than pLDDT can help in amyloid structure prediction with AF3. In the future, certain manipulations of the Multiple Sequence Alignments and determination of novel amyloid structures, particularly for longer proteins, could push this research forward.

The obtained results are disappointing and point toward the need for reliance on the sequence information of bacterial functional amyloids until more amyloid structures are experimentally determined. Therefore, the third hypothesis *that structures of amyloid fibrils can be predicted with AlphaFold 3* is rejected, and the corresponding goal *to investigate AlphaFold 3 performance on amyloid proteins, including bacterial functional amyloids* is realized. In the light of observations made in this chapter, it can be concluded that the structural predictions for bacterial functional amyloids are not a reliable source of information yet and will not help in understanding the molecular interactions between these proteins and human proteins. Hence, the next chapter uses only sequence data to give an overview of bacterial functional amyloids produced by microbes inhabiting humans and their potential interactions with the human proteome.

Results presented in this chapter are related to the preprint: ***Wojciechowska, A. W.****, Wojciechowski, J. W., & Kotulska, M. (2023). Non-standard proteins in the lenses of AlphaFold3 - a case study of amyloids. bioRxiv. This preprint was updated in line with this chapter. Submitted to the journal.*

Bacterial functional amyloids in the human microbiome

5.1 Introduction

The role of the gut microbiome in human health is undeniable by now. Intestinal bacteria supply essential substances such as vitamin B12 or short-chain fatty acids, maintain epithelial integrity, and regulate the host's immune response, influencing cytokine production and lymphocyte activation [255]. Changes in the human gut microbiome have been associated with the onset and progression of multiple diseases, including neurodegeneration. However, details of molecular cross-talk of the gut-brain axis remain elusive.

Bacterial functional amyloids could be important actors in this interplay. Bacterial functional amyloids are produced by bacterial strains that inhabit the human gut microbiome. Their structural similarity to pathological amyloids could lead to amyloid interactions between microbial and pathological amyloids.

It has been hypothesized that the pathological aggregation of α-synuclein in Parkinson's disease starts in the enteric nervous system. Amyloid structures of α-synuclein aggregation in the enteric nervous system appear at the beginning of this disorder and correlate with disease severity [92, 93]. The possible triggers of this pathological process could be microbial amyloids that inhabit the intestine and may have contact with the neuronal cells of the enteric nervous system. Once α-synuclein aggregation is triggered outside the central nervous system, according to Braak's hypothesis, it could propagate in a prion-like manner from the enteric nervous system through the vagus nerve to the brain [91]. Multiple studies have shown that microbial amyloids can affect the rates of aggregation of pathological ones [65]. The inhibition of bacterial functional amyloid CsgA protein in mouse models reduces aggregation of α-synuclein in the brain [97]. CsgA can also interact with Aβ and promote Alzheimer's pathology in C. elegans models, [98]. Biofilm-related bacterial functional amyloids are more abundant in Parkinson's disease patients than in healthy controls and can colocalize with α-synuclein in neurons, increasing its aggregation [101].

Microbial amyloids could have a broad impact on neurodegeneration. The intestines of patients with Parkinson's and Alzheimer's disease are often characterized by gut dysbiosis, dysfunction, and inflammation [73, 102]. In such conditions, the structural similarity of bacterial functional amyloids to pathological ones could trigger similar cytotoxic pathways and promote further inflammation and intestinal permeability [103]. Both microbial and human amyloids can activate immune receptors and inflammation pathways such as the NLRP3 inflammasome [106, 107].

Many of the previous studies focused on a single bacterial functional amyloid and its molecular interactions, providing evidence for the link between microbial amyloids and neurodegeneration. However, so far, no large-scale study that would gather information

from multiple resources has been performed on the microbiome scale. Here, this need is ansewered.

In preceding chapters, an overview of the available protein-protein interaction data in neurodegenerative diseases, sequence and structure analysis of bacterial functional amyloids were both provided. This chapter joins both views to investigate the potential clinical importance of bacterial functional amyloids in neurodegeneration. The fourth and last hypothesis that ***bacterial functional amyloids in the human intestine may influence neurodegeneration*** is investigated. For this goal, bacterial functional amyloids in the human gut microbiome are identified, and their interactions with human proteins are predicted. An atlas of known bacterial functional amyloids and their taxonomic origin is given. The prevalence of bacterial functional amyloids in the health and disease microbiome is discussed. Finally, a framework of potential molecular mechanisms of bacterial functional amyloids with human proteins in neurodegeneration is proposed.

5.2 Methods

Identification of bacterial functional amyloids in the human microbiome

Unified Human Gut proteome UHGP published by Almeida et al. clustered at 95% of identity (v1.0 edition, file name: uhgp-95.faa) was used as a reference dataset of the human gut microbiome proteome [256]. The dataset of bacterial functional amyloids BFA was described in the third chapter of this thesis.

To identify homologs of BFA in UHGP, a homology search with mmseqs was performed [245]; command: mmseqs search bfa uhgp-95 results tmp –comp-bias-corr 0 –mask 0. On purpose, the compositional correction was turned off, as multiple amyloids may contain repeated fragments or low-complexity regions.

The identified homologs were filtered with the entire sequence amyloid predictor AmyPred-FRL tool in the webserver version [118]. The AmyPred-FRL returns a prediction score from 0 to 1, where 1 means a highly probable amyloid and 0 a non-aggregating protein. The cutoff of 0.8 was applied. Sequences classified as amyloid-positive by AmyPred-FRL were additionally filtered with ArchCandy [112], to ensure they are likely to contain beta-arch motifs, typical for amyloids, in the sequences. To run ArchCandy, the following command was used: java -jar ArchCandyV2.jar –TMfilter -t=0.5 -i=SeqID Sequence. The resulting sequences were denoted as the UHGPAmyloids dataset. This approach uses predictors trained mostly on pathological amyloids and can miss many of the functional amyloids. However, it should limit the number of false positive examples, which is crucial when screening a big dataset like UHGP.

UHGPAmyloids was clustered with cd-hit [244], command: cd-hit -i inputfile.fasta -o outputfile.fasta -c 0.9/0.8/0.7/.

Taxonomy of the predicted bacterial functional amyloids

Taxonomic assignments were already provided by the authors of the UHGP dataset. Throughout the text, the same nomenclature as found in the UHGP files was used to minimize confusion. Shannon entropy of taxonomic assignments was calculated with the phyloseq package in R [257].

Cellular localization of the predicted bacterial functional amyloids

The cellular localizations of the UHGPAmyloids were predicted with the BUSCA webserver with a "taxonomic group" switched to "Prokarya - Other - 3 compartments" [258]. The BUSCA predicts the localization by identifying characteristic domains and motifs in the sequence.

Protein-protein interactions between predicted bacterial functional amyloids and human proteins

UHGPAmyloids proteins with *Extracellular* or *Plasma membrane* localization, according to BUSCA, were extracted and formed UHGPAmyloids_filtered dataset.

Human proteins were extracted from the Human Protein Atlas [259] (file: normal_tissue.tsv, version: 23.0, available in the *Downloadable* data section on the project website). To draw out proteins expressed in the intestine, The Human Protein Atlas was filtered so that the *Tissue* column, in the normal_tissue.tsv file, included at least one of the following words: *colon*, *small intestine*, *duodenum*, *rectum*. Then, proteins with *Subcellular localization [CC]* including one of the following expressions: *Cell membrane*, *cell membrane*, *secreted*, *Secreted*, *extracellular*, *Extracellular*, *cell surface*, *Cell surface*, *junction*, *Junction*, *secretory*, *Secretory*, *cell wall*, *Cell wall* were extracted. This dataset, referred to as HPA_filtered, should include proteins that are the most likely to interact with substances secreted by the microbiome.

Protein-protein interaction predictions between predicted bacterial functional amyloids from the UHGPAmyloids_filtered and human proteins from the HPA_filtered were predicted with ProteinPrompt [260] by Jakub Wojciechowski.

Overrepresentation Analysis

Human proteins from HPA_filtered with at least 5 interactors from UHGPAmyloids_filtered were considered for GO and KEGG terms overrepresentation analysis. The overrepresented terms were calculated with the ClusterProfiler package available in R [261].

The F-test was performed for the top 20 terms for *Biological Process*, *Molecular Function* and *KEGG Pathways* between proteins HPAIntestine_filtered, which had at least 5 interacting partners in UHGPAmyloids and the entire set *HPAIntestine_filtered* with *f.test* function in R programming language. P-values were corrected using Bonferroni correction.

Data analysis and visualization

All data analyses and visualizations were performed in Python 3 with the following packages: NumPy [155], Pandas [156], SciPy [157], Matplotlib [158], Seaborn [159].

5.3 Results

5.3.1 Presence of bacterial functional amyloids in the human gut microbiome

The aggregation propensity of bacterial functional amyloids should be evolutionarily conserved, as discussed in previous chapters. Hence, a pool of potential bacterial func-

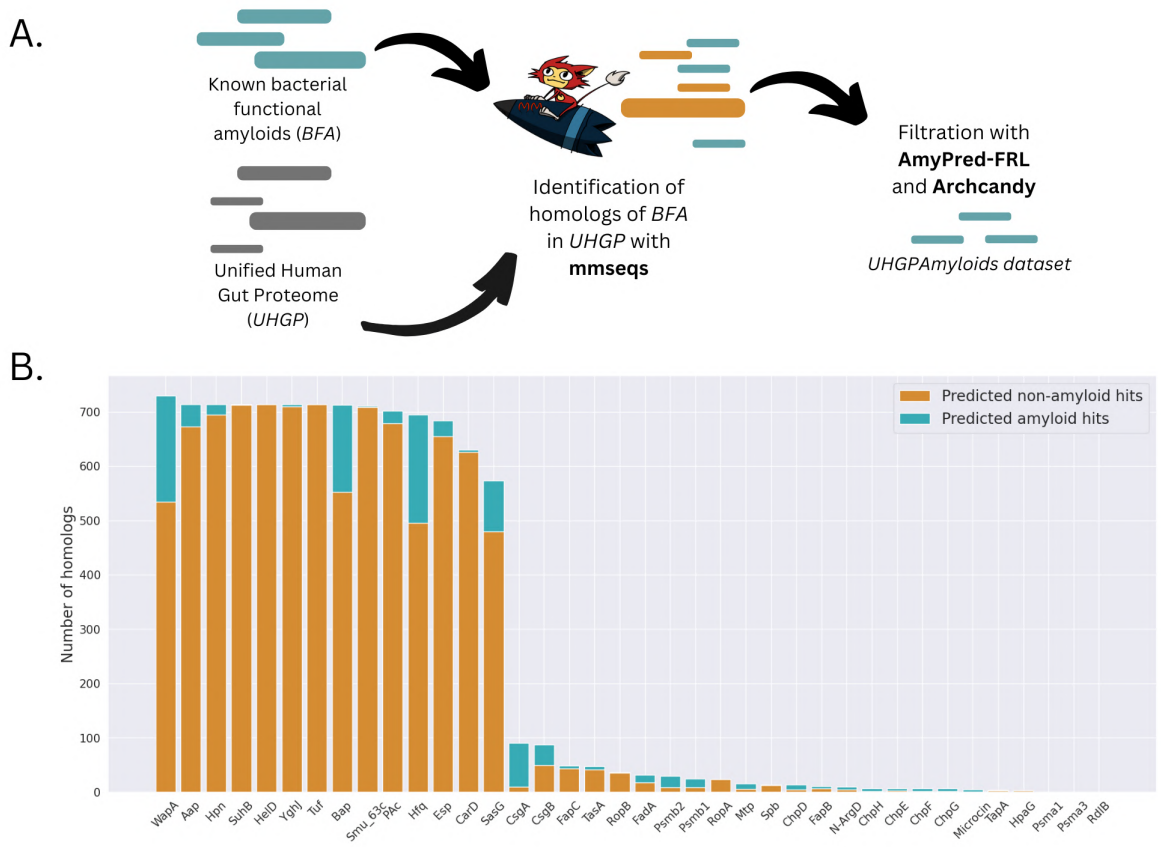

Figure 5.1: Identification of bacterial functional amyloids in human gut microbiome proteome. A. Pipeline for the search of bacterial functional amyloids in UHGP, B. Number of homologs found per BFA protein.

tional amyloids in the human microbiome proteome (UHGP) was created by searching for homologs of BFA proteins in the UHGP dataset [262]) (Figure 5.1). 10249 matches were found, which, after the removal of duplicates, gave 9541 unique sequences (mean sequence identity equal 45%, mean E-value was 5e-6). The homology search yielded multiple sequences for longer bacterial functional amyloids as a result of different domain matches (WapA: 730, HelD: 714, Aap: 714, YghJ: 714, SuhB: 714, Bap: 713, Smu63c: 711, PAc: 702, Tuf: 714). Three proteins, alpha phenol-soluble modulin and RdlB, had no homolog in the UHGP.

The aggregation propensity of the identified homologs was predicted with the full-length aggregation propensity predictor AMYPred-FRL to remove possible non-aggregating proteins. AMYPred-FRL was trained mostly on pathological amyloids and probably leads to multiple false negatives, but at the same time should reduce the false positive hits. AMYPred-FRL filtration discarded 90% of the sequences, leaving 855 proteins. No correlation was found between the AMYPred-FRL score and any of the homology parameters (Pearson's $|R| \leqslant 0.2$). The remaining 855 proteins were additionally tested with the ArchCandy tool that predicts the occurrence of beta-arch motifs characteristic of amyloid proteins. The next 50 sequences were discarded. This final dataset of 805 predicted bacterial functional amyloids from the human gut microbiome is referred to as UHGPAmyloids.

The UHGPAmyloids were clustered at the levels of 90%, 80% and 70% identity, giving

412, 302 and 243 sequences, respectively. The rather moderate decrease in the number of sequences left after clusterization indicates the relative diversity of the found bacterial functional amyloids.

Only UHGPAmyloids proteins are studied and homologs predicted as non-aggregating are ignored in the upcoming analyses.

5.3.2 Taxonomic origin of intestinal bacterial functional amyloids

The taxonomic origin of UHGPAmyloids was analyzed and Shannon's entropy for each group of homologs was calculated (Figure 5.3) to identify bacteria that produce predicted bacterial functional amyloids. Furthermore, the abundance of different taxonomic groups in UHGPAmyloids and UHGP datasets was compared with Fisher's exact test (Figure 5.2, Table 5.1). The proteome of our microbiome is mostly produced by Firmicutes, Bacteroidota, Proteobacteria and Actinobacteria. Despite that, the UHGPAmyloids proteins are not evenly distributed between these phyla. The enrichment in Firmicutes, Proteobacteria and Fusobacteria and the depletion in Actinobacteriota and Bacteroidota were found for UHGPAmyloids.

The taxonomic distribution of UHGPAmyloids indicates that these proteins are widely present in the bacterial tree of life. The RNA-binding Hfq protein and biofilm-related proteins had the highest taxonomic diversity. The long biofilm-related amyloids WapA, Bap, PAc, Aap and Esp could be found mostly in Firmictutes but also in other phyla, such as Proteobacteria, Myxococcota or Actinobacteria. Homologs of short biofilm-related proteins, which were predicted to be prone to aggregation, could also be found in other bacterial families, which are not typically associated with these proteins. For example, the CsgA and CsgB proteins had low diversity at the phylum level, but their aggregating homologs were found not only in Enterobacteriaceae. Low taxonomic diversity was found for the aggregating homologs of the Mtp protein, which were only identified in Streptomycetes.

Phylum	Percentage of UHGPAmyloids from the phylum	Fisher test p-value	Category
Firmicutes	70.8%	1.6e-10	Overrepresentation
Proteobacteria	24.8%	5e-41	Overrepresentation
Fusobacteria	1.7%	1.7e-4	Overrepresentation
Actinobacteriota	1.4%	1.5e-12	Underrepresentation
Bacteroidota	0.9%	7.5e-67	Underrepresentation
Campylobacterota	0.2%	0.45	-
Myxococcota	0.1%	0.2	-

Table 5.1: Comparison of the phylum prevalence for the UHGPAmyloids and UHGP datasets.

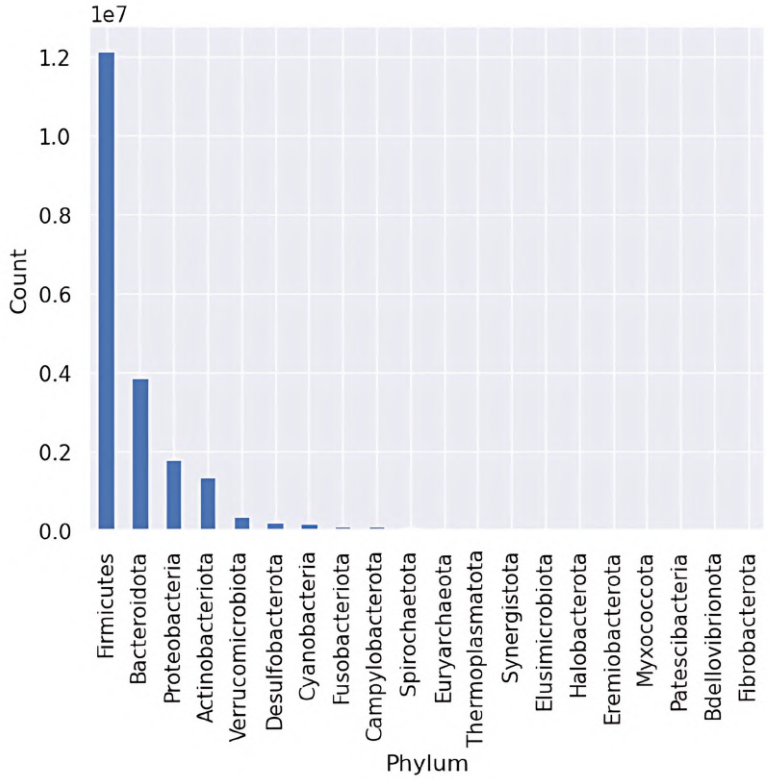

Figure 5.2: Phylum prevalence in the UHGP dataset.

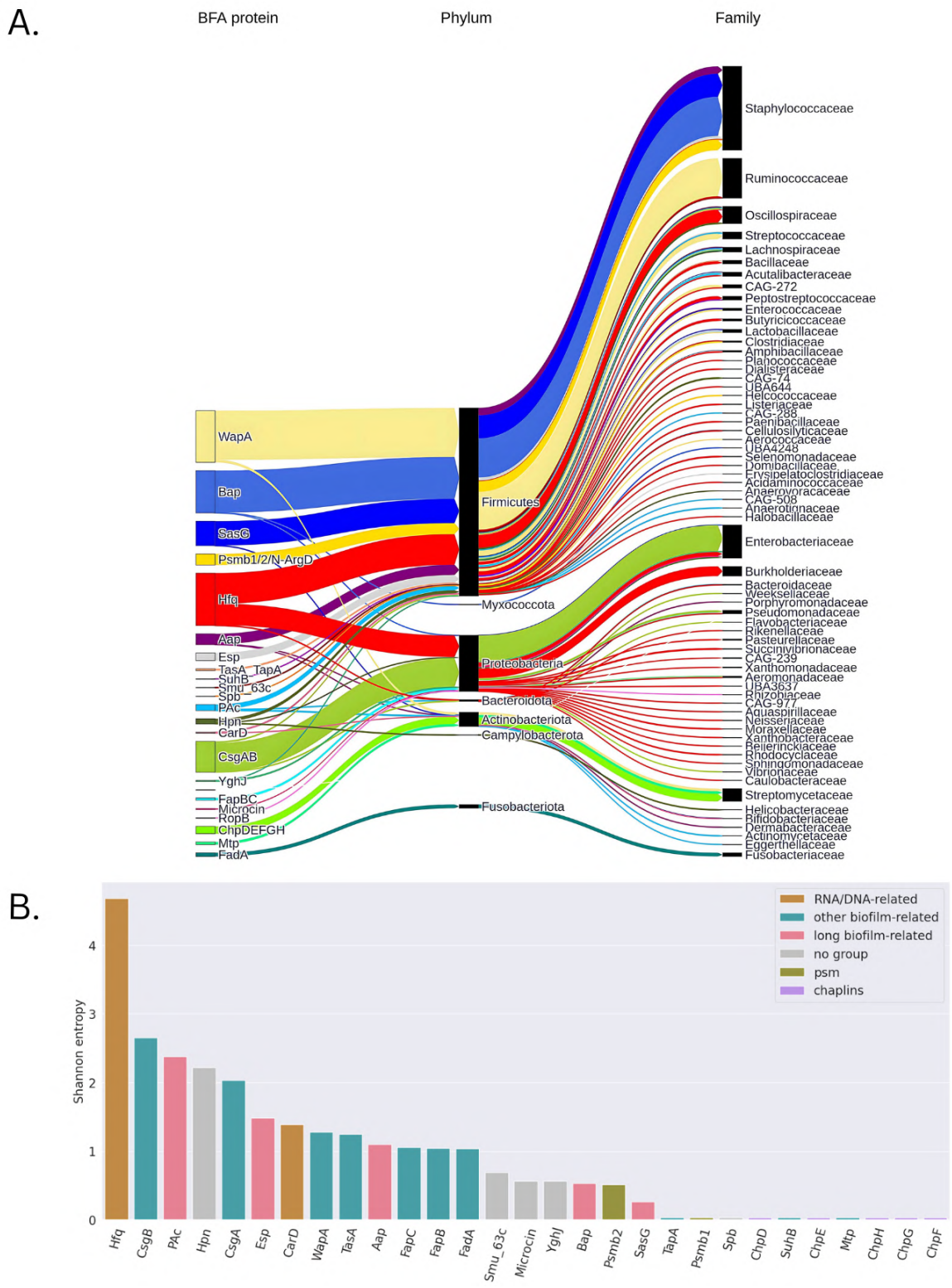

Figure 5.3: Taxonomic diversity of the UHGPAmyloids. A. Taxonomic origin of UHGPAmyloids proteins grouped by the BFA source protein. B. Shannon entropy for different groups of homologs in the UHGPAmyloids datasets.

5.3.3 Patient samples and the abundance of bacterial functional amyloids

The abundance of UHGPAmyloids in the gut microbiome samples of patients with neurodegenerative diseases was tested by Jakub Wojciechowski and Kinga Zielińska [263]. They studied three metagenomic datasets on Parkinson's disease, one metagenomic study on Alzheimer's disease, and one, as a negative control set, on Cryptococcal meningitis. Cryptococcal meningitis is an infectious brain disease not characterized by any amyloid deposition. Each dataset contained samples from patients with the disease and a respective control group. They performed a detailed processing of the raw reads with quality control and assembly procedures and searched for homologs of UHGPAmyloids in the samples.

The analysis showed a statistically significant greater presence of bacterial functional amyloids in samples from patients with Parkinson's disease than in the control in all three Parkinson's disease (Mann-Whitney U-test p-value $\leqslant 0.01$). This was not observed for the dataset on Alzheimer's disease or Cryptococcal meningitis.

The abundance of each bacterial functional amyloid family was analysed to identify the most important groups. Interestingly, in two out of three studies on Parkinson's disease, CsgA homologs appeared more frequently in patients than in healthy controls. The same was discovered for the dataset on Alzheimer's disease (Mann-Whitney U-test with Benjamini-Hochberg correction p-value $\leqslant 0.05$). These observations particularly point to the role of the curli protein in the pathology of amyloid disease. CsgA has already been found to interact with both α-synuclein and Aβ protein and promote neurodegeneration in Caenorhabditis elegans models [98], its apparent abundance could be of clinical significance.

5.3.4 Interactions of bacterial functional amyloids with human proteins

Cellular localization is of fundamental importance for the discussion of the molecular interactions between bacterial functional amyloids and human proteins. Hence, the cellular localization for both BFA proteins and UHGPAmyloids was predicted. 60% of the BFA source proteins were predicted as extracellular and 40% as cytoplasmatic. The UHGPAmyloids had similar proportions of predicted localizations: 43% were extracellular, 48% were cytoplasmatic, and 9% were membrane. The UHGPAmyloids proteins with extracellular localization were homologs of 25 different BFA proteins, mostly of biofilm-related ones like WapA (187 extracellular homologs) and Bap (129 extracellular homologs). In general, prokaryotic proteins are mostly cytoplasmatic (around 64%), sometimes membrane (around 20%) and less than 2% of them are extracellular. This leads to the conclusion that considered bacterial functional amyloids are more frequently extracellular or membrane than other bacterial proteins.

The frequently predicted non-plasmatic cellular localization of UHGPAmyloids gives grounds for the analysis of protein-protein interactions between bacterial functional and human proteins. To reveal potential molecular mechanisms of this interplay, only UHGPAmyloids proteins with predicted extracellular or membrane localization were considered (UHGPAmyloids_filtered, 417 proteins). To extract human proteins that could interact with bacterial functional amyloids, the Human Protein Atlas was used [264]. Proteins that are experimentally confirmed to be expressed in the human intestine and are biologically related to the extracellular membrane or the junction space were withdrawn

(HPAIntestine_filtered, 2,361 proteins). Then, possible interactions between proteins from HPAIntestine_filtered and UHGPAmyloids_filtered were computationally predicted.

183,742 interactions were predicted that affected 1,098 human proteins from HPAIntestine_filtered. As expected for protein-protein interaction systems, not all human proteins had an equal number of interactors from UHGPAmyloids_filtered and the distribution of this parameter was skewed (Figure 5.4). For each UHGPAmyloids_filtered protein, certain human protein interactors were predicted. Homologs of WapA had the highest number of interactors (114,058 in total), and homologs of TapA, had the lowest (135, in total).

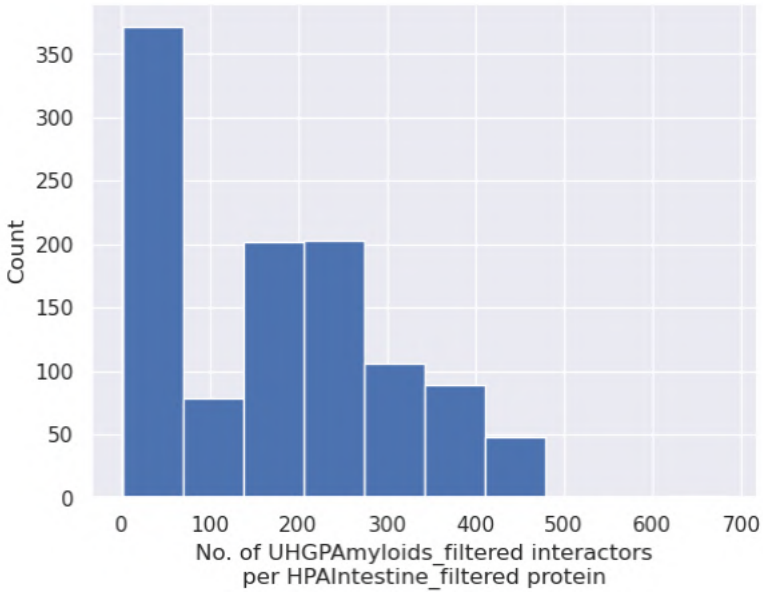

Figure 5.4: Histogram of the number of predicted human interactors (from the HPAintestine_filtered dataset) per UHGPAmyloids_filtered protein.

The most common interacting partner for bacterial functional amyloids was the human N-myc-interactor, Nmi protein (UniprotID: Q13287). It had 681 interactions with UHGPAmyloids_filtered proteins that spanned 20 different groups of homologs. The majority of Nmi interactions were with homologs of biofilm-related proteins (346 interactions with homologs of WapA, 92 with homologs of SasG, and 70 with homologs of CsgA). All predicted biofilm-related interactors of the Nmi protein originate from Proteobacteria, which is often associated with pro-inflammatory characteristics [265], or Firmicutes. Nmi protein is important from the immunological perspective as it interacts with interleukin-2 and STAT protein [266] and its level is often elevated in different cancers [267].

Among human proteins with the top five highest number of interactions with UHGPAmyloids_filtered, there were C-C motif chemokine 5, Interferon-induced transmembrane protein 1 and 2, both related to immune response, claudins that control the epithelial barrier, and ubiquitin ligases (see the full list in Table 5.2) [268]. Human pathological amyloids such as α-synuclein, major prion protein, and APP cutting protein presenilin-1 were also predicted to interact with 417 UHGPAmyloids_filtered proteins each. Tau and APP had 229 and 294 predicted interactions, respectively.

Uniprot ID	Description	Number of interactions
Q13287	N-myc-interactor	684
P18848	Cyclic AMP-dependent transcription factor ATF-4	460
Q9H6Y7	E3 ubiquitin-protein ligase RNF167	418
P13501	C-C motif chemokine 5	417
P30874	Somatostatin receptor type 2	417
O15551	Claudin-3	417
O14493	Claudin-4	417
P21926	CD9 antigen	417
Q9H3Z4	DnaJ homolog subfamily C member 5	417
P10082	Peptide YY	417
P49768	Presenilin-1	417
P04156	major prion protein	417
P02745	Complement C1q subcomponent subunit A	417
O00264	Membrane-associated progesterone receptor component 1	417
P37840	α-synuclein	417
P67809	Y-box-binding protein 1	417
Q01629	Interferon-induced transmembrane protein 2	417
P13164	Interferon-induced transmembrane protein 1	417
Q99942	E3 ubiquitin-protein ligase RNF5	417
P32241	Vasoactive intestinal polypeptide receptor 1	417
P08172	Muscarinic acetylcholine receptor M2	417
P63000	Ras-related C3 botulinum toxin substrate 1	416
P61586	Transforming protein RhoA	416
P60953	Cell division control protein 42 homolog	416
Q9Y328	Neuronal vesicle trafficking-associated protein 2	416
Q9UKJ5	Cysteine-rich hydrophobic domain-containing protein 2	416
P26583	High mobility group protein B2	416
P04899	Guanine nucleotide-binding protein G(i) subunit alpha-2	416
P63096	Guanine nucleotide-binding protein G(i) subunit alpha-1	416
P21453	Sphingosine 1-phosphate receptor 1	416

Table 5.2: Proteins from HPAIntestine_filtered representing the top five numbers of interactions from UHGPAmyloids_filtered (31 proteins in total).

5.3.5 Pathways potentially affected by intestinal bacterial functional amyloids

To get an overview of potential molecular pathways and processes affected by intestine bacterial functional amyloids, a two-step overrepresentation analysis was performed. Proteins from HPAIntestine_filtered with at least five predicted interactors from UHGPAmyloids_filtered were withdrawn. Then, GO terms, which were overrepresented in this group with respect to the entire human proteome, were extracted (Figure 5.5). Then it was checked if the identified go terms are also overrepresented with respect to the intestinal proteins HPAIntestine_filtered (Table 5.3 and 5.4). The performed analysis showed that bacterial functional amyloids could have a broad impact on the human proteome and interactome. Human proteins interacting with UHGPAmyloids_filtered were found to be enriched in GO Biological Process terms related to vesicle transport, protein localization, signaling, cell-cell and cell-matrix adhesion. UHGPAmyloids_filtered were also predicted to interact with claudins forming cell-cell junctions and maintaining the epithelial barrier function. The gut's impermeability is directly linked to these proteins, and their malfunctioning is associated with gastrointestinal diseases such as Inflammatory Bowel Disease (IBD), Ulcerative Colitis (UC), Crohn's Disease (CD) or Colorectal Cancer (CRC) [269, 270].

The predictions demonstrated that the intestinal bacterial functional amyloids could also affect signalling and transport in human cells. The enrichment in cytokine-binding proteins, growth factors and G-protein-coupled receptors, as well as tyrosine kinase receptors, all regulating cell response to internal and external stimuli, was found. Other overrepresented terms were related to transport and included endocytosis, exocytosis, and integrin-binding. The integrin-binding properties were noticed for human pathological amyloids such as Aβ-42 and tau, suggesting the overlap between the interactors of microbial and pathological amyloids.

The analogous two-step ORA was performed concerning KEGG pathways and Disease Ontology. As for GO analysis terms such as cell-to-cell adhesion, including focal adhesion, cell and adherens junctions, or transport were identified. In addition, multiple cancer-related pathways were observed. The "Proteoglycans in cancer" pathway regards proteins that are related to cell proliferation, migration and adhesion in cancerous cells. MAPK signalling pathway also influences cell proliferation, differentiation and death. The predicted interactions indicate that bacterial functional amyloids could generally affect immune-related pathways, as terms like chemokine signalling pathway, leukocyte migration, and response to viral infections were also noted.

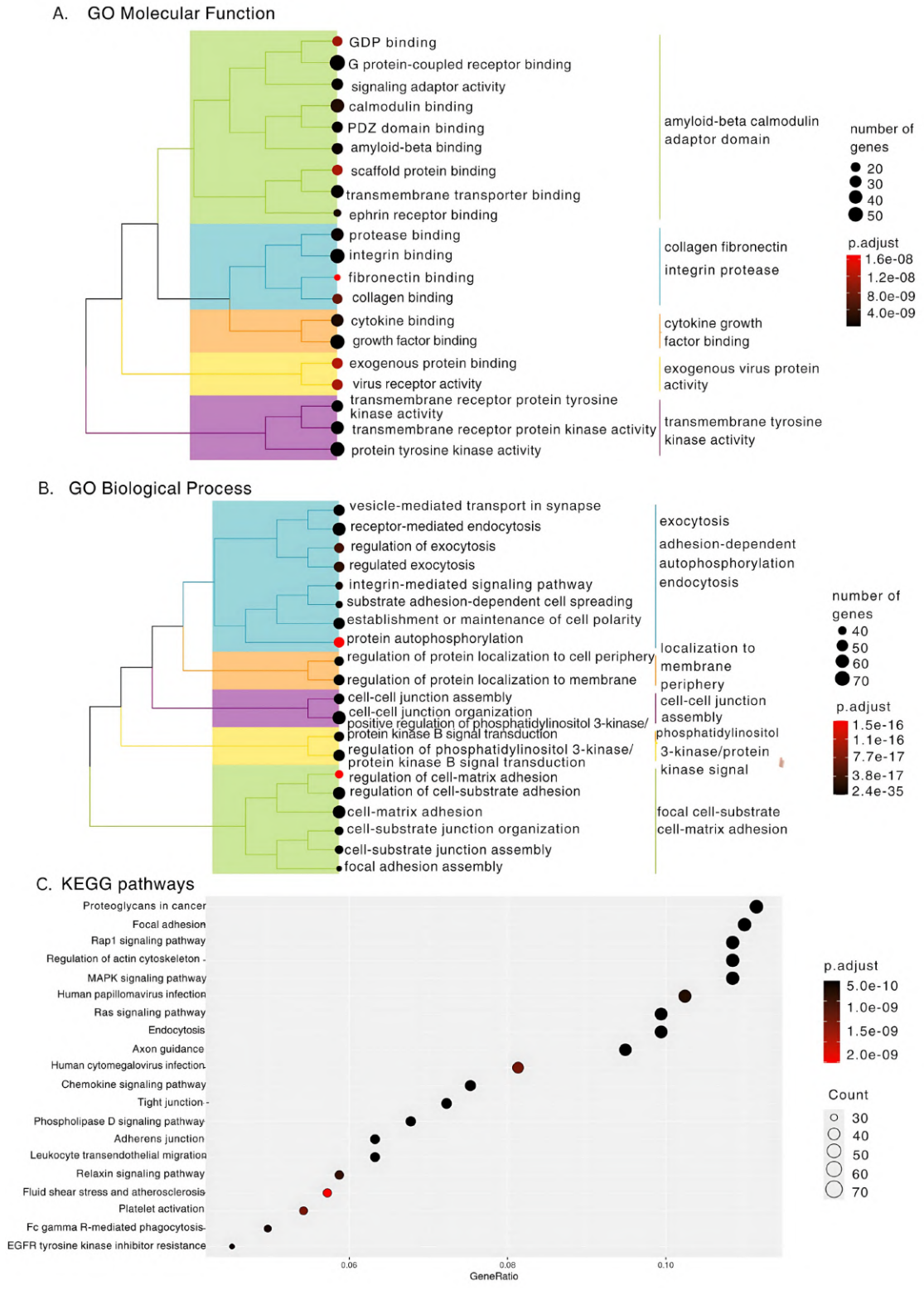

Figure 5.5: The results of the overrepresentation analysis for A. Gene Ontology Molecular Function, B. Gene Ontology Biological Process C. KEGG pathways.

GO/KEGG Term	Fisher Test's p-value after BH correction
growth factor binding	0.004
transmembrane receptor protein kinase activity	0.003
integrin binding	0.016
protein tyrosine kinase activity	0.002
G protein-coupled receptor binding	0.02
transmembrane receptor protein tyrosine kinase activity	0.006
protease binding	0.004
cytokine binding	0.035
PDZ domain binding	0.06
transmembrane transporter binding	0.021
amyloid-beta binding	0.016
signaling adaptor activity	0.009
collagen binding	0.06
GDP binding	0.016
virus receptor activity	0.068
exogenous protein binding	0.068
ephrin receptor binding	0.029
G protein activity	0.013
calmodulin binding	0.019
cell-cell junction organization	0.002
receptor-mediated endocytosis	0.002
cell-matrix adhesion	0.002
regulation of cell-substrate adhesion	0.002
cell-cell junction assembly	0.006
cell-substrate junction assembly	0.002
cell-substrate junction organization	0.002
regulation of protein localization to membrane	0.005
regulation of phosphatidylinositol 3-kinase/protein kinase B signal transduction	0.002
focal adhesion assembly	0.003
positive regulation of phosphatidylinositol 3-kinase/protein kinase B signal transduction	0.002

Table 5.3: Fisher tests results for each GO and KEGG term with Benjamini-Hochberg correction (number of tests = 60) with respect to intestinal proteins only (HPAIntestine_filtered).

GO/KEGG Term	Fisher Test's p-value after BH correction
regulation of protein localization to cell periphery	0.004
establishment or maintenance of cell polarity	0.013
substrate adhesion-dependent cell spreading	0.007
vesicle-mediated transport in synapse	0.004
integrin-mediated signaling pathway	0.002
regulated exocytosis	0.008
regulation of exocytosis	0.004
regulation of cell-matrix adhesion	0.006
regulation of epithelial cell migration	0.002
Proteoglycans in cancer	0.002
Focal adhesion	0.004
Rap1 signaling pathway	0.005
Axon guidance	0.007
Regulation of actin cytoskeleton	0.003
Ras signaling pathway	0.006
Adherens junction	0.013
Endocytosis	0.005
MAPK signalling pathway	0.003
Leukocyte transendothelial migration	0.02
Tight junction	0.053
Phospholipase D signaling pathway	0.006
Chemokine signaling pathway	0.051
Fc gamma R-mediated phagocytosis	0.028
EGFR tyrosine kinase inhibitor resistance	0.018
Human papillomavirus infection	0.016
Relaxin signaling pathway	0.038
Platelet activation	0.104
Human cytomegalovirus infection	0.013
Fluid shear stress and atherosclerosis	0.01

Table 5.4: (**Continued**) Fisher tests results for each GO and KEGG term with Benjamini-Hochberg correction (number of tests = 60) with respect to intestinal proteins only (HPAIntestine_filtered).

5.4 Discussion

The human microbiome is a complex system with a broad influence on our mood, dietary habits, health and incidence and progression of multiple diseases, including neurodegenerative ones. The interplay between gut and bacteria can involve a wide range of molecular mechanisms, one of them potentially involving bacterial functional amyloids. The structural similarity of microbial amyloids to pathological ones could lead to a two-step mode of action of these proteins. In line with the Braak hypothesis for Parkinson's disease, they could serve as triggers of pathological aggregation in the enteric nervous system via direct protein-protein interactions occurring in the colon neurons. The pathology could then spread to the brain and central nervous system through the vagus nerve [91]. Due to mechanisms of molecular mimicry, bacterial functional amyloids could activate the same molecular pathways as pathological amyloids and, for example, affect the immune system [103]. This is further supported by the fact that both types of amyloid have already been discovered to activate Toll-like receptors and signal inhibitory receptors in leukocytes 1 (SIRL-1) [104].

This part of the thesis focused on providing an overview of bacterial functional amyloids in the human microbiome and their potential impact on neurodegeneration. Previous studies often discussed theoretical considerations or experimental analyses of a single microbial amyloid. Here, the analysis of multiple available microbiome datasets and related predictions was conducted to provide the atlas of bacterial functional amyloids and their potential interactions with human proteins.

The search for potential bacterial functional amyloids in the human microbiome with homology methods provided a wide range of sequences in seven different bacterial phyla. Such proteins turned out to be more abundant in the microbiomes of Parkinson's disease patients than in healthy controls for all three considered metagenomic datasets. No such observation was made for either Alzheimer's disease datasets or Cryptococla meningitis, the latest also involving brain inflammation but not the presence of amyloid plaques. Despite that, a greater abundance of CsgA protein homologs seemed to be universal for samples from neurodegenerative patients. It was found for two out of three Parkinson's disease datasets, the Alzheimer's disease dataset, but not for the Cryptococla meningitis one. This gives the CsgA protein a special place in this discussion.

CsgA proteins have been shown to interact with both α-synuclein and Aβ-42 [271, 98, 240]. Furthermore, CsgA inhibition reduced neuronal death in Caenorhabditis elegans [98]. Curli proteins, similarly to Aβ, have a pathogen-associated molecular pattern that leads to the activation of immunological pathways [272]. Both amyloids are recognized by toll-like receptors (specifically, TLR2 / TLR1 immune sensor-receptor system) responsible for inflammation regulation [273]. Based on this evidence, the observed greater abundance of homologs of CsgA protein alone in samples from patients with Alzheimer's disease could be a promoter of inflammation, cytokine release, increased intestinal permeability and, consequently, of the onset and progression of neurodegeneration.

Identified bacterial functional amyloids (UHGPAmyloids) could have room for interaction with human proteins. The predictions of cellular localization of these proteins revealed a 20-fold more frequent occurrence of extracellular proteins than in the case of other bacterial proteins. Our limited knowledge of bacterial functional amyloids could be biased toward extracellular examples of these proteins, leading to this result. Even if so, many of the known and predicted bacterial functional amyloids are likely to interact with the extracellular environment and, consequently, with the human protein interactome.

Analysis of predicted protein-protein interactions revealed that intestinal bacterial functional amyloids could affect the functioning of multiple human proteins, including those responsible for endocytosis and exocytosis, signaling, and cell transport. The high incidence of CsgAB amyloid proteins in Proteobacteria could be the reason behind the pro-inflammatory properties of this bacterial group, as their populations are increased in Parkinson's disease, type II diabetes, and Alzheimer's disease [274, 275, 276]. Interestingly, according to the performed computational analysis, bacterial functional amyloids could also be present in anti-inflammatory bacteria, such as Lactobacillus and Bifidobacterium, which have clinical applications. Although the number of amyloid proteins identified in these groups was low, it could still be important in the cases of increased gut permeability.

The computational analysis revealed that bacterial functional amyloids, as pathological ones, could interfere with cell junctions. The negative impact of protein aggregation on epithelial cell integrity was previously discussed [41]. Human amyloids such as α-synculein, Aβ-42, tau and apolipoproteins can downregulate the expression of tight junctions that are responsible for the impermeability of the blood-brain barrier. [41, 42, 43] Analogously, the abundance of bacterial functional amyloids could have an unfavourable effect on gut permeability.

The presented results point to potential pro-inflammatory properties of intestinal bacterial functional amyloids. According to the predictions, they could interact with human proteins involved in chemokine signaling and leukocyte migration. Bacterial functional amyloids appear to be able to interact with integrin and calmodulin-binding proteins, which may have a neurotoxic effect.

The study showed that bacterial functional amyloids in the human gut microbiome have a great potential to play an important role in the gut-brain axis. The changes observed in the functionality of the microbiome in Parkinson's disease patients could be related to the different compositions of the microbiome, which may be characterized by an increase in bacterial species expressing functional amyloids. As a result, a pro-inflammatory positive feedback loop may appear. Abundant intestinal bacterial functional amyloids could affect the immune system and interact with tight junctions, leading to greater gut permeability. The greater permeability, on the other hand, could enable functional bacterial amyloids to penetrate the epithelial barrier and promote inflammation. The structural similarity of bacterial functional amyloids gives them the room to interact with pathological amyloids, such as α-synuclein, potentially triggering their aggregation in the enteric nervous system. Pathological aggregation could spread in a prion-like manner between the vagus nerve and then the brain, as suggested by Braak's hypothesis. The potential route of interactions discussed here, supported by a wide-scale computational analysis, sheds light on the clinical significance of these proteins, which still requires further in vivo experiments. This makes the last hypothesis of this work that *bacterial functional amyloids in the human intestine may influence neurodegeneration* accepted, and the corresponding goal *to computationally identify bacterial functional amyloids in the human gut proteome and their potential interactions with human proteins* realized.

Results presented in this chapter are available as a preprint: ***Wojciechowska, A. W.****, Wojciechowski, J. W., Zielinska, K., Soeding, J., Kosciolek, T., Kotulska, M. (2024). Aggregating gut: on the link between neurodegeneration and bacterial functional amyloids. bioRxiv. Submitted to the journal.*

Discussion and conclusion

6.0.1 Discussion

The medical and social revolution of the past century has completely changed our life expectancy. Unfortunately, the longer we live, the more likely we are to experience a loss of cognitive ability that may cast a shadow over the autumn of our lives. Neurodegenerative diseases and dementia pose a significant burden on the quality of life of millions of seniors and their families around the world, estimated to cost billions of euros each year only in Europe [277]. However, no effective treatment is available for these disorders.

Alzheimer's and Parkinson's disease are the most common neurodegenerative disorders. Both diseases are characterised by amyloid deposits, which are present in the brains of the patients. Amyloids are proteins or peptides that can form insoluble beta-sheet-rich regular structures via the aggregation process. Their presence in the course of neurodegeneration is found to be a hallmark of the disease. Hence, for decades, our neurodegeneration research focused on a few cases of amyloid proteins, such as Aβ, α-synuclein and tau. Although multiple studies pointed toward these proteins, exploring other scientific endeavours could shed new light on our understanding of these disorders. Therefore, *studying the understudied* in neurodegeneration was the main motif of this thesis.

We started by discussing the application of protein-protein interactions in neurodegeneration research. Protein-protein interaction systems are widely used to provide an overview of the disease process, identify molecular mechanisms underlying it, and propose drug targets. Their common representation as graph objects gives solid ground for the mathematical evaluation of such objects. The topological analysis of these systems was hypothesized to provide a new view of the molecular mechanisms of neurodegeneration.

The detailed analysis of three datasets of protein-protein interactions, regarding Parkinson's disease, cancer, and the entire human interactome, revealed that our scientific interests pose a significant bias on the available data. In consequence, a reliable and large-scale biological interpretation of such interactomes may not be possible. Chapter 2 showed that investigation of the topology of a protein-protein interaction network, which includes studying its centrality metrics, clusterisation, robustness, and performing failure cascade simulation, is strongly related to how the data were gathered. The disease networks, *Parkinson* and *Cancer* ones, were both built based on the literature. This had significant consequences, as the systems did not appear to be a random sample of the disease interactomes, but instead were biased in directions of the scientific interests. Therefore, it was suggested that the interpretation of such protein-protein interaction networks should be treated with caution. **In consequence, the first hypothesis that *Exploratory topological analysis of available protein-protein interaction datasets can provide a new perspective on neurodegeneration* was not accepted.**

A set of topological metrics was provided to help other researchers assess the quality of their protein-protein interaction networks. Investigation of the degree distribution should help identify how impactful the hubs are. Testing how their removal decomposes the network and how quickly the error propagates in the network, when the hubs are

the starting points, gives an idea of their importance in the network system. Having similarly important hubs as in *Parkinson* network, one may suspect potential interest bias towards certain proteins which stand behind the network topology. On the other hand, the relatively low hubs' importance and low network clusterization, as in the case of the Human Reference Interactome (*HuRI*), suggest that some data in the network may be missing. The applied topological analyses were encapsulated in one easy-to-use Colab notebook (an online programming environment) with a graphical user interface to help others in the quick topological analysis of their networks. **Hence, the first goal *to assess the quality of the real-world protein interaction networks through topological analysis and identify if the topological analysis can shed new light on disease mechanisms*** was realized.

The interest biases ruling the neurodegenerative studies were also observed in the graph of amyloid interactions, Amylograph [65]. Amyloid proteins can affect the rate of aggregation of one another and can even form fibrils that consist of both types of proteins [96]. Amyloid interactions are considered to play an important role in the comorbidity of amyloid-related diseases. However, the amyloid interaction network is also biased by community interests. The majority of links refer to the famous Aβ-42 peptide, related to Alzheimer's disease, while other amyloid proteins were not studied in this context at all.

Our data on neurodegeneration focus on the proteins of greatest interest for scientists. Although such proteins might be the most important ones for systems, it does not have to be so. Chapter 2 demonstrates that, so far, we have not obtained the full picture of neurodegeneration, and investigation of other topics could shed new light on the pathology of the disease. This motivated all the following chapters, which focused on microbiome-related bacterial functional amyloids.

The growing amount of research gives evidence that the microbiome plays a crucial role in our health. Our intestinal bacteria produce a wide variety of molecular entities that interact, among others, with our immune system, shaping our response to the environment. Multiple studies have shown that the microbiome of neurodegeneration patients and their healthy controls differs in taxonomy and related functionality of the bacteria.

One of the potentially interesting, though understudied, routes of gut-brain interactions points to microbial amyloids. These proteins are crafted by evolution to produce amyloid fibrils similar to those of pathological amyloids. The presence of pathological amyloids, such as α-synuclein or Aβ-42, is often cytotoxic. Functional amyloids, including microbial amyloids, form fibrils that the cell uses, for example, to protect itself from the environment. The similarity between pathological and functional amyloids prompted an interesting discussion about their role in neurodegeneration.

It was suggested that microbial amyloids could be able to activate similar undesired molecular pathways as pathological amyloids and interact with pathological amyloids, promoting their aggregation in the enteric nervous system, particularly in the case of Parkinson's disease due to molecular mimicry. The molecular studies provided evidence for the rationale of this hypothesis. However, they often focused on a certain group of microbial amyloids, not investigating them from a wider proteome-level perspective, which could give a large-scale view of the role of these proteins in neurodegeneration.

The knowledge of bacterial functional amyloids was dispersed in the literature. Hence, the first result of Chapter 3 is a dataset of these proteins based on meta-analysis. Indeed, such a dataset can also be biased by scientific interests, similarly to protein-protein interaction data, but bacterial functional amyloids are themselves an understudied topic in neurodegeneration, which should be enough to provide new insights.

The detailed analysis of the proteins in the dataset of bacterial functional amyloids revealed that many of the known examples of these proteins are related to biofilm formation, although cases of proteins with other functionalities, for example, DNA or RNA binding, are also present.

The sequence analysis of the dataset of bacterial functional amyloids focused on tandem repeats, regions of repeated sequences of amino acids in the protein sequence. We observed that bacterial functional amyloids are enriched in repeated regions. The repeats were often imperfect and consisted of a rather small number of units. They also had a specific amino acid profile. The simulations revealed that changing repeat characteristics in a protein potentially changes the stability of the fibril. The discovered characteristics of the repeats suggested that the occurrence of repeats in bacterial functional amyloids is not accidental. Repeats in different proteins often mediate various interactions inter- and intra-protein interactions [278]. Bacterial functional amyloids are no exceptions to this pattern. It seems that the repeated regions are not necessarily the drivers of the aggregation, but rather their regulators. The presence of repeated regions introduces multiple symmetric interactions in the protein structure, allowing better control of the aggregation. Multiple symmetrical interactions should also impact interactions of bacterial functional amyloids with other proteins, as repeated motifs are often stabilizing the heterocomplexes [278]. **As a result, the second hypothesis that *sequence tandem repeats influence the aggregation of bacterial functional amyloids* was accepted, and the corresponding goal *to examine the role of tandem repeats in bacterial functional amyloids* was realized.**

It can be argued that the enrichment of the specific types of repeats in bacterial functional amyloids could be related to a low number of known such proteins. Even if it is the case, identification of the repeats in a newly discovered microbial amyloid should provide insight into the aggregation mechanism, as repeats are likely to interfere with fibril formation.

Bioinformatics analysis of a protein is not complete without structural information. AlphaFold revolutionized protein structure prediction, allowing fast and accurate modelling. Given this spectacular breakthrough, the hope to predict, with this tool, amyloid fibrils for bacterial functional amyloids arose. In Chapter 4, AlphaFold was tested on a set of examples of amyloid proteins revealing that it is still not capable of predicting amyloid fibrils. The main reason for the problem was a low abundance of fibrillar structures in the PDB. When AlphaFold deals with an amyloid protein, it often prefers to produce a globular structure similar to a seen homolog during the training and give it a high score, instead of predicting a fibrillar model which does not resemble the training examples. Some improvement in the performance was observed for shorter sequences. **As the results were rather disappointing, the third hypothesis that *structure of fibrils of bacterial onal amyloids coids can be predwithd with AlphaFold 3* was rejected. The third goal *to investigate AlphaFold 3 performance on amyloid proteins, including bacterial functional amyloids* was realized.**

We discussed the potential clinical relevance of bacterial functional amyloids in the human gut microbiome in neurodegeneration as the final step of the analysis of these proteins. To search for bacterial functional amyloids in the human gut microbiome proteome, a pipeline for a proteome-wide screen based on the identification of homologs of known bacterial functional amyloids and amyloid predictors was proposed. With this method, 805 potential microbial amyloids were found in the microbiome proteome. Compared to the total number of microbial genes in the human gut microbiome, this number may seem

modest. However, the search was limited by our current knowledge of bacterial functional amyloids and amyloid predictors, which were trained on pathological amyloids. Consequently, multiple microbial amyloids were probably missed. In the light of these facts, the 805 identified bacterial functional amyloids, which were homologs of only 38 proteins, seem to be a reasonable number.

The predicted bacterial functional amyloids spanned the bacterial tree of life, reaching beyond the bacterial species known to produce functional amyloids. This highlights that the universe of bacterial functional amyloids is larger than we might suspect. The abundance of bacterial functional amyloids was shown to be larger in the case of metagenomic samples from patients with Parkinson's disease than in healthy controls. In the case of Alzheimer's disease, the overall abundance of functional bacterial amyloids in patients was not significantly different from that of the controls. However, the abundance of homologs of a specific bacterial functional amyloids, CsgA, was different in the case of patients with Alzheimer's disease and Parkinson's disease.

The analysis of samples from patients with neurodegeneration demonstrates that not only are bacterial functional amyloids widely present in our microbiome, but they also seem to correlate with the state of the disease. To propose the mechanism of influence of bacterial functional amyloids on human health, their interactions with human proteins were predicted. Multiple molecular pathways are suspected to be affected by bacterial functional amyloids, primarily including the inflammatory response, but also transport, cell signalling, and cell junctions. Although we evolved to live with microbes and tolerate them, in the course of neurodegeneration that is often accompanied by intestinal dysbiosis, the ability to interact with our immune system by bacterial functional amyloids could have a further negative effect on our health. Bacterial functional amyloids could also affect the rates of aggregation of pathological amyloids in the enteric nervous system and hence play an important role in the onset and progression of neurodegeneration. **Therefore, the fourth hypothesis that *bacterial functional amyloids in the human intestine may influence neurodegeneration* was accepted, and the corresponding goal *to computationally identify bacterial functional amyloids in the human gut proteome and their potential interactions with human proteins* was realized.**

6.0.2 Conclusion

Our scientific interests shape our view of biological processes. As trivial as it may seem, life is a complex thing, and our models are as good as the data that we have. At this point, our data about biology is biased. We gather more knowledge about processes that interest us and less about the unpopular ones. As a result, we do not learn about all aspects of biology in the same uniform way; to do so, we would have to choose the topics that we study in a random way. Only then would our knowledge be a uniform sample of the biology that remains behind. Such unbiased data could lead to better generalisation of biology and unveil novel yet unexplored ideas and solutions. However, it is only a purely theoretical concept, as convincing a scientist (and funders) to randomly pick a topic instead of the most fascinating or promising one is against human nature. Maybe, with the rise of artificial intelligence capable of doing science, this would be possible, but still, the limitation of financial resources would remain.

The neurodegeneration research suffers from the problem of biased data, which is a consequence of our interests. We know a lot about a couple of pathological amyloids, as they are prominent hallmarks of the disease that caught our attention. As important as

they are, focusing on pathological amyloids reduces our chances of finding novel, promising directions, which are needed, as currently, no cures for Alzheimer's and Parkinson's diseases are known. Choosing to study bacterial functional amyloids in this thesis and their role in neurodegeneration was also biased by scientific interests. However, it seems that any step away from the beaten track can give us a new perspective on the biology of disease, and in this case, on neurodegeneration.

Exploring the characteristics of bacterial functional amyloids was largely limited by the scarcity of data about them. The 38 examples of bacterial functional amyloids do not seem to be a lot to uncover general trends about these proteins. However, detailed analyses revealed an interesting characteristic of the frequent sequence repeats that could regulate their aggregation. The low abundance of data was particularly problematic in the case of structure prediction. Because AlphaFold was trained on mostly globular proteins, it struggles to differentiate when an amyloid fold should be predicted and when not. Despite that, with the analysis we could understand that globular models for amyloid proteins are related to the AlphaFold problem with predicting different folds for homologous proteins. Estimating the scale of bacterial functional amyloids in the microbiome was difficult, as few examples are known, and amyloid predictors do not adjust to such proteins. In spite of this, we could see that bacterial functional amyloids are more common than expected in the microbiome, relate to the disease, and due to mechanisms of molecular mimicry can have negative effect on neurodegeneration.

Computational analyses are fueled by experimental data. By studying bacterial functional amyloids, we clearly see that we still lack experimental methods to perform large-scale studies on amyloid proteins which limits our development of bioinformatics methods. As the data regards only pathological amyloids, computational methods mostly focus on these proteins as well, helping to study the pathological amyloids more. This leads to the interest loop that may be broken by high-throughput technologies which could guide the discovery of novel amyloid sequences and structures, and not only aggregation-prone regions, across different proteomes and push research forward in the future. Until this happens, this thesis demonstrates that studying even the scarce data on a seemingly niche topic can give a new and valuable insight for the broader context of human diseases. It seems that at this point our knowledge of biology is highly concentrated around our scientific interests. Such topics are well-studied and deepen our understanding of molecular mechanisms. But at the same time they are our local minima that do not allow us for full exploration of the biological landscape. To cross these knowledge barriers, we need to step into a new era of studies that focuses on large-scale and high-throughput methods that may eliminate the biases that we see in our biological data. And this is taking place on our eyes...

List of publications

The following publications and preprints are related to this thesis:

1. **Nowakowska, A. W.,** & Kotulska, M. (2022). Topological analysis as a tool for detection of abnormalities in protein–protein interaction data. *Bioinformatics*, *38*(16), 3968-3975.

2. **Nowakowska, A. W.**, Wojciechowski, J. W., Szulc, N., & Kotulska, M. (2023). The role of tandem repeats in bacterial functional amyloids. *Journal of Structural Biology*, *215*(3), 108002.

3. **Wojciechowska, A. W.**, Wojciechowski, J. W., & Kotulska, M. (2024). Non-standard proteins in the lenses of AlphaFold3-case study of amyloids. *bioRxiv*

4. **Wojciechowska, A. W.**, Wojciechowski, J. W., Zielinska, K., Soeding, J., Kosciolek, T., & Kotulska, M. (2024). Aggregating gut: on the link between neurodegeneration and bacterial functional amyloids. *bioRxiv*.

Other publications that I coauthored during the course of my PhD thesis are the following:

1. Guerlais, V., Allouch, N., Moseman, E. A., **Wojciechowska, A. W.**, Wojciechowski, J. W., & Marcelino, I. (2024). Transcriptomic profiling of "brain-eating amoeba" Naegleria fowleri infection in mice: the host and the protozoa perspectives. *Frontiers in Cellular and Infection Microbiology*, *14*, 1490280.

2. Kalitnik, A., Szefczyk, M., **Wojciechowska, A. W.**, Wojciechowski, J. W., Gąsior-Głogowska, M., Olesiak-Bańska, J., & Kotulska, M. (2024). Cytotoxic Staphylococcus aureus PSM3 inhibits the aggregation of human insulin in vitro. *Physical Chemistry Chemical Physics*, *26*(21), 15587-15599.

3. Polańska, O., Szulc, N., Dyrka, W., **Wojciechowska, A. W.**, Kotulska, M., Żak, A. M., Gąsior-Głogowska M.E. & Szefczyk, M. (2025). Environmental sensitivity of amyloidogenic motifs in fungal NOD-like receptor-mediated immunity: Molecular and structural insights into amyloid assembly. *International Journal of Biological Macromolecules*, 140773.

4. Kalitnik, O., Lassota, A., Polańska, O., Gąsior-Głogowska, M., Szefczyk, M., Barbach, A., Chilimoniuk, J., Jęśkowiak-Kossakowska, I., **Wojciechowska, A. W.**, Wojciechowski, J.W., Szulc, N., Kotulska, M., Burdukiewicz, M. (2025). Experimental methods for studying amyloid cross-interactions, *Protein Science*, 6(24), e70151.

References

[1] Frederico AC Azevedo, Ludmila RB Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata EL Ferretti, Renata EP Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513(5):532–541, 2009.

[2] Alzheimer's Association. Alzheimer's disease facts and figures.

[3] MC De Rijk, MMB Breteler, GA Graveland, A Ott, DE Grobbee, FGA Van der Meche, and A Hofman. Prevalence of parkinson's disease in the elderly: the rotterdam study. *Neurology*, 45(12):2143–2146, 1995.

[4] Michael G Erkkinen, Mee-Ohk Kim, and Michael D Geschwind. Clinical neurology and epidemiology of the major neurodegenerative diseases. *Cold Spring Harbor perspectives in biology*, 10(4):a033118, 2018.

[5] Lary C Walker and Mathias Jucker. The exceptional vulnerability of humans to alzheimer's disease. *Trends in molecular medicine*, 23(6):534–545, 2017.

[6] Eric Heuer, Rebecca F. Rosen, Amarallys Cintron, and Lary C. Walker. Nonhuman primate models of alzheimer-like cerebral proteopathy. *Current pharmaceutical design*, 18(8):1159–1169, 2012.

[7] Kartik Pattabiraman, Sydney Keaton Muchnik, and Nenad Sestan. The evolution of the human brain and disease susceptibility. *Current opinion in genetics & development*, 65:91–97, 2020.

[8] Nala Rogers. Alzheimer's origins tied to rise of human intelligence. *Nature. doi*, 10, 2015.

[9] Emiliano Bruner and Heidi IL Jacobs. Alzheimer's disease: the downside of a highly evolved parietal lobe? *Journal of Alzheimer's Disease*, 35(2):227–240, 2013.

[10] Nico J Diederich, Toshiki Uchihara, Sten Grillner, and Christopher G Goetz. The evolution-driven signature of parkinson's disease. *Trends in Neurosciences*, 43(7):475–492, 2020.

[11] Nico J Diederich, D James Surmeier, Toshiki Uchihara, Sten Grillner, and Christopher G Goetz. Parkinson's disease: Is it a consequence of human brain evolution? *Movement Disorders*, 34(4):453, 2019.

[12] Liyong Wu, Pedro Rosa-Neto, Ging-Yuek R Hsiung, A Dessa Sadovnick, Mario Masellis, Sandra E Black, Jianping Jia, and Serge Gauthier. Early-onset familial alzheimer's disease (eofad). *Canadian Journal of Neurological Sciences*, 39(4):436–445, 2012.

[13] Manabu Funayama, Kenya Nishioka, Yuanzhe Li, and Nobutaka Hattori. Molecular genetics of parkinson's disease: Contributions and global trends. *Journal of human genetics*, 68(3):125–130, 2023.

[14] Emma Nichols, Jaimie D Steinmetz, Stein Emil Vollset, Kai Fukutaki, Julian Chalek, Foad Abd-Allah, Amir Abdoli, Ahmed Abualhasan, Eman Abu-Gharbieh, Tayyaba Tayyaba Akram, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. *The Lancet Public Health*, 7(2):e105–e125, 2022.

[15] Caleb E Finch and Stanley M Burstein. Dementia in the ancient greco-roman world was minimally mentioned. *Journal of Alzheimer's Disease*, (Preprint):1–8, 2024.

[16] Monika Jankowska-Kieltyka, Adam Roman, and Irena Nalepa. The air we breathe: air pollution as a prevalent proinflammatory stimulus contributing to neurodegeneration. *Frontiers in Cellular Neuroscience*, 15:647643, 2021.

[17] Angelika Więckowska-Gacek, Anna Mietelska-Porowska, Małgorzata Wydrych, and Urszula Wojda. Western diet as a trigger of alzheimer's disease: From metabolic syndrome and systemic inflammation to neuroinflammation and neurodegeneration. *Ageing research reviews*, 70:101397, 2021.

[18] Mark Hamer and Yoichi Chida. Physical activity and risk of neurodegenerative disease: a systematic review of prospective evidence. *Psychological medicine*, 39(1):3–11, 2009.

[19] Brendon Stubbs, Li-Jung Chen, Chun-Yi Chang, Wen-Jung Sun, and Po-Wen Ku. Accelerometer-assessed light physical activity is protective of future cognitive ability: A longitudinal study among community dwelling older adults. *Experimental gerontology*, 91:104–109, 2017.

[20] Agusti Marfany, Cristina Sierra, Miguel Camafort, Monica Domenech, and Antonio Coca. High blood pressure, alzheimer disease and antihypertensive treatment. *Panminerva Medica*, 60(1):8–16, 2018.

[21] GO Skeie, B Muller, K Haugarvoll, JP Larsen, and OB Tysnes. Parkinson disease: associated disorders in the norwegian population based incident parkwest study. *Parkinsonism & Related Disorders*, 19(1):53–55, 2013.

[22] Giuseppe Tosto, Thomas D Bird, David A Bennett, Bradley F Boeve, Adam M Brickman, Carlos Cruchaga, Kelley Faber, Tatiana M Foroud, Martin Farlow, Alison M Goate, et al. The role of cardiovascular risk factors and stroke in familial alzheimer disease. *JAMA neurology*, 73(10):1231–1237, 2016.

[23] Jiao Luo, Jesper Qvist Thomassen, Céline Bellenguez, Benjamin Grenier-Boley, Itziar De Rojas, Atahualpa Castillo, Kayenat Parveen, Fahri Küçükali, Aude Nicolas, Oliver Peters, et al. Genetic associations between modifiable risk factors and alzheimer disease. *JAMA network open*, 6(5):e2313734–e2313734, 2023.

[24] Huan Song, Johanna Sieurin, Karin Wirdefeldt, Nancy L Pedersen, Catarina Almqvist, Henrik Larsson, Unnur A Valdimarsdóttir, and Fang Fang. Association of stress-related disorders with subsequent neurodegenerative diseases. *JAMA neurology*, 77(6):700–709, 2020.

[25] Charlotte Madore, Zhuoran Yin, Jeffrey Leibowitz, and Oleg Butovsky. Microglia, lifestyle stress, and neurodegeneration. *Immunity*, 52(2):222–240, 2020.

[26] Samuel Carvalho Dumith, Kevin Francisco Durigon Meneghini, and Lauro Miranda Demenech. Who are the individuals with the worst perceived quality of sleep? a population-based survey in southern brazil. *Preventive Medicine Reports*, 21:101288, 2021.

[27] Emily Simmonds, Kristin S Levine, Jun Han, Hirotaka Iwaki, Mathew J Koretsky, Nicole Kuznetsov, Faraz Faghri, Caroline Warly Solsberg, Artur Schuh, Lietsel Jones, et al. Sleep disturbances as risk factors for neurodegeneration later in life. *medRxiv*, 2023.

[28] Oscar L López and Steven T DeKosky. Clinical symptoms in alzheimer's disease. *Handbook of clinical neurology*, 89:207–216, 2008.

[29] Sigurlaug Sveinbjornsdottir. The clinical symptoms of parkinson's disease. *Journal of neurochemistry*, 139:318–324, 2016.

[30] David S Eisenberg and Michael R Sawaya. Structural studies of amyloid proteins at the molecular level. *Annual review of biochemistry*, 86(1):69–95, 2017.

[31] David Eisenberg and Mathias Jucker. The amyloid state of proteins in human diseases. *Cell*, 148(6):1188–1203, 2012.

[32] Claudio Soto. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nature Reviews Neuroscience*, 4(1):49–60, 2003.

[33] Qinxi Guo, Zilai Wang, Hongmei Li, Mary Wiese, and Hui Zheng. App physiological and pathophysiological functions: insights from animal models. *Cell research*, 22(1):78–89, 2012.

[34] Tania F Gendron and Leonard Petrucelli. The role of tau in neurodegeneration. *Molecular neurodegeneration*, 4:1–19, 2009.

[35] Mihael H Polymeropoulos, Christian Lavedan, Elisabeth Leroy, Susan E Ide, Anindya Dehejia, Amalia Dutra, Brian Pike, Holly Root, Jeffrey Rubenstein, Rebecca Boyer, et al. Mutation in the α-synuclein gene identified in families with parkinson's disease. *science*, 276(5321):2045–2047, 1997.

[36] Michael J Volles and Peter T Lansbury. Zeroing in on the pathogenic form of α-synuclein and its mechanism of neurotoxicity in parkinson's disease. *Biochemistry*, 42(26):7871–7878, 2003.

[37] Arjan Quist, Ivo Doudevski, Hai Lin, Rushana Azimova, Douglas Ng, Blas Frangione, Bruce Kagan, Jorge Ghiso, and Ratnesh Lal. Amyloid ion channels: a common structural link for protein-misfolding disease. *Proceedings of the National Academy of Sciences*, 102(30):10427–10432, 2005.

[38] Joseph B El Khoury, Kathryn J Moore, Terry K Means, Josephine Leung, Kinya Terada, Michelle Toft, Mason W Freeman, and Andrew D Luster. Cd36 mediates the innate host response to β-amyloid. *The Journal of experimental medicine*, 197(12):1657–1666, 2003.

[39] Alessandra Webers, Michael T Heneka, and Paul A Gleeson. The role of innate immune responses and neuroinflammation in amyloid accumulation and progression of alzheimer's disease. *Immunology and cell biology*, 98(1):28–41, 2020.

[40] Changyoun Kim, Dong-Hwan Ho, Ji-Eun Suk, Sungyong You, Sarah Michael, Junghee Kang, Sung Joong Lee, Eliezer Masliah, Daehee Hwang, He-Jin Lee, et al. Neuron-released oligomeric α-synuclein is an endogenous agonist of tlr2 for paracrine activation of microglia. *Nature communications*, 4(1):1562, 2013.

[41] Yu Yamazaki, Mitsuru Shinohara, Motoko Shinohara, Akari Yamazaki, Melissa E Murray, Amanda M Liesinger, Michael G Heckman, Elizabeth R Lesser, Joseph E Parisi, Ronald C Petersen, et al. Selective loss of cortical endothelial tight junction proteins during alzheimer's disease progression. *Brain*, 142(4):1077–1092, 2019.

[42] Chia-Chen Liu, Yu Yamazaki, Michael G Heckman, Yuka A Martens, Lin Jia, Akari Yamazaki, Nancy N Diehl, Jing Zhao, Na Zhao, Michael DeTure, et al. Tau and apolipoprotein e modulate cerebrovascular tight junction integrity independent of cerebral amyloid angiopathy in alzheimer's disease. *Alzheimer's & Dementia*, 16(10):1372–1383, 2020.

[43] Wei-Li Kuan, Neal Bennett, Xiaoling He, Jeremy N Skepper, Nataly Martynyuk, Ruwani Wijeyekoon, Prabhas V Moghe, Caroline H Williams-Gray, and Roger A Barker. α-synuclein pre-formed fibrils impair tight junction protein expression without affecting cerebral endothelial cell function. *Experimental neurology*, 285:72–81, 2016.

[44] Hamdam Hourfar, Farhang Aliakbari, Shabboo Rahimi Aqdam, Zahra Nayeri, Hassan Bardania, Daniel E Otzen, and Dina Morshedi. The impact of α-synuclein aggregates on blood-brain barrier integrity in the presence of neurovascular unit cells. *International Journal of Biological Macromolecules*, 229:305–320, 2023.

[45] Magdalini Polymenidou and Don W Cleveland. Prion-like spread of protein aggregates in neurodegeneration. *Journal of Experimental Medicine*, 209(5):889–893, 2012.

[46] David W Sanders, Sarah K Kaufman, Sarah L DeVos, Apurwa M Sharma, Hilda Mirbaha, Aimin Li, Scarlett J Barker, Alex C Foley, Julian R Thorpe, Louise C Serpell, et al. Distinct tau prion strains propagate in cells and mice and define different tauopathies. *Neuron*, 82(6):1271–1288, 2014.

[47] Zane Jaunmuktane, Annelies Quaegebeur, Ricardo Taipa, Miguel Viana-Baptista, Raquel Barbosa, Carolin Koriath, Raf Sciot, Simon Mead, and Sebastian Brandner. Evidence of amyloid-β cerebral amyloid angiopathy transmission through neurosurgery. *Acta neuropathologica*, 135:671–679, 2018.

[48] Stanley B Prusiner, Amanda L Woerman, Daniel A Mordes, Joel C Watts, Ryan Rampersaud, David B Berry, Smita Patel, Abby Oehler, Jennifer K Lowe, Stephanie N Kravitz, et al. Evidence for α-synuclein prions causing multiple system atrophy in humans with parkinsonism. *Proceedings of the National Academy of Sciences*, 112(38):E5308–E5317, 2015.

[49] Thomas G. Beach, Charles H. Adler, Lihfen Lue, Lucia I. Sue, Jyothi Bachalakuri, Jonette Henry-Watson, Jeanne Sasse, Sarah Boyer, Scophil Shirohi, Reed Brooks, Jennifer Eschbacher, Charles L. White, Haru Akiyama, John Caviness, Holly A. Shill, Donald J. Connor, and Marwan N. Sabbagh. Unified staging system for lewy body disorders: correlation with nigrostriatal degeneration, cognitive impairment and motor dysfunction. *Acta Neuropathologica*, 117(6):613–634, apr 2009.

[50] Chris Mezias and Ashish Raj. Analysis of amyloid- pathology spread in mouse models suggests spread is driven by spatial proximity, not connectivity. *Frontiers in Neurology*, 8, 2017.

[51] Kelly Servick. Another major drug candidate targeting the brain plaques of alzheimer's disease has failed. what's left? *Science*, mar 2019.

[52] Orestes Vicente Forlenza and Breno José Alencar Pires Barbosa. What are the reasons for the repeated failures of clinical trials with anti-amyloid drugs for ad treatment?, 2025.

[53] Gennaro Pagano, Kirsten I. Taylor, Judith Anzures-Cabrera, Maddalena Marchesi, Tanya Simuni, Kenneth Marek, Ronald B. Postuma, Nicola Pavese, Fabrizio Stocchi, Jean-Philippe Azulay, Brit Mollenhauer, Lydia López-Manzanares, David S. Russell, James T. Boyd, Anthony P. Nicholas, María R. Luquin, Robert A. Hauser, Thomas Gasser, Werner Poewe, Benedicte Ricci, Anne Boulay, Annamarie Vogt, Frank G. Boess, Jürgen Dukart, Giulia D'Urso, Rebecca Finch, Stefano Zanigni, Annabelle Monnet, Nathalie Pross, Andrea Hahn, Hanno Svoboda, Markus Britschgi, Florian Lipsmeier, Ekaterina Volkova-Volkmar, Michael Lindemann, Sebastian Dziadek, Štefan Holiga, Daria Rukina, Thomas Kustermann, Geoffrey A. Kerchner, Paulo Fontoura, Daniel Umbricht, Rachelle Doody, Tania Nikolcheva, and Azad Bonni. Trial of Prasinezumab in Early-Stage Parkinson's Disease. *The New England journal of medicine*, 387(5):421 – 432, 2022.

[54] Jorik Nonnekes, Monique H M Timmer, Nienke M de Vries, Olivier Rascol, Rick C Helmich, and Bastiaan R Bloem. Unmasking levodopa resistance in parkinson's disease. *Movement disorders : official journal of the Movement Disorder Society*, 31(11):1602—1609, November 2016.

[55] Angelo Antonini, Aron Emmi, and Marta Campagnolo. Beyond the dopaminergic system: lessons learned from levodopa resistant symptoms in parkinson's disease. *Movement disorders clinical practice*, 10(Suppl 2):S50, 2023.

[56] Rutvi Prajapati and Isaac Arnold Emerson. Gene prioritization in parkinson's disease using human protein–protein interaction network. *Journal of Computational Biology*, 27(11):1610–1621, 2020.

[57] Sreedevi Chandrasekaran and Danail Bonchev. A network view on parkinson's disease. *Computational and structural biotechnology journal*, 7(8):e201304004, 2013.

[58] Avijit Podder, Mansi Pandit, and Latha Narayanan. Drug target prioritization for alzheimer's disease using protein interaction network analysis. *OMICS: A Journal of Integrative Biology*, 22(10):665–677, 2018.

[59] Hindol Rakshit, Nitin Rathi, and Debjani Roy. Construction and analysis of the protein-protein interaction networks based on gene expression profiles of parkinson's disease. *PloS one*, 9(8):e103047, 2014.

[60] Jun Shen, Xiao-Chang Chen, Wang-Jun Li, Qiu Han, Chun Chen, Jing-Min Lu, Jin-Yu Zheng, and Shou-Ru Xue. Identification of parkinson's disease-related pathways and potential risk factors. *Journal of International Medical Research*, 48(10):0300060520957197, 2020.

[61] Negar Sadat Soleimani Zakeri, Saeid Pashazadeh, and Habib MotieGhader. Drug repurposing for alzheimer's disease based on protein-protein interaction network. *BioMed research international*, 2021(1):1280237, 2021.

[62] Ying Hu, Yichen Yang, Zhonghai Fang, Yan-Shi Hu, Lei Zhang, and Ju Wang. Detecting pathway relationship in the context of human protein-protein interaction network and its application to parkinson's disease. *Methods*, 131:93–103, 2017.

[63] Harriet Keane, Brent J Ryan, Brendan Jackson, Alan Whitmore, and Richard Wade-Martins. Protein-protein interaction networks identify targets which rescue the mpp+ cellular model of parkinson's disease. *Scientific reports*, 5(1):17004, 2015.

[64] Konstantina V Biza, Katerina C Nastou, Paraskevi L Tsiolaki, Chara V Mastrokalou, Stavros J Hamodrakas, and Vassiliki A Iconomidou. The amyloid interactome: exploring protein aggregation. *PloS one*, 12(3):e0173163, 2017.

[65] Michał Burdukiewicz, Dominik Rafacz, Agnieszka Barbach, Katarzyna Hubicka, Laura Bąkała, Anna Lassota, Jakub Stecko, Natalia Szymańska, Jakub W Wojciechowski, Dominika Kozakiewicz, et al. Amylograph: a comprehensive database of amyloid–amyloid interactions. *Nucleic Acids Research*, 51(D1):D352–D357, 2023.

[66] Jean Guy LeBlanc, Christian Milani, Graciela Savoy De Giori, Fernando Sesma, Douwe Van Sinderen, and Marco Ventura. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Current opinion in biotechnology*, 24(2):160–168, 2013.

[67] William Fusco, Manuel Bernabeu Lorenzo, Marco Cintoni, Serena Porcari, Emanuele Rinninella, Francesco Kaitsas, Elena Lener, Maria Cristina Mele, Antonio Gasbarrini, Maria Carmen Collado, et al. Short-chain fatty-acid-producing bacteria: key components of the human gut microbiota. *Nutrients*, 15(9):2211, 2023.

[68] Melanie Schirmer, Sanne P Smeekens, Hera Vlamakis, Martin Jaeger, Marije Oosting, Eric A Franzosa, Rob Ter Horst, Trees Jansen, Liesbeth Jacobs, Marc Jan Bonder, et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell*, 167(4):1125–1136, 2016.

[69] Ivaylo I Ivanov and Kenya Honda. Intestinal commensal microbes as immune modulators. *Cell host & microbe*, 12(4):496–508, 2012.

[70] Aafke WF Janssen and Sander Kersten. The role of the gut microbiota in metabolic health. *The FASEB Journal*, 29(8):3111–3123, 2015.

[71] Lea Ann Chen and Kaitlyn Boyle. The role of the gut microbiome in health and disease in the elderly. *Current gastroenterology reports*, pages 1–14, 2024.

[72] Annamaria Cattaneo, Nadia Cattane, Samantha Galluzzi, Stefania Provasi, Nicola Lopizzo, Cristina Festari, Clarissa Ferrari, Ugo Paolo Guerra, Barbara Paghera, Cristina Muscio, et al. Association of brain amyloidosis with pro-inflammatory gut bacterial taxa and peripheral inflammation markers in cognitively impaired elderly. *Neurobiology of aging*, 49:60–68, 2017.

[73] Fatemah Sadeghpour Heravi, Kaveh Naseri, and Honghua Hu. Gut microbiota composition in patients with neurodegenerative disorders (parkinson's and alzheimer's) and healthy controls: a systematic review. *Nutrients*, 15(20):4365, 2023.

[74] Rasoul Mirzaei, Behnaz Bouzari, Seyed Reza Hosseini-Fard, Maryam Mazaheri, Yaghoub Ahmadyousefi, Milad Abdi, Saba Jalalifar, Zahra Karimitabar, Ali Teimoori, Hossein Keyvani, et al. Role of microbiota-derived short-chain fatty acids in nervous system disorders. *Biomedicine & Pharmacotherapy*, 139:111661, 2021.

[75] Marcus M Unger, Jörg Spiegel, Klaus-Ulrich Dillmann, David Grundmann, Hannah Philippeit, Jan Bürmann, Klaus Faßbender, Andreas Schwiertz, and Karl-Herbert Schäfer. Short chain fatty acids and gut microbiota differ between patients with parkinson's disease and age-matched controls. *Parkinsonism & related disorders*, 32:66–72, 2016.

[76] Yuhai Zhao, Lin Cong, Vivian Jaber, and Walter J Lukiw. Microbiome-derived lipopolysaccharide enriched in the perinuclear region of alzheimer's disease brain. *Frontiers in Immunology*, 8:1064, 2017.

[77] Christin Bissig, Leila Rochin, and Guillaume Van Niel. Pmel amyloid fibril formation: the bright steps of pigmentation. *International journal of molecular sciences*, 17(9):1438, 2016.

[78] Daniel Otzen and Roland Riek. Functional amyloids. *Cold Spring Harbor perspectives in biology*, 11(12):a033860, 2019.

[79] Vassiliki A Iconomidou, Paul Cordopatis, Andreas Hoenger, and Stavros J Hamodrakas. The silkmoth eggshell as a natural amyloid shield for the safe development of insect oocyte and embryo: Insights from studies of silkmoth chorion protein peptide-analogues of the b famil. *Peptide Science*, 96(6):723–733, 2011.

[80] Sven J Saupe. Amyloid signaling in filamentous fungi and bacteria. *Annual review of microbiology*, 74(1):673–691, 2020.

[81] Frank Gondelaud, Pierre-Yves Lozach, and Sonia Longhi. Viral amyloids: New opportunities for antiviral therapeutic strategies. *Current Opinion in Structural Biology*, 83:102706, 2023.

[82] Natalia Szulc, Marlena Gąsior-Głogowska, Jakub W Wojciechowski, Monika Szefczyk, Andrzej M Żak, Michał Burdukiewicz, and Malgorzata Kotulska. Variability of amyloid propensity in imperfect repeats of csga protein of salmonella enterica and escherichia coli. *International journal of molecular sciences*, 22(10):5127, 2021.

[83] Anu Iswarya Jaisankar, AS Smiline Girija, Shoba Gunasekaran, and J Vijayashree Priyadharsini. Molecular characterisation of csga gene among esbl strains of a. baumannii and targeting with essential oil compounds from azadirachta indica. *Journal of King Saud University-Science*, 32(8):3380–3387, 2020.

[84] Faezeh Moghadas Kasani, Ali Salehzadeh, and Amir Jalali. Relationship between flu and csga virulence genes and biofilm production in uropathogenic escherichia coli. *Journal of Advanced Biomedical Sciences*, 10(4):2775–2785, 2020.

[85] Diego Romero, Hera Vlamakis, Richard Losick, and Roberto Kolter. Functional analysis of the accessory protein tapa in bacillus subtilis amyloid fiber assembly. *Journal of bacteriology*, 196(8):1505–1513, 2014.

[86] Madhu Nagaraj, Mumdooh Ahmed, Jeppe Lyngsø, Brian Stougaard Vad, Andreas Bøggild, Anne Fillipsen, Jan Skov Pedersen, Daniel Erik Otzen, and Ümit Akbey. Predicted loop regions promote aggregation: a study of amyloidogenic domains in the functional amyloid fapc. *Journal of molecular biology*, 432(7):2232–2252, 2020.

[87] Agustina Taglialegna, Leticia Matilla-Cuenca, Pedro Dorado-Morales, Susanna Navarro, Salvador Ventura, James A Garnett, Iñigo Lasa, and Jaione Valle. The biofilm-associated surface protein esp of enterococcus faecalis forms amyloid-like fibers. *npj Biofilms and Microbiomes*, 6(1):15, 2020.

[88] Patrick Di Martino. Bap: a new type of functional amyloid. *Trends in microbiology*, 24(9):682–684, 2016.

[89] Emilie Fortas, Federica Piccirilli, Antoine Malabirade, Valeria Militello, Sylvain Trépout, Sergio Marco, Aziz Taghbalout, and Véronique Arluison. New insight into the structure and function of hfq c-terminus. *Bioscience Reports*, 35(2):e00190, 2015.

[90] Florian Turbant, Pengzhi Wu, Frank Wien, and Véronique Arluison. The amyloid region of hfq riboregulator promotes dsra: rpos rnas annealing. *Biology*, 10(9):900, 2021.

[91] Heiko Braak, U Rüb, WP Gai, and Kelly Del Tredici. Idiopathic parkinson's disease: possible routes by which vulnerable neuronal types may be subject to neuroinvasion by an unknown pathogen. *Journal of neural transmission*, 110:517–536, 2003.

[92] Thibaud Lebouvier, Michel Neunlist, Stanislas Bruley des Varannes, Emmanuel Coron, Anne Drouard, Jean-Michel N'Guyen, Tanguy Chaumette, Maddalena Tasselli, Sébastien Paillusson, Mathurin Flamand, et al. Colonic biopsies to assess the neuropathology of parkinson's disease and its relationship with symptoms. *PloS one*, 5(9):e12728, 2010.

[93] Christopher B Forsyth, Kathleen M Shannon, Jeffrey H Kordower, Robin M Voigt, Maliha Shaikh, Jean A Jaglin, Jacob D Estes, Hemraj B Dodiya, and Ali Keshavarzian. Increased intestinal permeability correlates with sigmoid mucosa alpha-synuclein staining and endotoxin exposure markers in early parkinson's disease. *PloS one*, 6(12):e28032, 2011.

[94] Robert P Friedland and Matthew R Chapman. The role of microbial amyloid in neurodegeneration. *PLoS pathogens*, 13(12):e1006654, 2017.

[95] Molly Elkins, Neha Jain, and Çagla Tükel. The menace within: bacterial amyloids as a trigger for autoimmune and neurodegenerative diseases. *Current Opinion in Microbiology*, 79:102473, 2024.

[96] Sushma Subedi, Santanu Sasidharan, Niharika Nag, Prakash Saudagar, and Timir Tripathi. Amyloid cross-seeding: Mechanism, implication, and inhibition. *Molecules*, 27(6):1776, 2022.

[97] Timothy R Sampson, Collin Challis, Neha Jain, Anastasiya Moiseyenko, Mark S Ladinsky, Gauri G Shastri, Taren Thron, Brittany D Needham, Istvan Horvath, Justine W Debelius, et al. A gut bacterial amyloid promotes α-synuclein aggregation and motor impairment in mice. *elife*, 9:e53111, 2020.

[98] Chenyin Wang, Chun Yin Lau, Fuqiang Ma, and Chaogu Zheng. Genome-wide screen identifies curli amyloid fibril as a bacterial component promoting host neurodegeneration. *Proceedings of the National Academy of Sciences*, 118(34):e2106504118, 2021.

[99] Line Friis Bakmann Christensen, Kirstine Friis Jensen, Janni Nielsen, Brian Stougaard Vad, Gunna Christiansen, and Daniel Erik Otzen. Reducing the amyloidogenicity of functional amyloid protein fapc increases its ability to inhibit α-synuclein fibrillation. *Acs Omega*, 4(2):4029–4039, 2019.

[100] Ibrahim Javed, Zhenzhen Zhang, Jozef Adamcik, Nicholas Andrikopoulos, Yuhuan Li, Daniel E Otzen, Sijie Lin, Raffaele Mezzenga, Thomas P Davis, Feng Ding, et al. Accelerated amyloid beta pathogenesis by bacterial amyloid fapc. *Advanced Science*, 7(18):2001299, 2020.

[101] Ariadna Fernández-Calvet, Leticia Matilla-Cuenca, María Izco, Susanna Navarro, Miriam Serrano, Salvador Ventura, Javier Blesa, Maite Herráiz, Gorka Alkorta-Aranburu, Sergio Galera, et al. Gut microbiota produces biofilm-associated amyloids with potential for neurodegeneration. *Nature Communications*, 15(1):4150, 2024.

[102] Masaaki Hirayama and Kinji Ohno. Parkinson's disease and gut microbiota. *Annals of Nutrition and Metabolism*, 77(Suppl. 2):28–35, 2021.

[103] Monica Bucciantini, Elisa Giannoni, Fabrizio Chiti, Fabiana Baroni, Lucia Formigli, Jesús Zurdo, Niccolò Taddei, Giampietro Ramponi, Christopher M Dobson, and Massimo Stefani. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *nature*, 416(6880):507–511, 2002.

[104] Nimrod Golan, Yizhaq Engelberg, and Meytal Landau. Structural mimicry in microbial and antimicrobial amyloids. *Annual Review of Biochemistry*, 91(1):403–422, 2022.

[105] Maren Rautenberg, Hwang-Soo Joo, Michael Otto, and Andreas Peschel. Neutrophil responses to staphylococcal pathogens and commensals via the formyl peptide receptor 2 relates to phenol-soluble modulin release and virulence. *The FASEB Journal*, 25(4):1254, 2011.

[106] Çagla Tükel, R Paul Wilson, Jessalyn H Nishimori, Milad Pezeshki, Brett A Chromy, and Andreas J Bäumler. Responses to amyloids of microbial and host origin are mediated through toll-like receptor 2. *Cell host & microbe*, 6(1):45–53, 2009.

[107] Glenn J Rapsinski, Meghan A Wynosky-Dolfi, Gertrude O Oppong, Sarah A Tursi, R Paul Wilson, Igor E Brodsky, and Çagla Tükel. Toll-like receptor 2 and nlrp3 cooperate to recognize a functional bacterial amyloid, curli. *Infection and immunity*, 83(2):693–701, 2015.

[108] Jacinte Beerten, Joost Schymkowitz, and Frederic Rousseau. Aggregation prone regions and gatekeeping residues in protein sequences. *Current topics in medicinal chemistry*, 12(22):2470–2478, 2012.

[109] Sebastian Maurer-Stroh, Maja Debulpaep, Nico Kuemmerer, Manuela Lopez De La Paz, Ivo Cristiano Martins, Joke Reumers, Kyle L Morris, Alastair Copland, Louise Serpell, Luis Serrano, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature methods*, 7(3):237–242, 2010.

[110] Oscar Conchillo-Solé, Natalia S de Groot, Francesc X Avilés, Josep Vendrell, Xavier Daura, and Salvador Ventura. Aggrescan: a server for the prediction and evaluation of" hot spots" of aggregation in polypeptides. *BMC bioinformatics*, 8:1–17, 2007.

[111] Rafael Zambrano, Michal Jamroz, Agata Szczasiuk, Jordi Pujols, Sebastian Kmiecik, and Salvador Ventura. Aggrescan3d (a3d): server for prediction of aggregation properties of protein structures. *Nucleic acids research*, 43(W1):W306–W313, 2015.

[112] Abdullah B Ahmed, Nadia Znassi, Marie-Thérèse Château, and Andrey V Kajava. A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's & Dementia*, 11(6):681–690, 2015.

[113] Michał Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Anna Duda-Madej, Paweł Mackiewicz, and Małgorzata Kotulska. Amyloidogenic motifs revealed by n-gram analysis. *Scientific reports*, 7(1):12961, 2017.

[114] Jakub W Wojciechowski and Małgorzata Kotulska. Path-prediction of amyloidogenicity by threading and machine learning. *Scientific Reports*, 10(1):7721, 2020.

[115] Pawel Gasior and Malgorzata Kotulska. Fish amyloid–a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC bioinformatics*, 15:1–8, 2014.

[116] Ryann Perez, Xinning Li, Sam Giannakoulias, and E James Petersson. Aggbert: best in class prediction of hexapeptide amyloidogenesis with a semi-supervised protbert model. *Journal of chemical information and modeling*, 63(18):5727–5733, 2023.

[117] Runtao Yang, Jiaming Liu, and Lina Zhang. Ecamyloid: An amyloid predictor based on ensemble learning and comprehensive sequence-derived features. *Computational Biology and Chemistry*, 104:107853, 2023.

[118] Phasit Charoenkwan, Saeed Ahmed, Chanin Nantasenamat, Julian MW Quinn, Mohammad Ali Moni, Pietro Lio', and Watshara Shoombuatong. Amypred-frl is a novel approach for accurate prediction of amyloid proteins by using feature representation learning. *Scientific reports*, 12(1):7697, 2022.

[119] Peleg Ragonis-Bachar and Meytal Landau. Functional and pathological amyloid structures in the eyes of 2020 cryo-em. *Current Opinion in Structural Biology*, 68:184–193, 2021.

[120] Dan Ofer, Nadav Brandes, and Michal Linial. The language of proteins: Nlp, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758, 2021.

[121] Ritesh Sharma, Sameer Shrivastava, Sanjay Kumar Singh, Abhinav Kumar, Sonal Saxena, and Raj Kumar Singh. Deep-abppred: identifying antibacterial peptides in protein sequences using bidirectional lstm with word2vec. *Briefings in Bioinformatics*, 22(5):bbab065, 2021.

[122] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20:1–17, 2019.

[123] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[124] Wayland Yeung, Zhongliang Zhou, Sheng Li, and Natarajan Kannan. Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. *Briefings in Bioinformatics*, 24(1):bbac599, 2023.

[125] Christian Dallago, Konstantin Schütze, Michael Heinzinger, Tobias Olenyi, Maria Littmann, Amy X Lu, Kevin K Yang, Seonwoo Min, Sungroh Yoon, James T Morton, et al. Learned embeddings from deep learning to visualize and predict protein sets. *Current Protocols*, 1(5):e113, 2021.

[126] Sujeet Bhoite, Nani Van Gerven, Matthew R Chapman, and Han Remaut. Curli biogenesis: bacterial amyloid assembly by the type viii secretion pathway. *EcoSal Plus*, 8(2):10–1128, 2019.

[127] Helena Ø Rasmussen, Amit Kumar, Ben Shin, Fisentzos Stylianou, Lee Sewell, Yingqi Xu, Daniel E Otzen, Jan Skov Pedersen, and Steve J Matthews. Fapa is an intrinsically disordered chaperone for pseudomonas functional amyloid fapc. *Journal of Molecular Biology*, 435(2):167878, 2023.

[128] Morten S Dueholm, Daniel Otzen, and Per Halkjær Nielsen. Evolutionary insight into the functional amyloids of the pseudomonads. *PLoS One*, 8(10):e76630, 2013.

[129] Morten S Dueholm, Mads Albertsen, Daniel Otzen, and Per Halkjær Nielsen. Curli functional amyloid systems are phylogenetically widespread and display large diversity in operon and protein structure. *PloS one*, 7(12):e51274, 2012.

[130] Caleen B Ramsook, Cho Tan, Melissa C Garcia, Raymond Fung, Gregory Soybelman, Ryan Henry, Anna Litewka, Shanique O'Meally, Henry N Otoo, Roy A Khalaf, et al. Yeast cell adhesion molecules have functional amyloid-forming sequences. *Eukaryotic cell*, 9(3):393–404, 2010.

[131] Emma Lacroix, Lionel Pereira, Byoungjoo Yoo, Krysta M Coyle, Sahil Chandhok, Richard Zapf, Dane Marijan, Ryan D Morin, Stephanie Vlachos, Nicholas Harden, et al. Evolutionary conservation of systemic and reversible amyloid aggregation. *Journal of cell science*, 134(22):jcs258907, 2021.

[132] Michael R Sawaya, Michael P Hughes, Jose A Rodriguez, Roland Riek, and David S Eisenberg. The expanding amyloid family: Structure, stability, function, and pathogenesis. *Cell*, 184(19):4857–4873, 2021.

[133] Frank Shewmaker, Ryan P McGlinchey, and Reed B Wickner. Structural insights into functional and pathological amyloid. *Journal of Biological Chemistry*, 286(19):16533–16540, 2011.

[134] Matteo Delucchi, Elke Schaper, Oxana Sachenkova, Arne Elofsson, and Maria Anisimova. A new census of protein tandem repeats and their relationship with intrinsic disorder. *Genes*, 11(4):407, 2020.

[135] Kasper Holst Hansen, Chang Hyeock Byeon, Qian Liu, Taner Drace, Thomas Boesen, James F Conway, Maria Andreasen, and Ümit Akbey. Structure of biofilm-forming functional amyloid psmα1 from staphylococcus aureus. *Proceedings of the National Academy of Sciences*, 121(33):e2406775121, 2024.

[136] Fan Bu, Derek R Dee, and Bin Liu. Structural insight into escherichia coli csga amyloid fibril assembly. *MBio*, 15(4):e00419–24, 2024.

[137] Hamed Tabatabaei Ghomi, Elizabeth M Topp, and Markus A Lill. Fibpredictor: a computational method for rapid prediction of amyloid fibril structures. *Journal of molecular modeling*, 22:1–10, 2016.

[138] Stanislav A Bondarev, Olga V Bondareva, Galina A Zhouravleva, and Andrey V Kajava. Betaserpentine: a bioinformatics tool for reconstruction of amyloid structures. *Bioinformatics*, 34(4):599–608, 2018.

[139] Liangyue Guo, Qilin Yu, Di Wang, Xiaoyu Wu, Peter G Wolynes, and Mingchen Chen. Generating the polymorph landscapes of amyloid fibrils using ai: Ribbonfold. *Proceedings of the National Academy of Sciences*, 122(16):e2501321122, 2025.

[140] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xv. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1539–1549, 2023.

[141] Alessia David, Suhail Islam, Evgeny Tankhilevich, and Michael JE Sternberg. The alphafold database of protein structures: a biologist's guide. *Journal of molecular biology*, 434(2):167336, 2022.

[142] Feng Ren, Xiao Ding, Min Zheng, Mikhail Korzinkin, Xin Cai, Wei Zhu, Alexey Mantsyzov, Alex Aliper, Vladimir Aladinskiy, Zhongying Cao, et al. Alphafold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel cdk20 small molecule inhibitor. *Chemical Science*, 14(6):1443–1452, 2023.

[143] Andrius Bernatavicius, Martin Sicho, Antonius PA Janssen, Alan Kai Hassen, Mike Preuss, and Gerard JP van Westen. Alphafold meets de novo drug design: Leveraging structural protein information in multitarget molecular generative models. *Journal of Chemical Information and Modeling*, 64(21):8113–8122, 2024.

[144] Chenguang Zhao, Tong Liu, and Zheng Wang. Panda-3d: protein function prediction based on alphafold models. *NAR Genomics and Bioinformatics*, 6(3):lqae094, 2024.

[145] David F Burke, Patrick Bryant, Inigo Barrio-Hernandez, Danish Memon, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Alistair S Dunham, Pascal Albanese, Andrew Keller, et al. Towards a structurally resolved human protein interaction network. *Nature Structural & Molecular Biology*, 30(2):216–225, 2023.

[146] Fabian Hosp, Hannes Vossfeldt, Matthias Heinig, Djordje Vasiljevic, Anup Arumughan, Emanuel Wyler, Markus Landthaler, Norbert Hubner, Erich E Wanker, Lars Lannfelt, et al. Quantitative interaction proteomics of neurodegenerative disease proteins. *Cell reports*, 11(7):1134–1146, 2015.

[147] Ted G Lewis. *Network science: Theory and applications*. John Wiley & Sons, 2011.

[148] Márton Pósfai and Albert-László Barabási. *Network science*, volume 3. Citeseer, 2016.

[149] Swami Iyer, Timothy Killingback, Bala Sundaram, and Zhen Wang. Attack robustness and centrality of complex networks. *PloS one*, 8(4):e59613, 2013.

[150] Caroline C Friedel and Ralf Zimmer. Influence of degree correlations on network structure and stability in protein-protein interaction networks. *BMC bioinformatics*, 8:1–10, 2007.

[151] Tord Berggård, Sara Linse, and Peter James. Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842, 2007.

[152] Kumar Yugandhar, Shagun Gupta, and Haiyuan Yu. Inferring protein-protein interaction networks from mass spectrometry-based proteomic approaches: a mini-review. *Computational and structural biotechnology journal*, 17:805–811, 2019.

[153] Colin S Gillespie. Fitting heavy tailed distributions: the powerlaw package. *Journal of Statistical Software*, 64:1–16, 2015.

[154] Michel L Goldstein, Steven A Morris, and Gary G Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41:255–258, 2004.

[155] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[156] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[157] Pauli Virtanen, Ralf Gommers, Evgeni Burovski, Travis E Oliphant, Warren Weckesser, David Cournapeau, Pearu Peterson, Tyler Reddy, Matt Haberland, Josh Wilson, et al. scipy/scipy: Scipy 1.6. 0. *Zenodo*, 2021.

[158] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[159] Michael Waskom, Olga Botvinnik, Paul Hobson, Jordi Warmenhoven, John B Cole, Yaroslav Halchenko, Jake Vanderplas, Stephan Hoyer, Santi Villalba, Eric Quintero, et al. seaborn: v0. 6.0 (june 2015). *Zenodo*, 2015.

[160] Tiago P Peixoto. The graph-tool python library. *(No Title)*, 2017.

[161] Hadley Wickham. ggplot2. *Wiley interdisciplinary reviews: computational statistics*, 3(2):180–185, 2011.

[162] Noemi Del Toro, Anjali Shrivastava, Eliot Ragueneau, Birgit Meldal, Colin Combe, Elisabet Barrera, Livia Perfetto, Karyn How, Prashansa Ratan, Gautam Shirodkar, et al. The intact database: efficient access to fine-grained molecular interaction data. *Nucleic acids research*, 50(D1):D648–D653, 2022.

[163] Hema Adhikari and Christopher M. Counter. Interrogating the protein interactomes of ras isoforms identifies pip5k1a as a kras-specific vulnerability. *Nature Communications*, 9(1), September 2018.

[164] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E. Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J. Campos-Laborie, Benoit Charloteaux, Dongsic Choi, Atina G. Coté, Meaghan Daley, Steven Deimling, Alice Desbuleux, Amélie Dricot, Marinella Gebbia, Madeleine F. Hardy, Nishka Kishore, Jennifer J. Knapp, István A. Kovács, Irma Lemmens, Miles W. Mee, Joseph C. Mellor, Carl Pollis, Carles Pons, Aaron D. Richardson, Sadie Schlabach, Bridget Teeking, Anupama Yadav, Mariana Babor, Dawit Balcha, Omer Basha, Christian Bowman-Colin, Suet-Feung Chin, Soon Gang Choi, Claudia Colabella, Georges Coppin, Cassandra D'Amata, David De Ridder, Steffi De Rouck, Miquel Duran-Frigola, Hanane Ennajdaoui, Florian Goebels, Liana Goehring, Anjali Gopal, Ghazal Haddad, Elodie Hatchi, Mohamed Helmy, Yves Jacob, Yoseph Kassa, Serena Landini, Roujia Li, Natascha van Lieshout, Andrew MacWilliams, Dylan Markey, Joseph N. Paulson, Sudharshan Rangarajan, John Rasla, Ashyad Rayhan, Thomas Rolland, Adriana San-Miguel, Yun Shen, Dayag Sheykhkarimli, Gloria M. Sheynkman, Eyal Simonovsky, Murat Taşan, Alexander Tejeda, Vincent Tropepe, Jean-Claude Twizere, Yang Wang, Robert J. Weatheritt, Jochen Weile, Yu Xia, Xinping Yang, Esti Yeger-Lotem, Quan Zhong, Patrick Aloy, Gary D. Bader, Javier De Las Rivas, Suzanne Gaudet, Tong Hao, Janusz Rak, Jan Tavernier, David E. Hill, Marc Vidal, Frederick P. Roth, and Michael A. Calderwood. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, April 2020.

[165] Christian Haenig, Nir Atias, Alexander K Taylor, Arnon Mazza, Martin H Schaefer, Jenny Russ, Sean-Patrick Riechers, Shushant Jain, Maura Coughlin, Jean-Fred Fontaine, et al. Interactome mapping provides a network of neurodegenerative disease proteins and uncovers widespread protein aggregation in affected brains. *Cell reports*, 32(7), 2020.

[166] Susan A Kennedy, Mohamed-Ali Jarboui, Sriganesh Srihari, Cinzia Raso, Kenneth Bryan, Layal Dernayka, Theodosia Charitou, Manuel Bernal-Llinares, Carlos Herrera-Montavez, Aleksandar Krstic, et al. Extensive rewiring of the egfr network in colorectal cancer cells expressing transforming levels of krasg13d. *Nature communications*, 11(1):499, 2020.

[167] Rosalind F Roberts, Richard Wade-Martins, and Javier Alegre-Abarrategui. Direct visualization of alpha-synuclein oligomers reveals previously undetected pathology in parkinson's disease brain. *Brain*, 138(6):1642–1657, 2015.

[168] Jupyter widgets community. ipywidgets, a github repository. Retrieved from https://github.com/jupyter-widgets/ipywidgets, 2015.

[169] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[170] Peter Uetz, Loic Giot, Gerard Cagney, Traci A Mansfield, Richard S Judson, James R Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, et al. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[171] Soon-Hyung Yook, Zoltán N Oltvai, and Albert-László Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

[172] Jihua Ran, Hui Li, Jianfeng Fu, Ling Liu, Yanchao Xing, Xiumei Li, Hongming Shen, Yan Chen, Xiaofang Jiang, Yan Li, et al. Construction and analysis of the protein-protein interaction network related to essential hypertension. *BMC systems biology*, 7:1–12, 2013.

[173] Chen Chen, Hong Shen, Li-Guo Zhang, Jian Liu, Xiao-Ge Cao, An-Liang Yao, Shao-San Kang, Wei-Xing Gao, Hui Han, Feng-Hong Cao, et al. Construction and analysis of protein-protein interaction networks based on proteomics data of prostate cancer. *International journal of molecular medicine*, 37(6):1576–1586, 2016.

[174] Ryan P McGlinchey, Frank Shewmaker, Peter McPhie, Begoña Monterroso, Kent Thurber, and Reed B Wickner. The repeat domain of the melanosome fibril protein pmel17 forms the amyloid core promoting melanin synthesis. *Proceedings of the National Academy of Sciences*, 106(33):13731–13736, 2009.

[175] Reeba Susan Jacob, A Anoop, and Samir K Maji. Protein nanofibrils as storage forms of peptide drugs and hormones. *Biological and Bio-inspired Nanomaterials: Properties and Assembly Mechanisms*, pages 265–290, 2019.

[176] Xingmei Qi, Yu Wang, Hairui Yu, Ruifang Liu, Axel Leppert, Zihan Zheng, Xueying Zhong, Zhen Jin, Han Wang, Xiaoli Li, et al. Spider silk protein forms amyloid-like nanofibrils through a non-nucleation-dependent polymerization mechanism. *Small*, 19(46):2304031, 2023.

[177] Jakub W Wojciechowski, Emirhan Tekoglu, Marlena Gąsior-Głogowska, Virginie Coustou, Natalia Szulc, Monika Szefczyk, Marta Kopaczyńska, Sven J Saupe, and Witold Dyrka. Exploring a diverse world of effector domains and amyloid signaling motifs in fungal nlr proteins. *PLoS Computational Biology*, 18(12):e1010787, 2022.

[178] Ruben Hervas, Michael J Rau, Younshim Park, Wenjuan Zhang, Alexey G Murzin, James AJ Fitzpatrick, Sjors HW Scheres, and Kausik Si. Cryo-em structure of a neuronal functional amyloid implicated in memory persistence in drosophila. *Science*, 367(6483):1230–1234, 2020.

[179] Einav Tayeb-Fligelman, Orly Tabachnikov, Asher Moshe, Orit Goldshmidt-Tran, Michael R Sawaya, Nicolas Coquelle, Jacques-Philippe Colletier, and Meytal Landau. The cytotoxic staphylococcus aureus psmα3 reveals a cross-α amyloid-like fibril. *Science*, 355(6327):831–833, 2017.

[180] Shon A Levkovich, Ehud Gazit, and Dana Laor Bar-Yosef. Two decades of studying functional amyloids in microorganisms. *Trends in Microbiology*, 29(3):251–265, 2021.

[181] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

[182] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[183] Aric Hagberg and Drew Conway. Networkx: Network analysis with python. *URL: https://networkx. github. io*, pages 1–48, 2020.

[184] Andreas Heger and Liisa Holm. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Structure, Function, and Bioinformatics*, 41(2):224–237, 2000.

[185] Andreas Biegert and Johannes Söding. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, 24(6):807–814, 2008.

[186] Julien Jorda and Andrey V Kajava. T-reks: identification of tandem repeats in sequences with a k-means based algorithm. *Bioinformatics*, 25(20):2632–2638, 2009.

[187] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[188] Daniel WA Buchan and David T Jones. The psipred protein analysis workbench: 20 years on. *Nucleic acids research*, 47(W1):W402–W407, 2019.

[189] Alexander E Yarawsky, Stefanie L Johns, Peter Schuck, and Andrew B Herr. The biofilm adhesion protein aap from staphylococcus epidermidis forms zinc-dependent amyloid fibers. *Journal of Biological Chemistry*, 295(14):4411–4427, 2020.

[190] Joan A Geoghegan, Rebecca M Corrigan, Dominika T Gruszka, Pietro Speziale, James P O'Gara, Jennifer R Potts, and Timothy J Foster. Role of surface protein sasg in biofilm formation by staphylococcus aureus. *Journal of bacteriology*, 192(21):5663–5673, 2010.

[191] Mihaly Varadi, Greet De Baets, Wim F Vranken, Peter Tompa, and Rita Pancsa. Amypro: a database of proteins with validated amyloidogenic regions. *Nucleic acids research*, 46(D1):D387–D392, 2018.

[192] Elena Yarmola, Ivan P Ishkov, Nicholas M di Cologna, Megan Menashe, Robert L Whitener, Joanna R Long, Jacqueline Abranches, Stephen J Hagen, and L Jeannine Brady. Amyloid aggregates are localized to the nonadherent detached fraction of aging streptococcus mutans biofilms. *Microbiology Spectrum*, 10(4):e01661–22, 2022.

[193] Qing Meng, Qiuqiang Gao, Shebli Mehrazarin, Kamonchanok Tangwanichgapong, Yu Wang, Yiming Huang, Yutong Pan, Samuel Robinson, Ziwen Liu, Amirali Zangiabadi, et al. Fusobacterium nucleatum secretes amyloid-like fada to enhance pathogenicity. *EMBO reports*, 22(7):e52891, 2021.

[194] Diego Romero, Claudio Aguilar, Richard Losick, and Roberto Kolter. Amyloid fibers provide structural integrity to bacillus subtilis biofilms. *Proceedings of the National Academy of Sciences*, 107(5):2230–2234, 2010.

[195] Richard N Besingi, Iwona B Wenderska, Dilani B Senadheera, Dennis G Cvitkovitch, Joanna R Long, Zezhang T Wen, and L Jeannine Brady. Functional amyloids in streptococcus mutans, their use as targets of biofilm inhibition and initial characterization of smu_63c. *Microbiology*, 163(4):488–501, 2017.

[196] Christopher J Alteri, Juan Xicohténcatl-Cortes, Sonja Hess, Guillermo Caballero-Olín, Jorge A Giron, and Richard L Friedman. Mycobacterium tuberculosis produces pili during human infection. *Proceedings of the National Academy of Sciences*, 104(12):5145–5150, 2007.

[197] Elizabeth B Sawyer, Dennis Claessen, Maria Haas, Bhavna Hurgobin, and Sally L Gras. The assembly of individual chaplin peptides from streptomyces coelicolor into functional amyloid fibrils. *PLoS One*, 6(4):e18839, 2011.

[198] David Partouche, Valeria Militello, Andrea Gomez-Zavaglia, Frank Wien, Christophe Sandt, and Véronique Arluison. In situ characterization of hfq bacterial amyloid: a fourier-transform infrared spectroscopy study. *Pathogens*, 8(1):36, 2019.

[199] Anirudha Dutta, Sudipta Bhattacharyya, Anirban Kundu, Debabrata Dutta, and Amit Kumar Das. Macroscopic amyloid fiber formation by staphylococcal biofilm associated suhb protein. *Biophysical chemistry*, 217:32–41, 2016.

[200] Anastasiia O Kosolapova, Mikhail V Belousov, Anna I Sulatskaya, Maria E Belousova, Maksim I Sulatsky, Kirill S Antonets, Kirill V Volkov, Anna N Lykholay, Oksana Y Shtark, Ekaterina N Vasileva, et al. Two novel amyloid proteins, ropa and ropb, from the root nodule bacterium rhizobium leguminosarum. *Biomolecules*, 9(11):694, 2019.

[201] Ruiguang Ge, Xuesong Sun, Dongxian Wang, Qinglu Zhou, and Hongzhe Sun. Histidine-rich protein hpn from helicobacter pylori forms amyloid-like fibrils in vitro and inhibits the proliferation of gastric epithelial ags cells. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1813(8):1422–1427, 2011.

[202] Yan Wang, Jingbo Jiang, Yachao Gao, Yang Sun, Jianfeng Dai, Yang Wu, Di Qu, Gang Ma, and Xianyang Fang. Staphylococcus epidermidis small basic protein (sbp) forms amyloid fibrils, consistent with its function as a scaffolding protein in biofilms. *Journal of Biological Chemistry*, 293(37):14296–14311, 2018.

[203] Jonghee Oh, Jung-Gun Kim, Eunkyung Jeon, Chang-Hyuk Yoo, Jae Sun Moon, Sangkee Rhee, and Ingyu Hwang. Amyloidogenesis of type iii-dependent harpins from plant pathogenic bacteria. *Journal of Biological Chemistry*, 282(18):13601–13609, 2007.

[204] Masihuz Zaman and Maria Andreasen. Cross-talk between individual phenol-soluble modulins in staphylococcus aureus biofilm enables rapid and efficient amyloid formation. *Elife*, 9:e59776, 2020.

[205] Xuan Wang, Neal D Hammer, and Matthew R Chapman. The molecular basis of functional bacterial amyloid polymerization and nucleation. *Journal of Biological Chemistry*, 283(31):21530–21539, 2008.

[206] Neal D Hammer, Jens C Schmidt, and Matthew R Chapman. The curli nucleator protein, csgb, contains an amyloidogenic domain that directs csga polymerization. *Proceedings of the National Academy of Sciences*, 104(30):12494–12499, 2007.

[207] Christine R Langlois, Fen Pei, Suzanne S Sindi, and Tricia R Serio. Distinct prion domain sequences ensure efficient amyloid propagation by promoting chaperone binding or processing in vivo. *PLoS genetics*, 12(11):e1006417, 2016.

[208] Roland Riek and Sven J Saupe. The het-s/s prion motif in the control of programmed cell death. *Cold Spring Harbor perspectives in biology*, 8(9):a023515, 2016.

[209] Paraskevi L Tsiolaki, Nikolaos N Louros, and Vassiliki A Iconomidou. Hexapeptide tandem repeats dictate the formation of silkmoth chorion, a natural protective amyloid. *Journal of molecular biology*, 430(20):3774–3783, 2018.

[210] Casper B Rasmussen, Gunna Christiansen, Brian S Vad, Carina Lynggaard, Jan J Enghild, Maria Andreasen, and Daniel Otzen. Imperfect repeats in the functional amyloid protein fapc reduce the tendency to fragment during fibrillation. *Protein Science*, 28(3):633–642, 2019.

[211] Shuiliang Yu, Shaoman Yin, Chaoyang Li, Poki Wong, Binggong Chang, Fan Xiao, Shin-Chung Kang, Huimin Yan, Gengfu Xiao, Po Tien, et al. Aggregation of prion protein with insertion mutations is proportional to the number of inserts. *Biochemical Journal*, 403(2):343–351, 2007.

[212] J-M Lee, EM Ramos, J-H Lee, T Gillis, JS Mysore, MR Hayden, SC Warby, P Morrison, M Nance, CA Ross, et al. Cag repeat expansion in huntington disease determines age at onset in a fully dominant fashion. *Neurology*, 78(10):690–695, 2012.

[213] Bruce L Goode, Miu Chau, Paul E Denis, and Stuart C Feinstein. Structural and functional differences between 3-repeat and 4-repeat tau isoforms: implications for normal tau function and the onset of neurodegenerative disease. *Journal of Biological Chemistry*, 275(49):38182–38189, 2000.

[214] Jeffrey C Kessler, Jean-Christophe Rochet, and Peter T Lansbury. The n-terminal repeat domain of α-synuclein inhibits β-sheet and amyloid fibril formation. *Biochemistry*, 42(3):672–678, 2003.

[215] Chiharu Mizuguchi, Miho Nakagawa, Norihiro Namba, Misae Sakai, Naoko Kurimitsu, Ayane Suzuki, Kaho Fujita, Sayaka Horiuchi, Teruhiko Baba, Takashi Ohgita, et al. Mechanisms of aggregation and fibril formation of the amyloidogenic n-terminal fragment of apolipoprotein ai. *Journal of Biological Chemistry*, 294(36):13515–13524, 2019.

[216] Xiaodi Deng, Jamie Morris, James Dressmen, Matthew R Tubb, Patrick Tso, W Gray Jerome, W Sean Davidson, and Thomas B Thompson. The structure of dimeric apolipoprotein a-iv and its mechanism of self-association. *Structure*, 20(5):767–779, 2012.

[217] Robert Bussell Jr and David Eliezer. A structural and functional role for 11-mer repeats in α-synuclein and other exchangeable lipid binding proteins. *Journal of molecular biology*, 329(4):763–778, 2003.

[218] Xuan Wang and Matthew R Chapman. Sequence determinants of bacterial amyloid formation. *Journal of molecular biology*, 380(3):570–580, 2008.

[219] Caroline F Wright, Sarah A Teichmann, Jane Clarke, and Christopher M Dobson. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*, 438(7069):878–881, 2005.

[220] Karol Wróblewski, Mateusz Zalewski, Aleksander Kuriata, and Sebastian Kmiecik. Cabs-flex 3.0: an online tool for simulating protein structural flexibility and peptide modeling. *Nucleic Acids Research*, page gkaf412, 2025.

[221] Martin von Bergen, Stefan Barghorn, Shirley A Müller, Marcus Pickhardt, Jacek Biernat, Eva-Maria Mandelkow, Peter Davies, Ueli Aebi, and Eckhard Mandelkow. The core of tau-paired helical filaments studied by scanning transmission electron microscopy and limited proteolysis. *Biochemistry*, 45(20):6446–6457, 2006.

[222] Nikolaos N Louros and Vassiliki A Iconomidou. Identification of an amyloid fibril forming segment of human pmel17 repeat domain (rpt domain). *Peptide Science*, 106(1):133–139, 2016.

[223] Curran Oi, John D Treado, Zachary A Levine, Christopher S Lim, Kirsten M Knecht, Yong Xiong, Corey S O'Hern, and Lynne Regan. A threonine zipper that mediates protein–protein interactions: structure and prediction. *Protein Science*, 27(11):1969–1977, 2018.

[224] Benita Wiatrak, Janusz Piasny, Amadeusz Kuźniarski, and Kazimierz Gąsiorowski. Interactions of amyloid-β with membrane proteins. *International journal of molecular sciences*, 22(11):6075, 2021.

[225] Mikkel Christensen and Birgit Schiøtt. Revealing a dual role of ganglioside lipids in the aggregation of membrane-associated islet amyloid polypeptide. *The Journal of Membrane Biology*, 252:343–356, 2019.

[226] Michele FM Sciacca, Carmelo La Rosa, and Danilo Milardi. Amyloid-mediated mechanisms of membrane disruption. *Biophysica*, 1(2):137–156, 2021.

[227] Jingyao Li and Fuzhong Zhang. Amyloids as building blocks for macroscopic functional materials: designs, applications and challenges. *International journal of molecular sciences*, 22(19):10698, 2021.

[228] Dmitry Kurouski and Igor K Lednev. The impact of protein disulfide bonds on the amyloid fibril morphology. *International journal of biomedical nanoscience and nanotechnology*, 2(2):167–176, 2011.

[229] Jérôme Hennetin, Bérangère Jullian, Alasdair C Steven, and Andrey V Kajava. Standard conformations of β-arches in β-solenoid proteins. *Journal of molecular biology*, 358(4):1094–1105, 2006.

[230] Stanislav A Bondarev, Galina A Zhouravleva, Mikhail V Belousov, and Andrey V Kajava. Structure-based view on [psi+] prion properties. *Prion*, 9(3):190–199, 2015.

[231] Mike Sleutel, Brajabandhu Pradhan, Alexander N Volkov, and Han Remaut. Structural analysis and architectural principles of the bacterial amyloid curli. *Nature Communications*, 14(1):2822, 2023.

[232] Xuan Wang, Yizhou Zhou, Juan-Jie Ren, Neal D Hammer, and Matthew R Chapman. Gatekeeper residues in the major curlin subunit modulate bacterial amyloid fiber biogenesis. *Proceedings of the National Academy of Sciences*, 107(1):163–168, 2010.

[233] Fatemeh Nouri Emamzadeh. Role of apolipoproteins and α-synuclein in parkinson's disease. *Journal of Molecular Neuroscience*, 62(3):344–355, 2017.

[234] Thierry Pillot, Laurence Lins, Marc Goethals, Berlinda Vanloo, Johan Baert, Joel Vandekerckhove, Maryvonne Rosseneu, and Robert Brasseur. The 118–135 peptide of the human prion protein forms amyloid fibrils and induces liposome fusion. *Journal of molecular biology*, 274(3):381–393, 1997.

[235] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[236] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[237] Mara Zielinski, Christine Röder, and Gunnar F Schröder. Challenges in sample preparation and structure determination of amyloids by cryo-em. *Journal of Biological Chemistry*, 297(2), 2021.

[238] Brandon H Toyama and Jonathan S Weissman. Amyloid structure: conformational diversity and consequences. *Annual review of biochemistry*, 80(1):557–585, 2011.

[239] William G Tharp and Indra Neil Sarkar. Origins of amyloid-β. *BMC genomics*, 14:1–15, 2013.

[240] Sujeet S Bhoite, Yilin Han, Brandon T Ruotolo, and Matthew R Chapman. Mechanistic insights into accelerated α-synuclein aggregation mediated by human microbiome-associated functional amyloids. *Journal of Biological Chemistry*, 298(7), 2022.

[241] Peleg Ragonis-Bachar, Gabriel Axel, Shahar Blau, Nir Ben-Tal, Rachel Kolodny, and Meytal Landau. What can alphafold do for antimicrobial amyloids? *Proteins: Structure, Function, and Bioinformatics*, 92(2):265–281, 2024.

[242] Pawel P Wozniak and Malgorzata Kotulska. Amyload: website dedicated to amyloidogenic protein fragments. *Bioinformatics*, 31(20):3395–3397, 2015.

[243] Nikolaos Louros, Rob Van Der Kant, Joost Schymkowitz, and Frederic Rousseau. Stamp-db: a platform for structures of polymorphic amyloid fibril cores. *Bioinformatics*, 38(9):2636–2638, 2022.

[244] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[245] Milot Mirdita, Martin Steinegger, and Johannes Söding. Mmseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, 35(16):2856–2858, 2019.

[246] Ernst W Schmid and Johannes C Walter. Predictomes, a classifier-curated database of alphafold-modeled protein-protein interactions. *Molecular cell*, 85(6):1216–1232, 2025.

[247] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1):82–92, 2002.

[248] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[249] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

[250] Vivian Monzon, Typhaine Paysan-Lafosse, Valerie Wood, and Alex Bateman. Reciprocal best structure hits: using alphafold models to discover distant homologues. *Bioinformatics Advances*, 2(1):vbac072, 2022.

[251] Joseph W Schafer and Lauren L Porter. Alphafold2's training set powers its predictions of some fold-switched conformations. *Protein Science*, 34(4):e70105, 2025.

[252] Olivia S Pratt, Luc G Elliott, Margaux Haon, Shahram Mesdaghi, Rebecca M Price, Adam J Simpkin, and Daniel J Rigden. Alphafold 2, but not alphafold 3, predicts confident but unrealistic β-solenoid structures for repeat proteins. *Computational and structural biotechnology journal*, 2025.

[253] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

[254] Devlina Chakravarty, Joseph W Schafer, Ethan A Chen, Joseph F Thole, Leslie A Ronish, Myeongsang Lee, and Lauren L Porter. Alphafold predictions of fold-switched conformations are driven by structure memorization. *Nature communications*, 15(1):7296, 2024.

[255] Nikhil Aggarwal, Shohei Kitano, Ginette Ru Ying Puah, Sandra Kittelmann, In Young Hwang, and Matthew Wook Chang. Microbiome and human health: current understanding, engineering, and enabling technologies. *Chemical reviews*, 123(1):31–72, 2022.

[256] Leyuan Li, Tong Wang, Zhibin Ning, Xu Zhang, James Butcher, Joeselle M Serrana, Caitlin MA Simopoulos, Janice Mayne, Alain Stintzi, David R Mack, et al. Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics. *Nature Communications*, 14(1):3428, 2023.

[257] Paul J McMurdie and Susan Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4):e61217, 2013.

[258] Castrense Savojardo, Pier Luigi Martelli, Piero Fariselli, Giuseppe Profiti, and Rita Casadio. Busca: an integrative web server to predict subcellular localization of proteins. *Nucleic acids research*, 46(W1):W459–W466, 2018.

[259] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[260] Sebastian Canzler, Markus Fischer, David Ulbricht, Nikola Ristic, Peter W Hildebrand, and René Staritzbichler. Proteinprompt: a webserver for predicting protein–protein interactions. *Bioinformatics advances*, 2(1):vbac059, 2022.

[261] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, LI Zhan, et al. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The innovation*, 2(3), 2021.

[262] Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S Pollard, Ekaterina Sakharova, Donovan H Parks, Philip Hugenholtz, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology*, 39(1):105–114, 2021.

[263] Alicja W Wojciechowska, Jakub W Wojciechowski, Kinga Zielinska, Johannes Soeding, Tomasz Kosciolek, and Malgorzata Kotulska. Aggregating gut: on the link between neurodegeneration and bacterial functional amyloids. *bioRxiv*, pages 2024–11, 2024.

[264] Andreas Digre and Cecilia Lindskog. The human protein atlas—spatial localization of the human proteome in health and disease. *Protein science*, 30(1):218–233, 2021.

[265] Na-Ri Shin, Tae Woong Whon, and Jin-Woo Bae. Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in biotechnology*, 33(9):496–503, 2015.

[266] Ming-hua Zhu, Susan John, Maria Berg, and Warren J Leonard. Functional association of nmi with stat5 and stat1 in il-2-and ifn γ-mediated signaling. *Cell*, 96(1):121–130, 1999.

[267] Teng He, Yinbiao Qiao, Qi Yang, Jie Chen, Yongyuan Chen, Xiaoke Chen, Zhixing Hao, Mingjie Lin, Zheyu Shao, Pin Wu, et al. Nmi: a potential biomarker for tumor prognosis and immunotherapy. *Frontiers in pharmacology*, 13:1047463, 2022.

[268] K Sabrina Lynn, Raven J Peterson, and Michael Koval. Ruffles and spikes: Control of tight junction morphology and permeability by claudins. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1862(9):183339, 2020.

[269] Jonathan Landy, Emma Ronde, Nick English, Sue K Clark, Ailsa L Hart, Stella C Knight, Paul J Ciclitira, and Hafid Omar Al-Hassi. Tight junctions in inflammatory bowel diseases and inflammatory bowel disease associated colorectal cancer. *World journal of gastroenterology*, 22(11):3117, 2016.

[270] Sachiko Tsukita, Hiroo Tanaka, and Atsushi Tamura. The claudins: from tight junctions to biological systems. *Trends in biochemical sciences*, 44(2):141–152, 2019.

[271] Sergei Perov, Ofir Lidor, Nir Salinas, Nimrod Golan, Einav Tayeb-Fligelman, Maya Deshmukh, Dieter Willbold, and Meytal Landau. Structural insights into curli csga cross-β fibril architecture inspire repurposing of anti-amyloid compounds as anti-biofilm agents. *PLoS pathogens*, 15(8):e1007978, 2019.

[272] James M Hill and Walter J Lukiw. Microbial-generated amyloids and alzheimer's disease (ad), 2015.

[273] Yizhou Zhou, Luz P Blanco, Daniel R Smith, and Matthew R Chapman. Bacterial amyloids. *Amyloid Proteins: Methods and Protocols*, pages 303–320, 2012.

[274] Ali Keshavarzian, Stefan J Green, Phillip A Engen, Robin M Voigt, Ankur Naqib, Christopher B Forsyth, Ece Mutlu, and Kathleen M Shannon. Colonic bacterial composition in parkinson's disease. *Movement Disorders*, 30(10):1351–1360, 2015.

[275] Fen Zhang, Dominik Aschenbrenner, Ji Youn Yoo, and Tao Zuo. The gut mycobiome in health, disease, and clinical applications in association with the gut bacterial microbiome assembly. *The Lancet Microbe*, 3(12):e969–e983, 2022.

[276] Chun-Che Hung, Chiung-Chih Chang, Chi-Wei Huang, Rui Nouchi, and Chia-Hsiung Cheng. Gut microbiota in patients with alzheimer's disease spectrum: a systematic review and meta-analysis. *Aging (Albany NY)*, 14(1):477, 2022.

[277] Linus Jönsson. The personal economic burden of dementia in europe. *The Lancet Regional Health–Europe*, 20, 2022.

[278] Juan Mac Donagh, Abril Marchesini, Agostina Spiga, Maximiliano José Fallico, Paula Nazarena Arrías, Alexander Miguel Monzon, Aimilia-Christina Vagiona, Mariane Gonçalves-Kulik, Pablo Mier, and Miguel A Andrade-Navarro. Structured tandem repeats in protein interactions. *International Journal of Molecular Sciences*, 25(5):2994, 2024.