

UCLA

UCLA Electronic Theses and Dissertations

Title

Computing amyloid: approaches for evaluating, discovering, and targeting amyloids

Permalink

<https://escholarship.org/uc/item/6ch1d5pb>

Author

Zink, Samantha

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computing amyloid: approaches for evaluating, discovering, and targeting amyloids

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of
Philosophy in Biochemistry, Molecular and Structural Biology

by

Samantha Anne Zink

2024

© Copyright by
Samantha Anne Zink
2024

ABSTRACT OF THE DISSERTATION

Computing amyloid: approaches for evaluating, discovering, and targeting amyloids

by

Samantha Anne Zink

Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

University of California, Los Angeles, 2024

Professor Jose Alfonso Rodriguez, Chair

Amyloid proteins call into question our long-standing notions of the relationship between a protein sequence and its three-dimensional fold. These proteins exhibit distinct conformations associated with soluble and fibrillar states and potentially various fibril polymorphs. This conformational plasticity contributes to the diverse roles of amyloids, which fall into two distinct categories within biological systems. Pathogenic amyloids are observed in patients suffering from debilitating neurodegenerative and systemic disorders such as Alzheimer's Disease and Parkinson's Disease and can be infectious as is the case with prion-related diseases. In contrast to pathogenic amyloids, functional amyloid assemblies have a defined function in living systems. Such functional amyloids are prevalent across both eukaryotic and prokaryotic domains of life, implying the fold may be an early and conserved mechanism with value to cellular function. This dichotomy raises an important question: What dictates the fold adopted by amyloid-forming sequences and in turn, what determines the function of a particular amyloid fold? This thesis

attempts to address this puzzle with computational tools and experimental validation, as a stepping stone toward a broader understanding of amyloids across the tree of life.

Advances in structural biology have provided atomic-level views of amyloids, from short 'steric zipper' segments and to full amyloid fibril structures, offering researchers detailed insights into the features of amyloid proteins. Concurrently, the advent of machine learning techniques has allowed for the exploration of data that previously had been unfeasible. Importantly, despite their differences, pathogenic and functional amyloids have been found to share a common architecture: long unbranched fibrils with polypeptide chains folded nearly two-dimensionally within a layer, and with adjacent layers stacked parallel in register. The work presented in Chapter 2 of this dissertation explores the application of machine learning strategies to expand the capabilities of steric zipper prediction and investigates their utility in mining organism proteomes for high zipper propensity proteins that may harbor amyloid-forming domains. The approach offers a fast way to survey proteomes for putative amyloid-forming proteins or domains, providing insight into the prevalence and distribution of steric zippers across proteomes. Chapter 3 presents an alternative application of the methods developed in Chapter 2. Here, we employ computational approaches to design novel peptide sequences indented to bind a known fibril surface. Subsequently, we investigate the effects of these designed peptides on amyloid fibril growth and their resulting structures. By combining structural biology, machine learning, and targeted peptide design, the work presented in this thesis aims to deepen our understanding of amyloid proteins across diverse biological contexts, potentially paving the way for novel therapeutic strategies and insights into both pathogenic and functional amyloid formations.

The dissertation of Samantha Anne Zink is approved.

David S. Eisenberg

Joseph Ambrose Loo

Lin Jiang

Anastassia N. Alexandrova

Jose Alfonso Rodriguez, Committee Chair

University of California, Los Angeles

2024

DEDICATION

For Jon, whose belief in me became my strength – this journey was possible because of you.

For Julien, the heart of everything that I do.

TABLE OF CONTENTS

Abstract of the Dissertation	ii-iii
Dedication	v
List of Figures	vii
List of Tables	viii
Acknowledgements	ix-x
Curriculum Vitae	xi-xii
 Chapter 1: Introduction	 1-11
References	12-16
 Chapter 2: Using Structure-Informed Machine Learning to Predict Putative Amyloid-Like Zippers Across Entire Proteomes	 17-58
Introduction	18-20
Methods	20-25
Results	25-31
Hexapeptide Analysis	25-28
Proteome Analysis	28-31
Amyloid Atlas	28-30
Evaluating zipper-forming propensity across <i>Saccharomyces cerevisiae</i> proteome	30-31
Discussion	32-35
References	36-39
 Chapter 3: Machine learning-aided design of type-1 homozippers with fibril-surface binding potential	 59-72
Introduction	59-60
Materials and Methods	61-62
Results and Discussion	62-66
High-throughput computational peptide design	62-63
A preliminary structure solution of peptide B1_R115	63
Kinetic effects of designed peptides on fibrillization	64
Morphological screening of fibrils grown in the presence of peptide	64-65
Peptide docking	65-66
Concluding Remarks	66
References	67
 Appendix Chapter 1	 73-101
 Appendix Chapter 2	 102-130

List of Figures

Figure 2-1:	Structure-informed machine learning overview and top yeast candidates	41
Figure 2-2:	Score distribution histograms for machine learning datasets	43
Figure 2-3:	Correlation between Zif and percentage of amino acid type in <i>S. cerevisiae</i> proteome	44
Figure 2-4:	AlphaFold3 structure predictions in <i>S. cerevisiae</i> and <i>A. thaliana</i>	45
Figure 2S-1:	Description and validation of the structure-informed machine learning model	46
Figure 2S-2:	Relationship between amino acid composition and zipper propensity score	47
Figure 2S-3:	Impact of individual amino acid counts on predicted amyloid propensity scores in hexapeptides	48
Figure 2S-4:	Impact of charged residues in a hexapeptide sequence on zipper propensity scores	49
Figure 2S-5:	Charged residues at the zipper interface and the solvent interface have similar effects on the hexapeptides overall zipper propensity score	50
Figure 2S-6:	Mean zipper propensity scores for hexapeptides based on amino acid type and position	51
Figure 2S-7:	Score distribution for amyloid atlas and Zif comparison for notable amyloid proteins	52
Figure 2S-8:	Distribution of ranked Zif and pAD for proteins with solved amyloid fibril structures	53
Figure 2S-9:	Distribution of ranked Zif and pAD for proteins in the <i>S. cerevisiae</i> proteome	54
Figure 2S-10:	Correlation between Zif and amino acid content for proteins in the <i>S. cerevisiae</i> proteome	55
Figure 2S-11:	Metric calculation methods	56
Figure 2S-12:	Correlation between Zif and physical metrics for proteins with known amyloid structures	57
Figure 3-1:	Computational design and identification of candidate surface binders	69
Figure 3-2:	Crystallization of GNAQVY peptide	70
Figure 3-2:	Fibrillization kinetics of alpha-synuclein in the presence of designed peptides	71
Figure 3-4:	Dose-dependent fibrillization kinetics of alpha-synuclein and designed peptides	72

List of Tables

Table 2S-1: Metrics for proteins in Amyloid Atlas	58
Table 3S-1: Preliminary structure solution for B1_R115 peptide	73

ACKNOWLEDGMENTS

I would like to express my profound gratitude to those whose support and guidance were instrumental in the successful completion of this dissertation.

First, I would like to thank my advisor, Prof. Jose Rodriguez for his unwavering support and invaluable insights. His perspective, intellectual rigor and genuine investment in my growth as a researcher has left an indelible mark on this work and my development as a scholar. Similarly, I want to thank my committee, Prof. David Eisenberg, Prof. Joseph Loo, Prof. Lin Jiang, and Prof. Anastassia Alexandrova for their time, advice, and helpful feedback at many group meetings. I would also like to thank Dr. Michael Sawaya and Dr. Duilio Cascio for sharing their vast reservoir of knowledge in structural biology and for their thoughtful discussions and feedback during group meetings.

I would like to thank all of my lab mates, both past and present members for contributing to thoughtful discussions regarding the work presented here. In particular, Becky Jenkins, for her time and expertise in the collaboration of the peptide binding experiments discussed in Chapter 3. Jeff Qu, a very talented undergraduate researcher, for developing the machine learning methods presented in Chapter 2 as well as the crystallization studies presented in Chapter 3. Dr. Maria Flores and Dr. Logan Richards for their intellectual support and enriching my graduate school experience in immeasurable ways. And Dr. Marcus Gallagher-Jones for thoughtful discussions on protein modeling.

Finally, I would like to thank my friends and family for their love and support. My parents, for always encouraging me to pursue whatever I wanted to in life and for working hard to grant me all the opportunities I could dream of. My group of mom friends for their support and camaraderie in my journey through motherhood. To my husband, Jon: without your encouragement, this work would not have been possible. Thank you for believing in me, even when the path was difficult and uncertain. And to Julien, my sweet boy: you enrich my life in ways words cannot explain. I love you so much, and I am so lucky that I get to be your mom. Through everything, being your mom will always be my greatest achievement.

Chapter 2 is a draft manuscript by Samantha Zink, Songrong Qu, Michael Sawaya, David Eisenberg, and Jose A. Rodriguez. This project was conceived and designed by SZ, SQ, and JAR. SQ and JAR designed and trained the machine learning model. SZ built data processing pipeline and SQL storage database. SZ performed hexapeptide and proteome-level analysis.

Chapter 3 is a draft manuscript by Samantha Zink, Becky Jenkins, Songrong Qu, and Jose A. Rodriguez. This project was conceived and designed by SZ and JAR. SZ designed and analyzed peptides. SQ performed crystallization and peptide structure determination. BJ performed fibrillization kinetics and negative-stain EM. Molecular docking was performed by SZ.

CURRICULUM VITAE

Samantha Anne Zink

Education

- 2015-2017 University of California, Los Angeles
Bachelor of Science in Biochemistry
- 2018-2020 University of California, Los Angeles
Master of Science in Biochemistry, Molecular and Structural Biology
- 2018-2024 University of California, Los Angeles
Doctor of Philosophy in Biochemistry, Molecular and Structural Biology,
Expected

Honors and Awards

- 2020-2021 Molecular Biology Institute Whitcome Fellow, UCLA
- 2020-2021 UCLA Cellular and Molecular Biology Training Grant, UCLA

Publications: Original Research

- (1) Ferrero, S., Flores, M. D., Short, C., Vazquez, C. A., Clark, L. E., Ziegenbein, J., **Zink, S.**, Fuentes, D., Payes, C., Batto, M. V., Collazo, M., García, C. C., Abraham, J., Cordo, S. M., Rodriguez, J. A., & Helguera, G. (2021). Antibody-Based Inhibition of Pathogenic New World Hemorrhagic Fever Mammarenaviruses by Steric Occlusion of the Human Transferrin Receptor 1 Apical Domain. *Journal of Virology*, 95(17), 1868–1888.

- (2) Hickerson, B.T., Daniels-Wells, T.R., Payes, C., Clark, L.E., Candelaria, P.V., Bailey, K.W., Sefing, E.J., **Zink, S.**, Ziegenbein, J., Abraham, J., Helguera, G., Penichet, M.L., Gowen, B.B. (2022). Host receptor-targeted therapeutic approach to counter pathogenic New World mammarenavirus infections. *Nature Communications*, 13(1), 558.
- (3) Flores, M. D., Sawaya, M. R., Boyer, D. R., **Zink, S.**, Tovmasyan, S., Saucedo, A., Zee, C.-T., Cardenas, J., Fioriti, L., & Rodriguez, J. A. (2022). Structure of a reversible amyloid fibril formed by the CPEB3 prion-like domain reveals a core sequence involved in translational regulation. *BioRxiv : The Preprint Server for Biology*
- (4) Richards, L. S., Flores, M. D., **Zink, S.**, Schibrowsky, N. A., Sawaya, M. R., & Rodriguez, J. A. (2023). Cryo-EM Structure of a Human LECT2 Amyloid Fibril Reveals a Network of Polar Ladders at its Core. *Structure*, 31(11) 1386-1393.
- (5) **Zink, S.**, Qu, S., Sawaya, M., Eisenberg, D., Rodriguez, J.A. Leveraging Structure-Informed Machine Learning for Fast Steric Zipper Propensity Prediction Across Whole Proteomes. *Manuscript submitted, September 2024.*

Chapter 1

INTRODUCTION

A Historical View of Amyloid Studies

The initial identification of amyloid proteins dates to the mid-19th century, marking the beginning of a complex scientific narrative. In 1854, physician scientist Rudolf Virchow first described unusual deposits in human brain tissues that exhibited a reaction to iodine staining that was reminiscent of starch.¹ This observation led Virchow to coin the term “amyloid” – meaning starch-like. However, this classification was short-lived. Merely four years later, scientists Friedreich and Kekule upended Virchow’s assumption by demonstrating the presence of nitrogen in these peculiar masses, revealing them to be proteinaceous.¹ This pivotal discovery was just the first step in deconstructing the structural and functional complexities of amyloids, setting the stage for a scientific journey that would eventually unveil a diverse world of protein aggregates.

As the field of amyloid biology expanded, it became evident that these proteins were ubiquitous across the biological spectrum – from those involved in devastating diseases to others performing crucial biological functions across the tree of life. In humans, the discovery of amyloid involvement in diseases like Alzheimer’s, Parkinson’s and type-II diabetes dominated much of the early research. In 1984, Glenner and Wong’s isolation and purification of amyloid- β (A β) was the first instance linking an amyloid disease to a specific protein.² This led to the formulation of the ‘amyloid hypothesis’ and sparked a surge in research focused on amyloid formation, aggregation, and toxicity. In the 2002, however, Chapman et al. demonstrated that the extracellular fibrils produced by the bacteria *E. coli* are amyloid.³ The discovery of amyloid-like structures in bacteria

expanded the research beyond a medical context. Subsequent investigations revealed amyloid presence in fungi with the identification of yeast prions, as well as in higher order organism such as Pmel17 in mammals or Orb2 in *Drosophila*.⁴⁻⁶

The trajectory of amyloid research has been profoundly influenced by technological advancements, particularly in the realms of structural and computational methods. Astbury's application of X-ray fiber diffraction in 1935 provided the first insights into the cross- β structure of amyloid fibrils.⁷ Subsequently, X-ray diffraction and microcrystal electron diffraction (MicroED) enabled identification of atomic structures of amyloid fibril building blocks from short peptide segments.^{8,9} More recently, advancements in cryogenic electron microscopy (cryo-EM) and helical reconstruction have allowed us to visualize full-length amyloid fibrils in unprecedented detail, providing a comprehensive view of their structures.^{10,11}

In parallel, advancements in computational biology have played an increasingly crucial role in amyloid research. The development of parallel computing in the 1990s and GPU acceleration in the 2000s dramatically increased simulation timescales and system sizes. These technological leaps have enabled the application of molecular dynamics simulations to large protein systems, including amyloids, providing insights into the mechanics of amyloid formation and stability that experimental efforts alone cannot achieve. Such simulations have significantly enhanced our understanding of the hydrophobic affect in amyloid formation and elucidated the impact of mutations on fibril stability.^{12,13} Advancements in computing have also revolutionized high-throughput protein design and molecular modeling, further enhancing amyloid research. Protein design software, particularly Rosetta, has become a powerful tool for engineering second-

generation amyloid inhibitors.¹⁴ Rosetta's capabilities extend beyond design, forming the foundation of the 3D-profiling method crucial for predicting hexapeptide sequences with a high propensity to adopt steric zipper motifs – the fundamental structural element of amyloid fibrils.¹⁵ Finally, the impact of machine learning advances on protein research is only beginning to be realized. In 2020, AlphaFold harnessed the power of deep learning to achieve structure prediction accuracies surpassing those of traditional physics-based force fields.¹⁶ This breakthrough has been remarkable for canonical proteins, revolutionizing the field of protein structure prediction. However, this technology has not yet been successfully adapted for amyloid proteins, highlighting a potential area for future development in computational approaches to amyloid research.

As we continue to elucidate the complexities of amyloid biology, interdisciplinary approaches combining structural biology, biophysics, and computational methods promise to yield further insight into amyloid proteins.

An introduction to the amyloid fold

The amyloid fold is unique and differs remarkably from canonical globular and membrane protein folds. Canonical proteins typically employ a mix of varying structural motifs such as α -helices, β -sheets, and loops. They utilize all three dimensions in their folded state and generally adopt a single 3D structure. Amyloids, on the other hand, are protein assemblies comprised largely of β -sheets. Individual protein monomers fold in 2 dimensions which then stack and form stabilizing interactions with identical residues above and below. The resultant structure is a long, unbranched fibril with β strands stacked 4.8 Å apart and β sheets running perpendicular to the fibril axis and

spaced $\sim 8\text{-}12$ Å. It is this combination of motifs gives rise to the characteristic cross- β diffraction pattern. Within β strands, stabilizing hydrogen bonds form between identical backbone amide groups along the length of the fibril. Between β sheets, side chains interdigitate to form a dry interface called a steric zipper.⁸ Within the steric zipper, most amino acid side chains contribute stabilizing features between β sheets. The stacking of hydrophobic residues provides favorable van der Waals contacts and/or π - π stacking while the stacking of polar residues can form hydrogen bond ladders.^{8,17,18} Alternatively, uncomplimented charged residues can be destabilizing. Analysis on solved amyloid fibrils, however, reveal that uncomplimented charges are often located on the surface and exposed to solvent molecules, mitigating some of the electrostatic repulsion.¹⁹ It is the combination of these factors across many fibril layers that contributes to the marked stability of the amyloid fold.

Sequence Composition

Despite the similarities in their ultra structure, the sequences of amyloid proteins exhibit remarkable diversity. In fact, it is thought that under the correct conditions, most proteins can form fibrils due to the presence of amyloid-driving hexapeptide segments.²⁰ This fact makes it difficult to assess a protein's ability to form amyloid or predicting the fold itself difficult. The importance of polar residues in the stabilization of amyloid cores was shown early on in the cases of both neurodegenerative disease and functional yeast prions.^{18,21} Early bioinformatic approaches have identified sequence patterns that promote the formation of β -sheet rich fibril structures. In a set of *de novo* designed proteins, alternating patterns of hydrophobic and polar residues in the proteins sequence form structures resembling those of amyloid proteins.²² Subsequent studies revealed that these alternating patterns occur less often in natural proteins and suggest these sequences have been

disfavored by evolution.²³ Additionally, work from the Linquist lab revealed the expansion of sequence repeats enhances the aggregation of amyloid-like fibrils in the yeast prion Sup35.²⁴ However, work by Goldschmidt et al. showed that sequence composition alone is not enough to promote fibrillization but the particular order of residues imparts major determinants in amyloid formation.²⁰

Amyloid Polymorphism

Protein folding is conceptualized as a search for the global free energy minimum conformation for a given primary sequence. This process can be visualized as a funnel, where x and y axes represent all possible conformations given a sequence and z represents the free energy of each state. Under normal conditions, intramolecular contacts drive folding towards a single functional conformation at the energetic minima resulting in a smooth energy landscape. However, certain conditions can destabilize this functional form, promoting an aggregation pathway driven by intermolecular contacts. This shift results in a more rugged energy landscape, where amyloid folds and other off-target aggregates predominate.

The rugged landscape of off-target aggregates reflects the propensity of amyloid proteins to form diverse structural polymorphs. These polymorphisms manifest not in the gross morphology of the amyloid fibril, but in the fine details of backbone and side chain geometries at their cores, impacting overall symmetry, and the number/type of contacts at the interface. Referencing the zipper alone, the side chains of zippers can pack differently (packing polymorphism), the zipper(s) can form from different segments of the protein (segmental polymorphism), or two segments with different sequences can come together to form a heterozymer.²⁵ With a single protein having

multiple potential zipper segments, the possibilities for potential folds are numerous. Because of this, the environment in which the protein exists has the potential to influence the polymorph(s) that is ultimately formed. Environmental factors, such as the presence of complementary molecules like ionic salts or cofactors, can mitigate electrostatic repulsion due to close contact with like charges. In α -synuclein (α Syn) isolated from multiple system atrophy patients, positively charged lysine residues are closely clustered.²⁶ Unresolved density in this region suggests an anionic ligand, potentially mitigating charge repulsion.²⁶ Additionally, post-translational modifications can exhibit similar charge dampening effects as is seen in the structures of pY39 α Syn.²⁷ Mutations in the protein, particularly those associated with familial forms of amyloid diseases, can alleviate electrostatic repulsions or steric issues, thereby lowering the energy barrier required to access specific amyloid conformations.

Amyloids exhibit a distinctive nucleation-driven growth mechanism

It is well-documented that the process of amyloid formation proceeds through a common nucleated growth mechanism.²⁸ When depicted graphically, it assumes a sigmoidal shape, encompassing an initial lag period followed by a phase of exponential increase ultimately flattening off at saturation levels. The lag phase is thought to be the span of time in which a nucleus, the smallest unit capable of templating fibril growth, can form. Once formed, monomers add to the nucleus and fibril elongation occurs quickly. The formation of a nucleus is typically a rare event, but several phenomena have been observed to abbreviate the initial lag period. One method is through the introduction of a preformed, stable nucleus, known as secondary nucleation or seeding. This process can occur in both a homogenous manner, where monomers add to fibrils of the same protein, or a heterogenous manner, where fibrils of one protein induce the fibrillization of a

different, but compatible, protein.^{29,30} Additionally, epitaxial nucleation is another possible mechanism to which fibril growth can be initiated. Here, the presence of a full-length fibril provides a surface to which a separate fibril grows, either in a homo- or heterogenous manner. It is important to note that not every type of seed/fibril is capable of cross-seeding or templating the growth of all amyloid fibrils. Understanding the compatibility requirements and the molecular determinants of these types of interactions is important, however, for understanding mechanisms in amyloid-related diseases and in the development of treatments.

Nature has taken advantage of the amyloid fold to perform a broad set of biological functions

Amyloids have been observed throughout the tree of life. In these cases, amyloid assemblies are employed by the organism to serve a diverse set of functions. For example, the amyloid Curli provides anchoring and structural support to bacterial biofilms.³ In mammals, the protein CPEB3 is thought to modulate the formation of long-term memories.³¹ Furthermore, functional amyloids can serve as protein storage repositories, exhibit regulatory roles as suppressors and activators of various processes, and act as vehicles for information transition in the form of prions.¹⁹ The breadth and scope of functional amyloid structures, as well as their associated roles, represent an emerging domain that is only beginning to be elucidated.

Amyloid fibrils are a pathological indicator of a wide variety of diseases

Amyloid proteins are found deposited into plaques as fibrillar aggregates in several human diseases where their presence is correlated with pathophysiology. These diseases are both neurological and systemic in nature, as aggregates containing specific proteins can be found across the body in organs such as the brain, kidney, liver, and heart. There are approximately 50 protein misfolding

disorders associated with the formation of amyloids, among the most notable are Alzheimer's Disease, Parkinson's Disease, and diabetes.³² These disorders are often diseases of aging, in part, due to the slow templating process of amyloid growth (see previous section). Factors such as familial inherited mutations can influence fibril growth patterns and are sometimes correlated with early disease onset.

In a group of neurodegenerative diseases, called synucleinopathies, there is an abnormal accumulation and deposition of the protein α -synuclein (α Syn) into fibrillar plaques. This group includes diseases such as Dementia with Lewy bodies, multiple system atrophy, and Parkinson's Disease (PD), the second leading cause of neurodegeneration after Alzheimer's Disease.³³ In PD, clinical studies have noted that sporadic PD often begins in a pre-symptomatic phase with the presence of α Syn in the dorsal motor nucleus of the vagus nerve, a region outside of the central nervous system.³⁴ In late stages of the disease, aggregates move into the central nervous system (CNS), specifically to the substantia nigra and cerebral cortex, in a manner suggestive of template-based spreading.^{34,35} The molecular mechanisms underlying the formation of α Syn deposits in the gut, and their subsequent migration to the brain, remain elusive; however, unraveling this phenomenon could provide valuable insights into the mechanisms driving amyloid fibril growth and seeding processes.

Introduction to in vitro and in silico methods in amyloid structural biology

Experimentally, the gold standard for characterizing high-resolution structure information of full-length amyloid fibrils is single-particle Cryo-EM. However, this process is slow, cumbersome, and restrictive as it cannot account for all polymorphs in a sample. Crystallography is an alternative to

Cryo-EM and has revealed high-resolution structural information of key zipper forming regions. This approach is powerful but is limited to small segments of the protein and thus cannot reveal full fibril folds. Atomic force microscopy (AFM) has proven invaluable in amyloid research, enabling both nanoscale visualization of fibril formation and assessment of material properties. Notably, AFM studies have revealed that amyloid fibrils possess strength comparable to steel, underscoring their remarkable structural integrity.³⁶ While experimental techniques have provided crucial insight into amyloid structures, the complexity and dynamic nature of amyloid formation benefits from complementary computational approaches. These methods as applied to amyloids are summarized below.

First principles calculations can offer structural predictions in lieu of experiments. Molecular Dynamics (MD) simulations have been employed to study the relative stability of wild-type and mutant amyloid fibrils, providing insight into how mutations can influence the strength of a fibril core.^{13,37} MD simulations have also been used in conjunction with homology-based modeling to predict the structure of bacterial-based amyloid fibrils.³⁸ Additionally, the application of MD simulations allows for the assessment of stability profiles of computationally designed mini-protein inhibitors aimed at preventing amyloid formation by proteins such as tau, amyloid- β ($\text{a}\beta$), and αSyn .¹⁴ Due to the prohibitive computational costs associated with MD simulations, alternative approaches are required for high-throughput evaluation and screening processes in full length fibrils.

Knowledge-based potentials, in contrast, offer powerful predictive power at a computationally affordable cost. These potentials are derived from previously determined structures and are based

on probabilities rather than approximations of the true quantum mechanical wavefunction. A front-runner in this space is the software suite, (Py)Rosetta.³⁹ Here, the stability of protein models is assessed through a combination of physical and statistically-derived terms.⁴⁰ With the energy function serving as the foundation, this approach has demonstrated successful application in various modeling domains, including protein design, molecular docking, and density-based modeling techniques.^{14,40–42}

Recent breakthroughs in machine learning have propelled the field of protein structure prediction and assessment into a new era. Amongst the top performers, DeepMind's AlphaFold and Rosetta (trRosetta, roseTTAfold) have consistently outperformed others in the latest Critical Assessment of protein Structure Prediction (CASP14) competition.^{16,43,44} No longer relying on sequence- or physical-based quantification metrics explicitly, these deep-learning based approaches effectively learn the structure of proteins in an end-to-end manner directly from the Protein Data Bank (PDB). The reliance of multiple sequence alignments makes predicting the folds of amyloid via this method unfeasible due to the low number of solved amyloid folds.

An introduction to amyloid specific protein-prediction methods

Many amyloid-specific computational approaches have been developed to aid in the assessment and prediction of amyloid proteins. These approaches can be broadly categorized into two main methods: sequence-based and structure-based. Sequence-based prediction techniques evaluate the intrinsic tendencies of amino acid sequences to form amyloid fibrils, without considering structural information. In this space, there has been significant focus on predicting regions prone to forming β -sheets, the building block of amyloid fibrils.^{45–47} These algorithms rely on the physiochemical

properties and arrangement of amino acids within the peptide. As an alternative approach, machine learning techniques have been leveraged to discover previously unidentified proteins in the yeast genome that display prion-like properties.⁴⁸ This approach is limited in its application due to the small and homogeneous training set, but is useful in identifying proteins of a specific amino acid make (see chapter 2). Alternatively, structure-based approaches rely on physical modeling and energetic evaluation to assess a sequences' likelihood to form fibrils, as is done in the 3D-profiling method.¹⁵ Here, the energetic fit of a hexapeptide sequence modeled onto the backbone of a known steric zipper is evaluated. This method has been successful at predicting sequences prone to forming type-1 homozippers both in individual proteins as well as across proteomes.²⁰

Overview

This thesis demonstrates the power of interdisciplinary approaches in advancing our understanding of the amyloid fold, with a particular focus on computational methods. In Chapter 2, I harness machine learning and big data techniques to mine proteomes for amyloid-forming segments, providing unprecedented insights into the prevalence and diversity of potential amyloidogenic sequence across various organisms. Chapter 3 extends this computational approach by applying design principles to engineer peptides predicted to interact with preformed fibrils, shedding light on the complex dynamics of fibril growth. Together, these studies highlight the significant potential of integrating computational and traditional biochemical methods to unravel the mysteries of amyloid formation and behavior, paving the way for future advances in both basic science and therapeutic interventions.

References

1. Sipe, J. D. & Cohen, A. S. Review: History of the Amyloid Fibril. *J. Struct. Biol.* **130**, 88–98 (2000).
2. Glenner, G. G. & Wong, C. W. Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochem. Biophys. Res. Commun.* **120**, 885–890 (1984).
3. Chapman, M. R. *et al.* Role of Escherichia coli Curli Operons in Directing Amyloid Fiber Formation. *Science* **295**, 851–855 (2002).
4. Shorter, J. & Lindquist, S. Prions as adaptive conduits of memory and inheritance. *Nat. Rev. Genet.* **6**, 435–450 (2005).
5. Fowler, D. M., Koulov, A. V., Balch, W. E. & Kelly, J. W. Functional amyloid – from bacteria to humans. *Trends Biochem. Sci.* **32**, 217–224 (2007).
6. Hervas, R. *et al.* Cryo-EM structure of a neuronal functional amyloid implicated in memory persistence in Drosophila. *Science* **367**, 1230–1234 (2020).
7. Astbury, W. T., Dickinson, S. & Bailey, K. The X-ray interpretation of denaturation and the structure of the seed globulins. *Biochem. J.* **29**, 2351-2360.1 (1935).
8. Nelson, R. *et al.* Structure of the cross- β spine of amyloid-like fibrils. *Nature* **435**, 773–778 (2005).
9. Rodriguez, J. A. *et al.* Structure of the toxic core of α -synuclein from invisible crystals. *Nature* **525**, 486–490 (2015).
10. Fitzpatrick, A. W. P. *et al.* Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature* **547**, 185–190 (2017).

11. He, S. & Scheres, S. H. W. Helical reconstruction in RELION. *J. Struct. Biol.* **198**, 163–176 (2017).
12. Krone, M. G. *et al.* Role of Water in Mediating the Assembly of Alzheimer Amyloid- β A β 16–22 Protofilaments. *J. Am. Chem. Soc.* **130**, 11066–11072 (2008).
13. Lemkul, J. A. & Bevan, D. R. Assessing the Stability of Alzheimer’s Amyloid Protofibrils Using Molecular Dynamics. *J. Phys. Chem. B* **114**, 1652–1660 (2010).
14. Murray, K. A. *et al.* De novo designed protein inhibitors of amyloid aggregation and seeding. *Proc. Natl. Acad. Sci.* **119**, e2206240119 (2022).
15. Thompson, M. J. *et al.* The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci.* **103**, 4074–4078 (2006).
16. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 1–11 (2021) doi:10.1038/s41586-021-03819-2.
17. Hughes, M. P. *et al.* Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science* **359**, 698–701 (2018).
18. Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 5355–5358 (1994).
19. Sawaya, M. R., Hughes, M. P., Rodriguez, J. A., Riek, R. & Eisenberg, D. S. The expanding amyloid family: Structure, stability, function, and pathogenesis. *Cell* **184**, 4857–4873 (2021).
20. Goldschmidt, L., Teng, P. K., Riek, R. & Eisenberg, D. Identifying the amyloids, proteins capable of forming amyloid-like fibrils. *Proc. Natl. Acad. Sci.* **107**, 3487–3492 (2010).

21. DePace, A. H., Santoso, A., Hillner, P. & Weissman, J. S. A Critical Role for Amino-Terminal Glutamine/Asparagine Repeats in the Formation and Propagation of a Yeast Prion. *Cell* **93**, 1241–1252 (1998).
22. West, M. W. *et al.* De novo amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11211–11216 (1999).
23. Broome, B. M. & Hecht, M. H. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis¹. *J. Mol. Biol.* **296**, 961–968 (2000).
24. Liu, J.-J. & Lindquist, S. Oligopeptide-repeat expansions modulate ‘protein-only’ inheritance in yeast. *Nature* **400**, 573–576 (1999).
25. Eisenberg, D. S. & Sawaya, M. R. Structural Studies of Amyloid Proteins at the Molecular Level. *Annu. Rev. Biochem.* **86**, 69–95 (2017).
26. Schweighauser, M. *et al.* Structures of α -synuclein filaments from multiple system atrophy. *Nature* **585**, 464–469 (2020).
27. Zhao, K. *et al.* Parkinson’s disease-related phosphorylation at Tyr39 rearranges α -synuclein amyloid fibril structure revealed by cryo-EM. *bioRxiv* 2020.04.13.040261 (2020) doi:10.1101/2020.04.13.040261.
28. Jarrett, J. T. & Lansbury, P. T. Seeding “one-dimensional crystallization” of amyloid: A pathogenic mechanism in Alzheimer’s disease and scrapie? *Cell* **73**, 1055–1058 (1993).
29. Andreetto, E. *et al.* Identification of Hot Regions of the A β –IAPP Interaction Interface as High-Affinity Binding Sites in both Cross- and Self-Association. *Angew. Chem. Int. Ed.* **49**, 3081–3085 (2010).
30. Sampson, T. R. *et al.* A gut bacterial amyloid promotes α -synuclein aggregation and motor impairment in mice. *eLife* **9**, e53111 (2020).

31. Si, K. & Kandel, E. R. The Role of Functional Prion-Like Proteins in the Persistence of Memory. *Cold Spring Harb. Perspect. Biol.* **8**, a021774 (2016).
32. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).
33. Elbaz, A., Carcaillon, L., Kab, S. & Moisan, F. Epidemiology of Parkinson's disease. *Rev. Neurol. (Paris)* **172**, 14–26 (2016).
34. Braak, H., Rüb, U., Gai, W. P. & Del Tredici, K. Idiopathic Parkinson's disease: possible routes by which vulnerable neuronal types may be subject to neuroinvasion by an unknown pathogen. *J. Neural Transm.* **110**, 517–536 (2003).
35. Jucker, M. & Walker, L. C. Pathogenic protein seeding in alzheimer disease and other neurodegenerative disorders. *Ann. Neurol.* **70**, 532–540 (2011).
36. Smith, J. F., Knowles, T. P. J., Dobson, C. M., MacPhee, C. E. & Welland, M. E. Characterization of the nanoscale properties of individual amyloid fibrils. *Proc. Natl. Acad. Sci.* **103**, 15806–15811 (2006).
37. Buchete, N.-V., Tycko, R. & Hummer, G. Molecular Dynamics Simulations of Alzheimer's β -Amyloid Protofilaments. *J. Mol. Biol.* **353**, 804–821 (2005).
38. Garnham, C. P., Campbell, R. L., Walker, V. K. & Davies, P. L. Novel dimeric β -helical model of an ice nucleation protein with bridged active sites. *BMC Struct. Biol.* **11**, 36 (2011).
39. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
40. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

41. Sahtoe, D. D. *et al.* Design of amyloidogenic peptide traps. *Nat. Chem. Biol.* 1–10 (2024)
doi:10.1038/s41589-024-01578-5.
42. Wang, R. Y.-R. *et al.* Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife* **5**, e17219 (2016).
43. Baek, M. & Baker, D. Deep learning and protein structure modeling. *Nat. Methods* **19**, 13–14 (2022).
44. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* **117**, 1496–1503 (2020).
45. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).
46. Trovato, A., Chiti, F., Maritan, A. & Seno, F. Insight into the Structure of Amyloid Fibrils from the Analysis of Globular Proteins. *PLOS Comput. Biol.* **2**, e170 (2006).
47. Tartaglia, G. G. *et al.* Prediction of Aggregation-Prone Regions in Structured Proteins. *J. Mol. Biol.* **380**, 425–436 (2008).
48. Alberti, S., Halfmann, R., King, O., Kapila, A. & Lindquist, S. A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins. *Cell* **137**, 146–158 (2009).

Title: Leveraging Structure-Informed Machine Learning for Fast Steric Zipper Propensity Prediction Across Whole Proteomes

Authors: Samantha Zink^{1#}, Songrong Qu^{1#}, Tom Holton², Eesha Shankar¹, Paulina Stanley¹, David Eisenberg², Michael R. Sawaya^{2*} and Jose A Rodriguez^{1*}

Affiliations: ¹Department of Chemistry and Biochemistry; UCLA-DOE Institute for Genomics and Proteomics; STROBE, NSF Science and Technology Center; University of California, Los Angeles (UCLA); Los Angeles, CA 90095, USA.

²Departments of Chemistry and Biochemistry and Biological Chemistry, UCLA, Los Angeles, CA 90095, USA; Howard Hughes Medical Institute, UCLA, Los Angeles, CA 90095, USA; UCLA-DOE Institute, UCLA, Los Angeles, CA 90095, USA; Molecular Biology Institute, UCLA, Los Angeles, CA 90095, USA

Denotes equal contribution

* Correspondence to: Jose A. Rodriguez (jrodriguez@mbi.ucla.edu), Michael R. Sawaya (sawaya@mbi.ucla.edu)

ABSTRACT

Predicting the amyloid fold and the propensity of peptide segments to adopt amyloid-like structures remain a challenge. However, recent progress has facilitated structure-based prediction of steric zipper propensity and the use of machine learning to accelerate the calculation of predictive models across many scientific areas. Leveraging these advances, we have developed a new approach for rapid proteome-wide assessment of zipper profiles that is informed by four million steric zipper predictions collected over ten years. This collection is used to build a machine learning model capable of predicting steric zipper propensity at rapid speed and allowing for the assessment of zippers at both the protein and proteome level. Our predictions show enrichment for zipper forming segments in proteins involved in cell wall reorganization in yeast, highlighting a potential category of interest for experimental characterization. Overall, our predictive model allows for the exploration of amyloid formation across the tree of life and provides a tool for rapid assessment of both novel and designed sequences for zipper density.

INTRODUCTION

The amyloid fold appears across the tree of life. While many amyloids are associated with protopathic diseases,¹ non-pathogenic amyloids have emerged as a counterpoint to the idea that amyloid assemblies result from misfolding errors. Non-pathogenic amyloids can be functional, or fulfill a specific role; for example, these can range from structural to regulatory.² Ongoing efforts have significantly broadened our understanding of the universe of amyloid proteins, but our perspective of amyloid-forming proteins or domains across entire proteomes is still limited.

Despite their unique function or association with disease, amyloid proteins are defined by a set of structural features that are common to all amyloid folds. Amyloid fibrils are typically

elongated and unbranched. Within them, individual protein monomers fold in roughly 2-dimensions, stacked $\sim 4.8\text{\AA}$ apart.³ These fibrils can be composed of thousands of monomer layers. Stacked β -sheets run parallel and in-register to the fibril axis while extended protein strands outspread perpendicularly from the axis. This scaffold is the source of the diagnostic cross- β diffraction pattern observed in pathogenic amyloids.^{4,5} In amyloid fibrils, stacked β -strands are stabilized through backbone hydrogen bonding which extend the length of the fibril. In amyloid fibril cores, the interfaces of mating β -sheets show tight interdigitation of peptide side chains, often so close in space to leave no room for water molecules. This structural motif was termed the steric-zipper, and is thought to be a major contribution to the stability of amyloid fibrils.⁶

It has been proposed that any protein can adopt an amyloid fold, making it difficult to predict amyloid propensity based on sequence alone. This idea is highlighted by the remarkable sequence diversity seen in the set of ~ 30 known amyloid proteins.³ However, some categories of proteins have a unique sequence composition that might predispose proteins to form functional amyloids; this is the case with prions. An enrichment of branched polar residues (Q/N) is a hallmark of prion-like proteins.⁷ This has allowed for the development of sequence-based machine learning approaches, such as that developed by Alberti et al., that resulted in the identification of 29 new yeast prion candidates that matched a particular sequence profile.⁸ It is currently not known, however, if the unique link between sequence composition and prion or amyloid folds extends beyond yeast prions or if this is diagnostic of functional prion behavior across organisms.

Structure-based approaches are an alternative method of identifying novel sequences that are predicted to adopt an amyloid-like fold. The online zipper database (zipperDB) is an example of structure-based amyloid predictors.⁹ Instead of only looking at the sequence, zipperDB models six residue segments of a protein sequence onto a type 1 steric zipper backbone and uses the

Rosetta energy to evaluate the likelihood of that segment's ability to adopt a steric zipper.⁹ This approach has been remarkably successful in identifying segments prone to forming zippers and has been used to query millions of hexapeptide sequences. However, its dependence on a multi-parameter search and optimization necessarily makes zipperDB computationally expensive and thus limits the scale of its deployment and the investigation of zipper forming potential on a proteome level.

Here, we aim to leverage the pre-existing scores generated by zipperDB over its 10+ years of existence by using machine learning to dramatically accelerate the assessment of potential amyloid-forming sequences. We train a fully-connected neural network on 4.1 million zipperDB sequence-score pairs and pre-compute zipper propensity scores for all possible 64 million hexapeptide sequences. We then investigate the utility of using type-1 zipper prediction scores to identify protein domains or regions with putative amyloid-forming potential across entire proteomes. This tool is publicly available as a webserver (<https://zipperdb.mbi.ucla.edu>) that uniquely enables the exploration of zipper forming potential across the tree of life.

METHODS

Zipper propensity prediction:

We treated the prediction problem as a regression to minimize the differences between the score predicted for a given hexapeptide and the score calculated for it by the original ZipperDB protocol. The set of available training data, obtained from queries submitted to the ZipperDB database, contained a total of 4,144,232 hexamer sequences and their corresponding scores. Segments were randomly partitioned into the training, test, and validation set with 3,276,800 segments included in the training set, 733,184 segments in the validation set, and 134,248 in the test set (including 310 test amyloid segments will be 134,558).

We constructed a network whose structure was composed of three fully connected layers. In it, hexapeptide sequences were one-hot-encoded to a 6×20 array, and then passed onto a 120×120 fully connected hidden layer (Figure S1a). Each node was connected to every node in the output nodes through linear operations, so each node produced an output equivalent to:

$$y_j = \sum w_i x + b_i$$

After the first layer of the network, a rectified linear unit (ReLU) was adopted as the activation function to introduce non-linearity to the model as follows:

$$f(x) = \max(0, x)$$

Outputs from the hidden layer were then passed through a 120×1 FC layer to produce a single number as a predicted score. Differences between predicted scores and those in the training set were minimized by reducing the mean squared loss (MSE) loss between the two. The loss function is calculated as follows:

$$\sum \frac{(P - T)^2}{N}$$

Where P is the predicted score, T is the original ZipperDB score, and N is the number of segments evaluated. Training converged after ~1,000 epochs, after which the loss score had decayed to ~0.2 REU² (Figure S1b). With this model, we computed scores for all possible permutations of hexapeptide sequences, totaling 20⁶ = 64,000,000 segments. Scores for this number of segments were produced within 45 minutes on an NVIDIA GeForce RTX 3090 GPU. This included the time required to load the model, generate predictions, and write all outputs into an output CSV file.

We also evaluated the accuracy of the model by evaluating the performance on the test set. Among all 134,248 segments, 90.6% of segments fell within the error range of ±0.5 REU while 98.0% of segments fell in the range of ±1 REU, indicating a high degree of accuracy achieved by

the model (Figure S1d). Additionally, we evaluated the network's steric zipper prediction accuracy using a confusion matrix on the test set (Figure S1c). The original protocol contended that any hexamer that has a REU score ≤ -23 are predicted to form steric zippers. In this case, positive was considered as $\text{REU} \leq -23.0$ while $\text{REU} > -23.0$ was considered negative. The confusion matrix was built as follows while the criteria is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. The model had an accuracy of 0.980, a sensitivity of 0.926, and a specificity of 0.992. It is worth noting that there was no clear distinction between some positive and negative segments. That is, for example, there was not a significant difference between a hexamer that had a score of -23.1 REU and another that has a score of -22.9 REU, though the latter was considered a negative while the former one a positive score.

Overall, we demonstrated that a simple FC network could replicate the time-consuming 3D profiling protocols of ZipperDB with a high accuracy and a significant increase in efficiency. Through the prediction of all possible permutations, we were then be able to carry out proteome-level analysis of steric zipper formation within minutes or seconds.

Calculated metrics

A number of metrics were calculated for each protein. Molecular weight was calculated as a sum of average isotopic masses of each amino acid in the sequence and the average isotopic mass

of one H₂O molecule (as in ExPASy ProtParam). GRand AVerage hydropathY (GRAVY) was calculated as a sum of Kyte-Doolittle hydrophobicity values across the sequence. Percent composition by amino acid and by type is the percent based off of the total sequence length. The amino acid types were split as follows: hydrophobic = (A, V, I, L, M), aromatic = (F, W, 0.5Y), positive = (H, K, R), negative = (D, E), polar = (N, Q, S, T, 0.5Y), special = (C, G, P). If an amino acid is in the sequence but not in the previous list, it was counted as non-canonical. Note tyrosine was assigned both polar and aromatic type and thus contributed a weight of 0.5 in each category.

Two additional metrics are calculated: the aliphatic index and instability index. The aliphatic index was defined as the relative volume of a protein occupied by aliphatic side chains (A, V, I, L) and was calculated as in Ikai, 1980.¹⁰ The instability index is an estimate of the relative stability of a protein in a test tube. A value < 40 was predicted to be stable and a value of > 40 predicted to be unstable.¹¹

Developed metrics

In addition to published metrics, we developed two protein level amyloid-specific metrics, Zif (Zipper fraction) and pAD (putative Amyloid-forming Domain). Zif represents the fraction of hexapeptides in the protein sequence that meet the threshold of qualifying as a zipper segment: any hexapeptide with a propensity score ≤ -23 .

pAD corresponds to any stretch of 75 residues within a sequence that contains the highest density of zipper segments in a protein; this was calculated using a sliding window approach. It reports the number of zipper segments present within the 75-residue window with the highest number of zipper segments. If a protein is smaller than 75 residues, the number of zipper segments in the whole protein was reported.

Transmembrane helix masking

Initial analysis of top scoring Zif/pAD sequences revealed proteins that contained long, repetitive stretches of hydrophobic residues. Gene Ontology (GO) analysis revealed a large majority of these sequences encoded proteins with many transmembrane (TM) domains. Because of this, we implemented a masking protocol that allowed us to view proteome-level zipper predictions without the overrepresentation of transmembrane domain-containing proteins, alongside the raw zipper predictions. To identify putative transmembrane domains we used the protocol proposed by Kyte and Doolittle, 1982, where protein segments of 19 residues long with GRAVY scores greater than 1.6 are assumed to have a high chance of being transmembrane domains.¹² In the TM-masked set, any identified zipper segments that overlapped with an assumed transmembrane domain were not considered when calculating proteome-wide Zif and pAD metrics. We recognize that the TM-masked set may suffer from false negatives, but consider that integral membrane segments are likely to be sequestered by the membrane and thus may be less likely to freely form amyloid fibrils.

Consideration of non-canonical amino acids and splice variants

Sequences containing non-canonical amino acids were omitted from our training and calculations due to the sparse sampling of these with predictive modeling programs. Thus, any proteins that contain non-canonical amino acids were not explicitly scored for overall amyloid propensity, any calculated metrics (instability index, GRAVY, etc.), or developed metrics (Zif, pAD). Proteins containing non-canonical amino acids make up a small fraction of the total number of proteins in the proteome (0% in yeast, 0.02% in Arabidopsis).

When analyzing whole proteomes, all splice variants are considered and plotted individually. For proteins in the amyloid atlas, the splice variant with the longest polypeptide sequence (i.e. unprocessed) was chosen, unless otherwise noted.

RESULTS

Hexapeptide Analysis

We trained a neural network on 4.1 million hexapeptide sequence-score pairs that were originally generated via the structure-based threading method of zipperDB.⁹ We then used this network to generate scores for all possible 64 million hexapeptides and performed a thorough assessment on the features of that dataset. We first explored the properties of the full set of 64 million possible hexapeptides, compared to the sets used in the training of the neural network. We separated the data into groups: all possible hexapeptides (allHex, $n = 64\text{M}$), the machine-learning set, which includes all segments submitted to the original zipperDB (train = 3.28M, test = 0.134M, and validation = 0.733M), and the set of hexapeptides excluded from training (excl = 59.86M). The hexapeptide scores are related to ΔG , where a negative score [reported in Rosetta Energy Units (REU)] indicates increased stability and a more favorable arrangement. Histograms of the scores are multimodal in the allHex and excl set with a majority of the density falling between -50 and 0 REU (Figure 2). Two additional subpopulations are apparent: one centered around 180 REU and one present between 0 and 75 REU. These populations represent hexapeptides that, when modeled in Rosetta, exhibit either unrealistically large scores indicative of physically impossible clashes or reveal potential biases in the scoring function for specific residues. The histograms from each machine learning set (train/test/val) appear unimodal with most of the data falling between -50 and 0 REU. For scores < 0 , the averages across all five datasets are overlapping, indicating

agreement across sets included and excluded from training. The differences in modality are likely due to bias in the original zipperDB toward protein sequences suspected to form amyloid; the sequences that constitute the training data were selected by users interested in amyloid, likely reducing sequence diversity. Therefore, the training data likely does not exhibit the diversity in all possible hexapeptide sequences, but the volume of data should be enough to mitigate the present biases.

We then looked at the behavior of the hexapeptide scores across various sequence compositions. There are numerous cases where an increase in the number of the specific amino acid type leads to a more favorable zipper score. As the number of hydrophobic residues in the hexapeptide increases, there is a higher likelihood of adopting a steric zipper motif, as reported by a linear decrease in the propensity score (Figure S2). We see the same linear dependency in the number of β -branched residues (I, T, V) present in the hexapeptide (Figure S2); as the number of β -branched residues increase, the propensity score becomes more favorable. Finally, the number of polar residues present is correlated to an increase in zipper propensity (Figure S2). Polar residues contain additional hydrogen bonding capabilities through their side chain atoms, which are thought to further stabilize the steric zipper motif. These trends are true across all five data subsets.

Other amino acid types are correlated to a decrease in zipper propensity. Aromatic residues are thought to stabilize the steric zipper through π - π stacking VDW interactions. In our analysis, however, we see a slight shift to less favorable as the number of aromatic residues in the hexapeptide is increased (< 1 REU) (Figure S2). Additionally, there is decrease in zipper propensity when adding charged residues (both negative and positive) as well as special residues (C, G, P) to the hexapeptide sequence (Figures S3 and S4). Charged residues tend to destabilize the zipper arrangement due to electrostatic repulsion and/or unmatched charged complementarity.

Sequences containing proline residues were omitted from the original zipperDB due to their unusually high propensity scores. Our method similarly yields elevated scores for proline-containing sequences, leading us to regard these scores as unreliable. These sequences drive the decrease in zipper propensity seen with the addition of special residues (Figure S3).

To understand this further, we examined how charge influences the propensity score of the hexapeptide, independent of the amino acid composition. As the total number of charged residues increases, the average propensity score becomes less favorable (Figure S4). This is also true when looking at the total charge of the hexapeptide. The resulting data takes on a parabolic shape with the minima at zero total charge and increasing as the total charge moves towards ± 6 (Figure S4). Interestingly, there is a linear increase in score when adding charged residues to both the zipper interface and the solvent interface (Figure S5). We would expect adding charged residues to the solvent interface to have less of a detrimental effect to the score than adding a charge to the zipper interface. This highlights potential biases in the solvation parameters of Rosetta.

Lastly, we assessed how zipper propensity changes when each amino acid occupies each position in the hexapeptide. We took the average score of all hexapeptides that had the amino acid of interest in the position of interest and expanded across all amino acids and positions in the hexapeptide. A majority of the amino acids had average scores of ~ -20 REU across all positions in the hexapeptide with the exception of arginine and proline (Figure S6). Scores of arginine-containing hexapeptides remain stable constant across positions, regardless of arginine position; however, the average score of ~ -15 REU is higher than other residue types. Arginine is a large, positively charged amino acid and is not expected to pack well in the zipper interface. Additionally, proline is inconsistent across positions. This is likely due to limitations in the training data as the original zipperDB did not routinely include scores for prolines. Additionally, the sequences with

prolines that were included in training were inconsistent and scored very high, indicating Rosetta has difficulty in modeling proline residues in the type-1 zipper backbone. In conclusion, arginine is a consistent driver in unfavorable scores and the inconsistency of proline score likely makes it unreliable.

Proteome Analysis

Amyloid Atlas

We next sought to develop metrics that describe the density of zipper segments in both the whole protein and specific protein domains. Zif is the fraction of protein that contains hexapeptides predicted to form zippers (see Methods). pAD, a domain specific metric, reports the total number of zipper segments per 75 residues in the sequence block with the most zipper segments. Hydrophobic residues pack well at the zipper interface and are also prominent in the membrane spanning regions of membrane proteins. Therefore, we implemented a masking procedure based on a segment's GRAVY score to reduce interference between favorable zipper segments and transmembrane domains (see Methods). These metrics highlight proteins with a high number of zipper-prone segments, and we hypothesize that these proteins are prone to forming amyloid fibrils.

To investigate the behavior of these metrics, we examined the properties of protein sequences with experimentally-determined amyloid structures (see Amyloid Atlas). This set of 36 unique protein sequences is populated mainly with human amyloids that have disease association but contains some examples across function and organism. The hexapeptide score distribution of these proven amyloid-forming proteins resembles that of the all-Hex set with no statistically

significant differences (Figure S7). The average Zif score is 0.15 ± 0.07 while the average pAD is 17.92 ± 8.70 (Figure S8). Among the top ranked candidates are prominent amyloid proteins such as tau and α -synuclein. There are only seven examples of functional amyloid proteins, but when ranked by Zif and pAD, their scores are evenly distributed throughout the dataset, suggesting that both functional and pathogenic amyloids contain a similar number of zipper segments (Table 1). While the two groups have a similar number of zipper segments, the way in which the adhesive interfaces come together to form the core is what dictates the behavior of that fibril species. The number of zipper segments is not indicative of how it will behave, just that it has the ability to form a fibril. When comparing Zif's in the fibril core with the proteoform sequence, we saw a higher density of zipper segments in the core (Figure S7b).

We next looked at correlations between the protein's properties and Zif score to gain a better understanding of features potentially correlated to zipper formation. We see moderately strong correlations with Zif scores and a protein's GRAVY and aliphatic index (both measures of hydrophobicity) (Figure S12). This makes sense given the strong positive correlations to hydrophobic and β -branched sequence composition and agrees with previously described literature on amyloid prone sequences.

We wanted to compare the performance of pAD on the set of observed amyloid structures. There are 13 observed polymorphs of the disease-associated, fully processed protein Islet Amyloid Polypeptide (IAPP). All observed cores overlap in sequence space (residues 34-70). pAD reports 17 of the 75 residues (0.23) are predicted to form zippers in the region 1-85 (multiple start positions with the same pAD score are reported) in unprocessed IAPP. If the window length is shortened to 40 residues, pAD reports a score of 14 out of 40 residues (0.35) are predicted to form zippers in

the sequence region of residue 27-76. Both of these sets overlap with the observed regions of the fibril core.

In another example, we looked at Nup98, the large nucleoporin protein that is involved in transport across the nuclear pore complex but is found phase-separated in certain cancers and deposited in amyloid fibrils in Alzheimer's disease where tau is also present.¹³ Despite the large size of this protein, pAD reports the top domain between residues 34-128 with 27 of the 75 residues (0.36) predicted to have high propensity for forming zippers. The observed fibril core is between residues 84-124 and overlaps with the predicted pAD.

This dataset gives us an idea as to what these metrics look like for a set of proteins with known amyloid folds. However, given the small number of proteins in this set as well as their bias towards human disease state, we wondered what this type of analysis would look like at the proteome level.

Evaluating zipper-forming propensity across the *Saccharomyces cerevisiae* proteome

We next wanted to understand the behavior of these metrics on a proteome level. As a starting point, we compared the behavior of the *Saccharomyces cerevisiae* proteome to the dataset of predicted yeast prion proteins in Alberti et al., 2009.⁸ The average Zif and pAD for the yeast proteome are 0.15 ± 0.05 and 19.5 ± 6.73 , respectively. About 65% of the Alberti dataset lies above the proteome average for both metrics (Figure S9).

Zif is negatively correlated to high percentages of charged residue content in the protein sequence (Figure 3). This is expected as charged residues are not predicted to be compatible with adopting a steric zipper motif. Interestingly, there is a positive correlation with the percentage of polar residues in the sequence, but the positive correlation comes with increased percentage of S

and T rather than Q and N (Figures 3 and S10). This is unexpected given the fact that Q/N are typically enriched in known amyloid-based prions, a feature that was prominent amongst the Alberti data set. Thus, this approach likely highlights yeast amyloid proteins that are dissimilar to the Q/N-rich prions found in Alberti et al.

Gene Ontology (GO) term analysis on the top 100 candidate proteins ranked by Zif are enriched for proteins involved in flocculation, cell wall/organelle organization and protein metabolic processes. Similarly, top ranked candidates for pAD are enriched for proteins involved in cell adhesion in biofilm formation, flocculation, and cell wall organization. This is in agreement with amyloid screening studies that noted enrichment in cell wall proteins.¹⁴

The top ranked protein according to pAD is the cell wall protein CRH1. This protein functions as a chitin transglycosylase and is upregulated upon an increase in temperature. pAD reveals a window between residue 299 and 374 with 67 predicted zipper segments (0.89) (Figure 1b). There are no experimental structures of this protein, but according to the AlphaFold3 (AF3) model, the N-terminal half of the protein is β -sheet rich with pLDDT scores > 90 indicating very high confidence in the prediction (Figure 1b).¹⁵ The C-terminal half, however, is unstructured and coincides with the putative amyloid domain (Figure 1b).

Wojciechowski et al. showed amyloid-like folds were predicted using AF3 under certain conditions.¹⁶ To gain insight into the potential amyloid fold of CRH1, we modeled 20 copies of the putative amyloid domain. AF3 indeed produces a model with amyloid-like features; mainly β -sheet rich, vertically stacked, with the correct distances present (Figure 4a). The 2-dimensional fold within the fibril core, however, is lost. It is in these intricacies in the heterozyper arrangements that dictates how the fibril folds in on itself.

DISCUSSION

The identification of amyloid-forming segments across entire proteomes is a computationally intensive problem. However, we have capitalized on the advent of readily accessible ML protocols and a decade of scored zipper predictions generated by the ZipperDB webserver to implement an approach to this problem that relies on the ML-driven scoring of zipper-forming propensities for hexapeptides. This implementation is orders of magnitude faster than existing approaches for scoring zipper formation propensity, though it yields a simple predicted score rather than a structural model of a predicted zipper complex.⁹

Given the speed of the updated ML-driven zipperDB, we considered it ideal to determine scores for all possible individual hexapeptide sequences. This analysis in turn yielded insights into potential biases in its scoring and enables proteome-wide characterization of zipper-forming propensities. As anticipated, an increasing number of hydrophobic residues led to a more stabilized zipper arrangement. These residues, especially Y, are known to stabilize amyloid fibrils due to favorable π - π stacking in the fibril core.^{6,17} In our analysis, however, the average score for hexapeptide sequences remained relatively unchanged regardless of the number of aromatic residues present. In contrast, increasing numbers of branched polar residues in the hexapeptide sequence also produced more favorable zipper scores.^{6,17,18} This was especially true for the β -branched amino acids (I, T, V), but also surprisingly true for serine. Conversely, we confirmed that an increased number of charged residues in the hexapeptide sequence results in a decrease in favorability. Unexpectedly, this decrease in favorability is present when adding charged residues to either the zipper interface or the solvent interface.³ These discrepancies likely reflect the opportunity for improvement in the solvation parameters in the Rosetta score function, the basis for our ML training. A further limitation is our restriction of peptide backbones to a type-1

homozipper arrangement. Theoretically, however, our analysis could be expanded to all ten symmetry patterns of steric homozippers, as well as to heterozippers, where the mating sheets contain different amino acid sequences.^{6,18,19} With our success in training of type-1 zippers, we believe that our structure-informed ML approach can be utilized to model and predict other steric zippers and the accompanied boost in computational speed could provide information on potential bias in the simulation protocol, especially for the implementation of the Rosetta score function.

Based on our analysis of zipper scores, we developed two predictive metrics that help assess proteome level zipper predictions. Zif describes the frequency of zipper segments on a whole protein, while pAD does so for a protein sub-domain. Our initial analysis performed on the *S. cerevisiae* proteome showed favorable agreement with previously identified putative prions or amyloids. The set of candidate yeast prions in the Alberti dataset are well distributed throughout the proteome according to both Zif and pAD, with about 70% residing above the proteome average for both metrics. However, while yeast prions are expected to have high Q and N content,^{7,8} high zipper propensity is not strictly correlated with Q/N content. This difference in propensity might reflect a residual bias in the energy function used for scoring in the original zipperDB.²⁰ This is not entirely surprising, since that scoring function did not explicitly account for hydrogen bond networks that might stabilize zipper cores.

GO term analysis of the top 100 ranked proteins in yeast revealed enrichment for proteins that localize to the cell surface, extracellular space, and cell septum/bud scar. In particular, GO term analysis revealed groups of proteins involved in biofilm formation, flocculation, and cell wall remodeling. This is in agreement with literature from other fungal species where functional amyloids have been shown to be involved in processes at the cell wall such as hydrophobins in filamentous fungi.^{21–23} Moreover, this is reminiscent of functional amyloids in bacterial species

that are present at the cell wall such as the reproductive proteins of the chaplin family in *Streptomyces coelicolor* and the biofilm proteins present in *Escherichia coli* and *Salmonella* spp.²⁴⁻

²⁸ The proteins Gas1 and Ygp1 are both involved in cell wall function and have been shown to have amyloid properties in vitro and were ranked 24th and 72nd by pAD. Originally identified in a yeast screen using PSIA-LC-MALDI, these proteins form SDS-resistant puncta in yeast cells and were shown to have amyloid properties via the bacteria-based system C-DAG¹⁴ (Figure 1b).

Intriguingly, among the top ranked protein domains is a family of serine-rich cell wall components. The TIR family (TIR1, TIR2, TIR3, TIR4) is upregulated upon an anaerobic shift in the environment where they move to the cell wall and replace existing cell wall proteins.²⁹ TIR2, TIR3, and TIR4 are all ranked within the top 10 with pAD scores of > 55 segments in the domain (0.73). The putative amyloid domain is predicted to start in the middle of the protein sequence, overlapping with a predicted unstructured region (AF3 pLDDT < 50).

In addition to our analysis of the yeast proteome, we pursued an additional model organism that is significantly understudied in amyloid biology, *Arabidopsis thaliana*. Proteome-wide zipper prediction in *Arabidopsis* revealed top ranked candidates are proteins that contain a high density of repeat segments and are predicted to have β -solenoid structures via AF3 (Figure 4b). This type of structure has been adopted by some prions and has amyloid characteristics. Most notably, the fungal prion Het-S uses a β -solenoid fold in its prion state to induce heterokaryon incompatibility.^{30,31} Additionally, the bacterial amyloid, Curli, uses a β -solenoid fold as a structural component of a biofilm.³² Most similar to the predicted fold, however, is the proposed amyloid-like β -solenoid fold of the ice nucleating proteins in bacteria.³³⁻³⁵ This is supported by work from the Linquist lab, which showed potential prion behavior in proteins involved in the autonomous flowering pathway of this species.³⁶ Luminidependens (LD), Flowering Locus PA

(FPA), and Flowering Locus CA (FCA) each possess prion-like properties. However, those proteins did not rank highly in our analysis likely due to the specific scoring of Q and N residues. Interestingly, it is suggested in the paper that these proteins could act as non-canonical prions due to lack of high molecular weight aggregates.

We have made our platform accessible to facilitate prediction and experimental validation of steric zipper forming segments and putative amyloidogenic proteins. We expect the speed and capacity of the new zipperDB to aid in the identification of proteins with a high proportion of zipper segments, which can narrow down the potential candidates for experimental testing. It remains to be seen whether a high density of zipper forming segments translates to high propensity to form amyloid fibrils *in vitro* and *in vivo*.

REFERENCES

1. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat Rev Mol Cell Biol* **15**, 384–396 (2014).
2. Fowler, D. M., Koulov, A. V., Balch, W. E. & Kelly, J. W. Functional amyloid – from bacteria to humans. *Trends in Biochemical Sciences* **32**, 217–224 (2007).
3. Sawaya, M. R., Hughes, M. P., Rodriguez, J. A., Riek, R. & Eisenberg, D. S. The expanding amyloid family: Structure, stability, function, and pathogenesis. *Cell* **184**, 4857–4873 (2021).
4. Astbury, W. T., Dickinson, S. & Bailey, K. The X-ray interpretation of denaturation and the structure of the seed globulins. *Biochem J* **29**, 2351-2360.1 (1935).
5. Sunde, M. *et al.* Common core structure of amyloid fibrils by synchrotron X-ray diffraction1. *Journal of Molecular Biology* **273**, 729–739 (1997).
6. Sawaya, M. R. *et al.* Atomic structures of amyloid cross- β spines reveal varied steric zippers. *Nature* **447**, 453–457 (2007).
7. Sondheimer, N. & Lindquist, S. Rnq1: An Epigenetic Modifier of Protein Function in Yeast. *Molecular Cell* **5**, 163–172 (2000).
8. Alberti, S., Halfmann, R., King, O., Kapila, A. & Lindquist, S. A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins. *Cell* **137**, 146–158 (2009).
9. Goldschmidt, L., Teng, P. K., Riek, R. & Eisenberg, D. Identifying the amyloids, proteins capable of forming amyloid-like fibrils. *Proc Natl Acad Sci USA* **107**, 3487–3492 (2010).
10. IKAI, A. Thermostability and Aliphatic Index of Globular Proteins. *The Journal of Biochemistry* **88**, 1895–1898 (1980).

11. Guruprasad, K., Reddy, B. V. & Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng* **4**, 155–161 (1990).
12. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132 (1982).
13. Ibáñez de Opakua, A. *et al.* Molecular interactions of FG nucleoporin repeats at high resolution. *Nat. Chem.* **14**, 1278–1285 (2022).
14. Ryzhova, T. A. *et al.* Screening for amyloid proteins in the yeast proteome. *Curr Genet* **64**, 469–478 (2018).
15. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 1–3 (2024) doi:10.1038/s41586-024-07487-w.
16. Wojciechowska, A. W., Wojciechowski, J. W. & Kotulska, M. Non-standard proteins in the lenses of AlphaFold3 - case study of amyloids. 2024.07.09.602655 Preprint at <https://doi.org/10.1101/2024.07.09.602655> (2024).
17. Nelson, R. *et al.* Structure of the cross- β spine of amyloid-like fibrils. *Nature* **435**, 773–778 (2005).
18. Riek, R. & Eisenberg, D. S. The activities of amyloids from a structural perspective. *Nature* **539**, 227–235 (2016).
19. Eisenberg, D. S. & Sawaya, M. R. Structural Studies of Amyloid Proteins at the Molecular Level. *Annu. Rev. Biochem.* **86**, 69–95 (2017).
20. Thompson, M. J. *et al.* The 3D profile method for identifying fibril-forming segments of proteins. *PNAS* **103**, 4074–4078 (2006).

21. Mackay, J. P. *et al.* The Hydrophobin EAS Is Largely Unstructured in Solution and Functions by Forming Amyloid-Like Structures. *Structure* **9**, 83–91 (2001).
22. de Vocht, M. L. *et al.* Structural and Functional Role of the Disulfide Bridges in the Hydrophobin SC3*. *Journal of Biological Chemistry* **275**, 28428–28432 (2000).
23. Macindoe, I. *et al.* Self-assembly of functional, amphipathic amyloid monolayers by the fungal hydrophobin EAS. *Proceedings of the National Academy of Sciences* **109**, E804–E811 (2012).
24. Claessen, D. *et al.* A novel class of secreted hydrophobic proteins is involved in aerial hyphae formation in *Streptomyces coelicolor* by forming amyloid-like fibrils. *Genes Dev.* **17**, 1714–1726 (2003).
25. Chapman, M. R. *et al.* Role of *Escherichia coli* Curli Operons in Directing Amyloid Fiber Formation. *Science* **295**, 851–855 (2002).
26. Collinson, S. K., Parker, J. M. R., Hodges, R. S. & Kay, W. W. Structural predictions of AgfA, the insoluble fimbrial subunit of *Salmonella* thin aggregative fimbriae1. *Journal of Molecular Biology* **290**, 741–756 (1999).
27. Larsen, P. *et al.* Amyloid adhesins are abundant in natural biofilms. *Environmental Microbiology* **9**, 3077–3090 (2007).
28. Jordal, P. B. *et al.* Widespread Abundance of Functional Bacterial Amyloid in Mycolata and Other Gram-Positive Bacteria. *Applied and Environmental Microbiology* **75**, 4101–4110 (2009).
29. Kowalski, L. R. Z., Kondo, K. & Inouye, M. Cold-shock induction of a family of TIP1-related proteins associated with the membrane in *Saccharomyces cerevisiae*. *Molecular Microbiology* **15**, 341–353 (1995).

30. Ritter, C. *et al.* Correlation of structural elements and infectivity of the HET-s prion. *Nature* **435**, 844–848 (2005).
31. Wasmer, C. *et al.* Amyloid Fibrils of the HET-s(218–289) Prion Form a β Solenoid with a Triangular Hydrophobic Core. *Science* **319**, 1523–1526 (2008).
32. Sleutel, M., Pradhan, B., Volkov, A. N. & Remaut, H. Structural analysis and architectural principles of the bacterial amyloid curli. *Nat Commun* **14**, 2822 (2023).
33. Graether, S. P. & Jia, Z. Modeling *Pseudomonas syringae* Ice-Nucleation Protein as a β -Helical Protein. *Biophysical Journal* **80**, 1169–1173 (2001).
34. Garnham, C. P., Campbell, R. L., Walker, V. K. & Davies, P. L. Novel dimeric β -helical model of an ice nucleation protein with bridged active sites. *BMC Structural Biology* **11**, 36 (2011).
35. Hansen, T. *et al.* Ice nucleation proteins self-assemble into large fibres to trigger freezing at near 0 °C. *eLife* **12**, RP91976 (2023).
36. Chakrabortee, S. *et al.* Luminidependens (LD) is an Arabidopsis protein with prion behavior. *Proceedings of the National Academy of Sciences* **113**, 6065–6070 (2016).

FIGURES

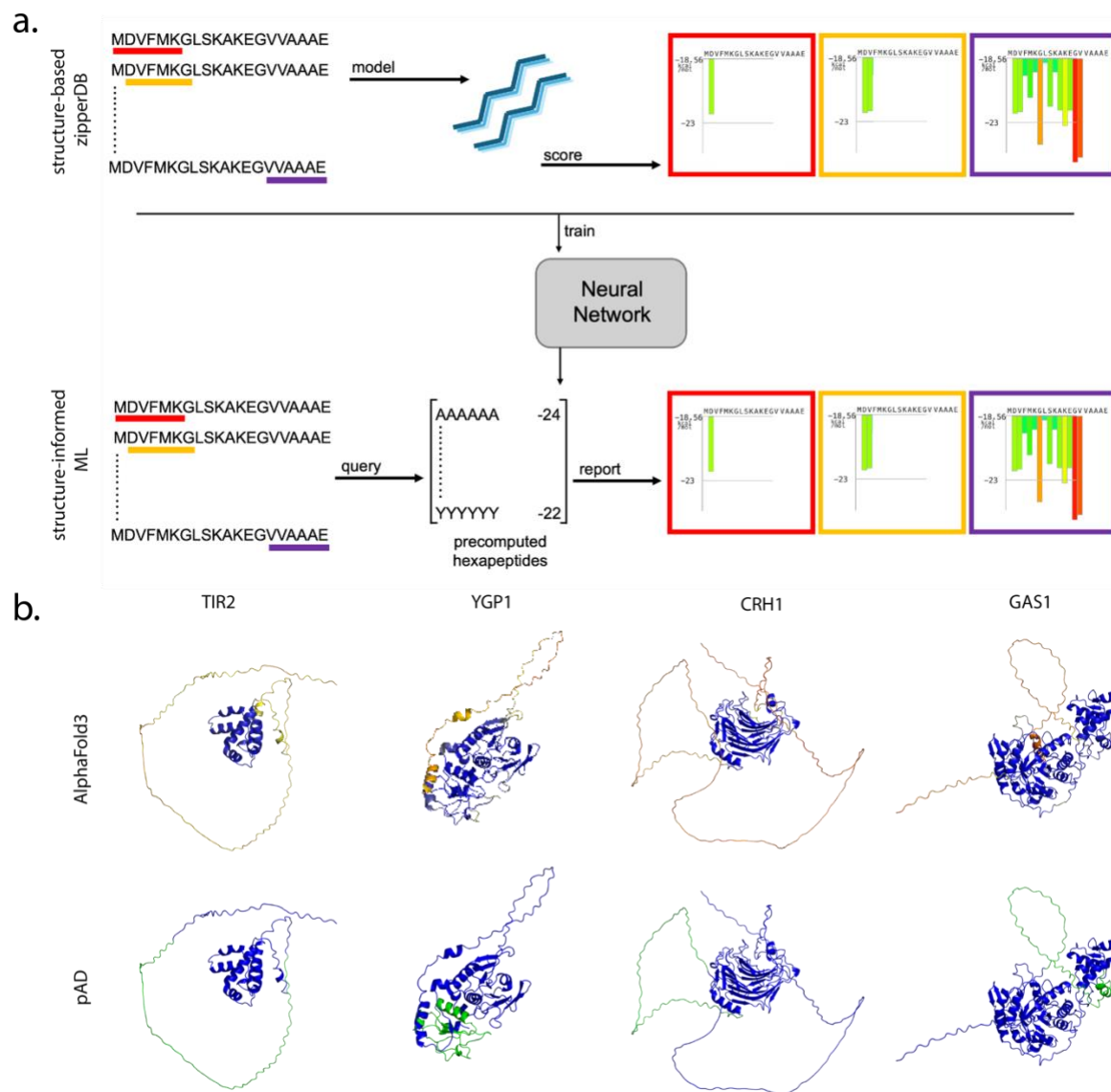


Figure 2-1: a. Comparison of structure-based 3D profiling method to structure-informed machine learning-based approach. Top panel shows structure-based zipperDB methodology. Sequence is divided into hexapeptide segments, individually modeled onto NNQQNY backbone where the energetics are evaluated and finally are reported in the zipper profile. A decade of data from this process (4.1 million hexapeptide sequence/score pairs) were collected and fed into a neural

network which then produced scores for all 64 million possible hexapeptide sequences. The structure-informed method will query the precomputed table to produce the zipper profile. **b.** Candidate examples from *S. cerevisiae* proteome. AlphaFold3 models are shown on top colored by pLDDT scores (blue > 90, yellow < 50). Bottom panel shows pAD highlighted in green.

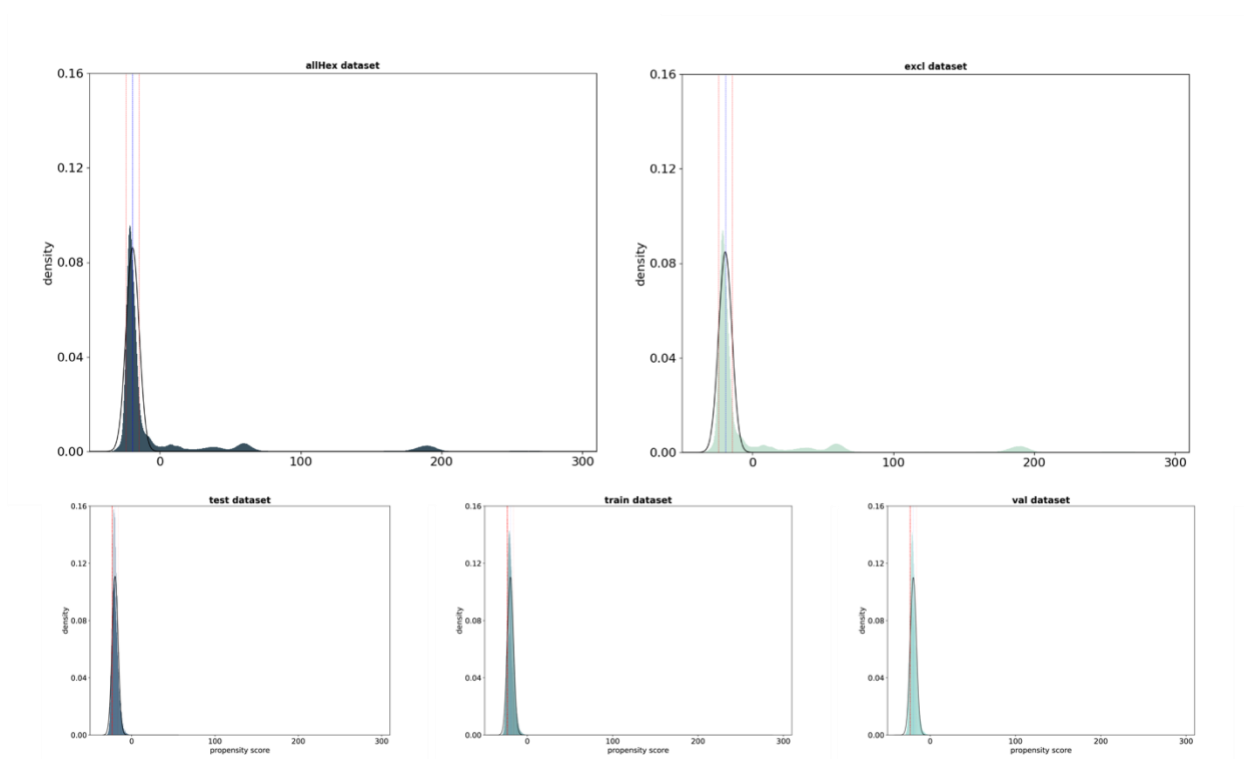


Figure 2-2: Score distribution histograms for each dataset. Histograms were plotted across all five datasets: all hexapeptides (allHex), sequences excluded from the machine learning set (excl) and machine learning sets (train, val, test). The means for scores < 0 across all five sets exhibited overlapping ranges (allHex = -19.27 ± 4.62 , excl = -19.23 ± 4.7 , train = -19.79 ± 3.62 , val = -19.79 ± 3.62 , test = -19.81 ± 3.6).

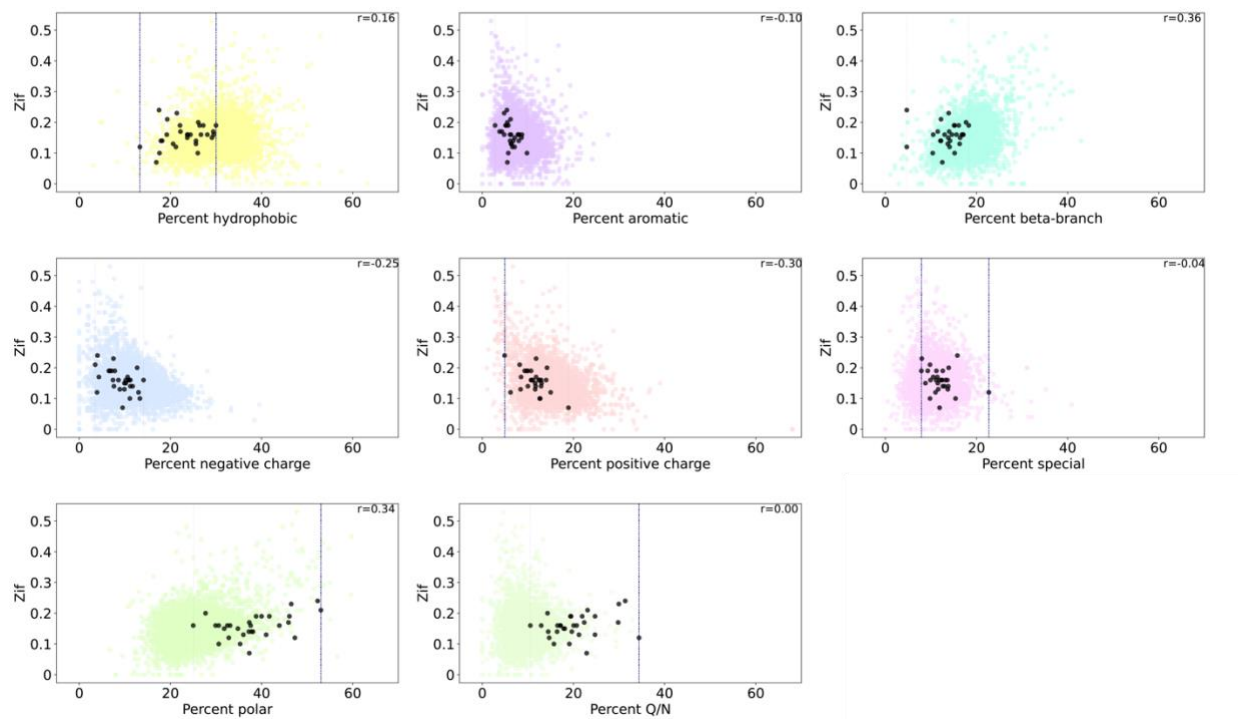


Figure 2-3: Correlation between Zif and percentage of amino acid type for proteins in the *S. cerevisiae* proteome. Black dots represent prospective yeast prion proteins from Alberti dataset. Pearson's correlation coefficients shown. Bounds represent minimum and maximum values.

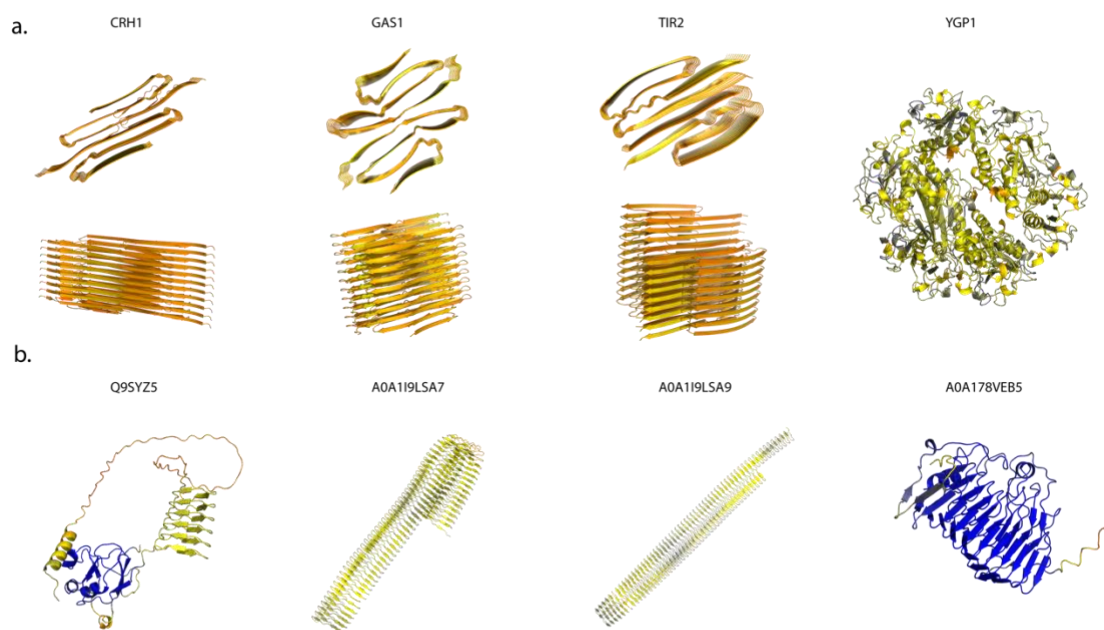


Figure 2-4: AlphaFold3 Structure Predictions. **a.** AF3 models of predicted amyloid domains in top *S. cerevisiae* pAD candidate proteins. **b.** AF3 models of top pAD candidates in *A. thaliana*.

SUPPLEMENTARY FIGURES

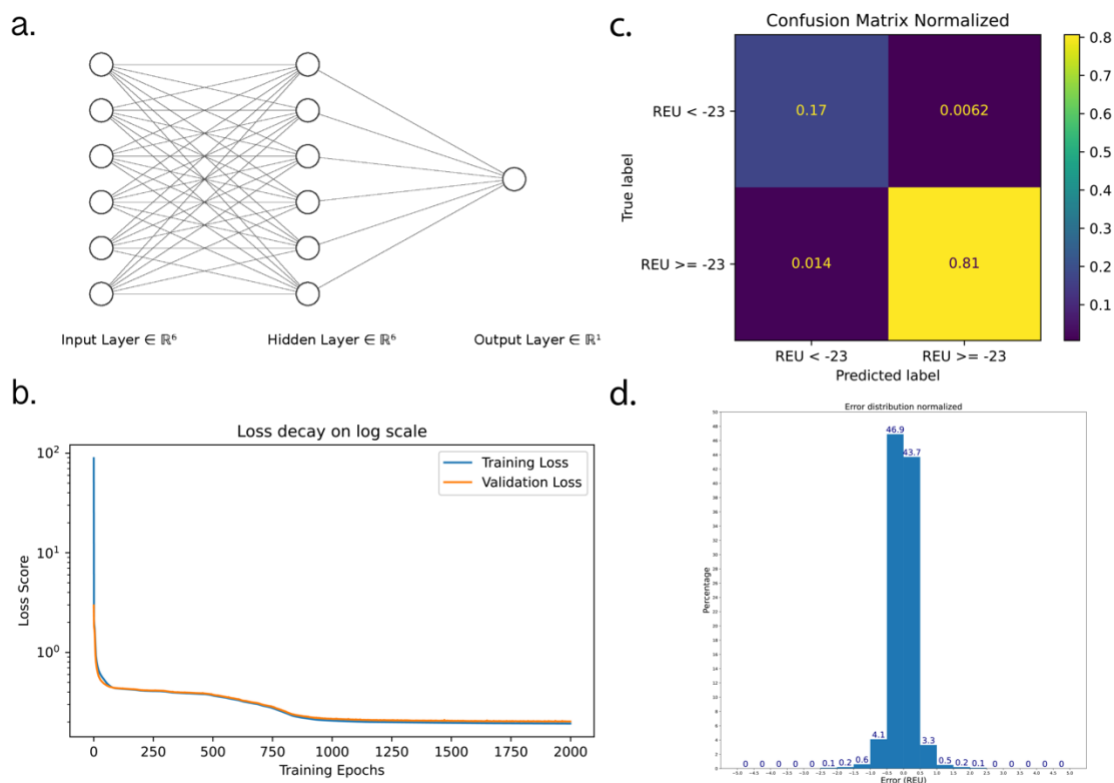


Figure 2S-1: Description and validation of the Structure-Informed Machine Learning Model **a.** schematic diagram depicting architecture of the fully-connected neural network **b.** The decay of loss during training (log scale) shows the model reached convergence after ~1000 epochs of training **c.** Normalized confusion matrix shows the model achieved 98% accuracy. **d.** Normalized error distribution performed on the test set shows that 98% of the error is within ± 1 REU of true score.

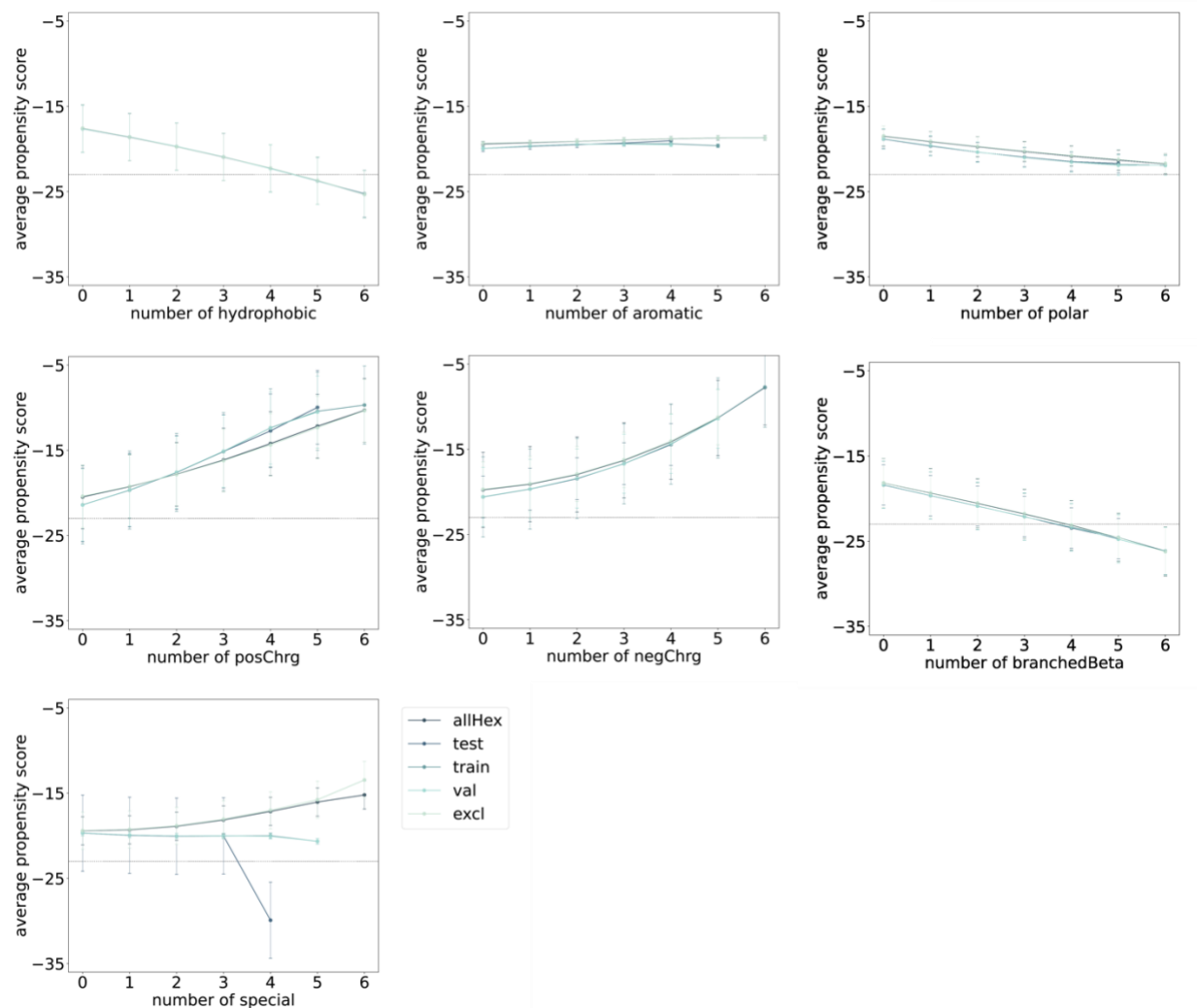


Figure 2S-2: Relationship between amino acid composition and zipper propensity score. The number of hydrophobic, polar, and β -branched residues are correlated with a more favorable zipper arrangement whereas the number of charged and special residues are correlated with a less favorable zipper score.

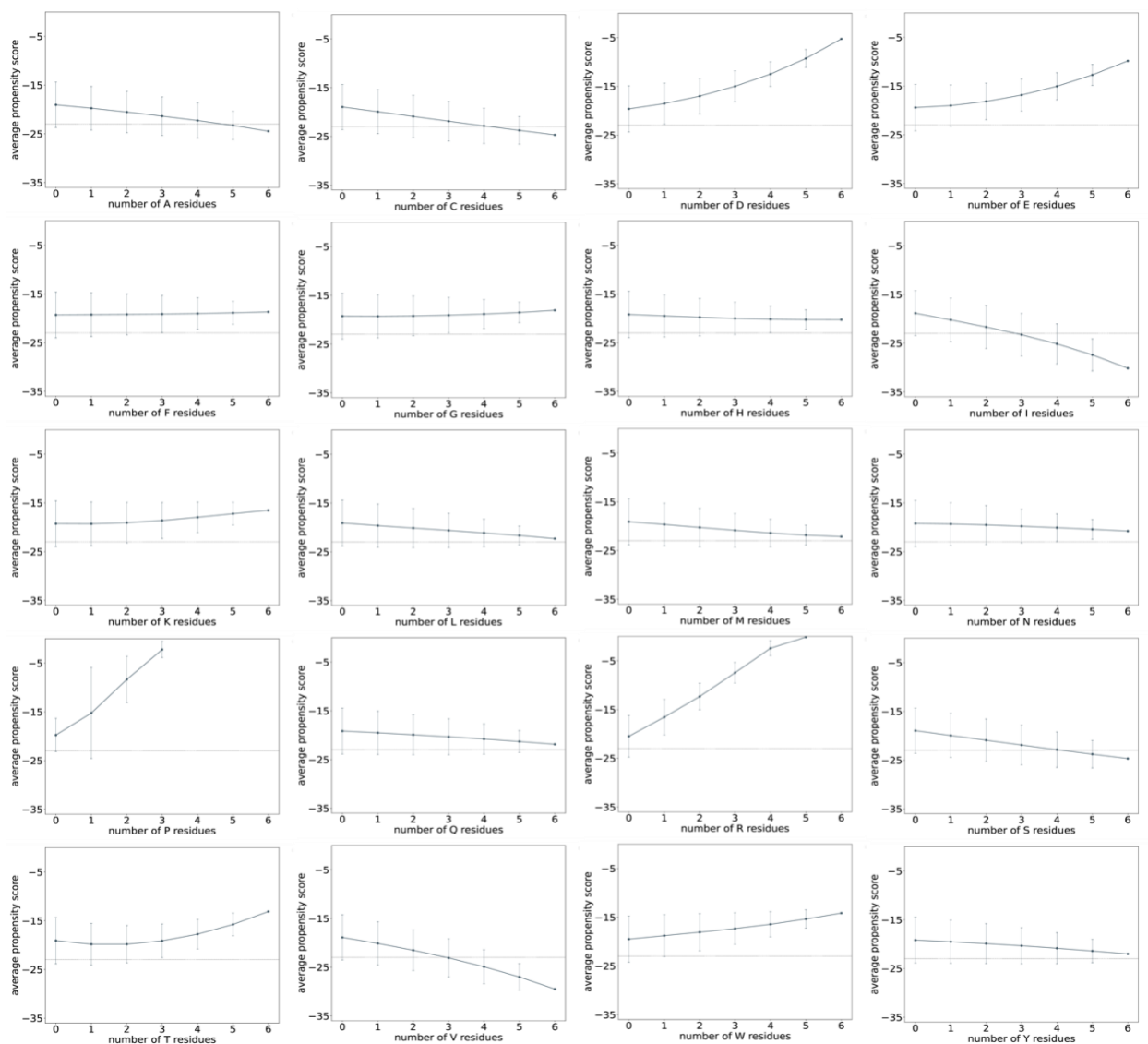


Figure 2S-3: Impact of individual amino acid counts on predicted amyloid propensity scores in hexapeptides. Averaged scores taken for hexapeptides with scores < 0.

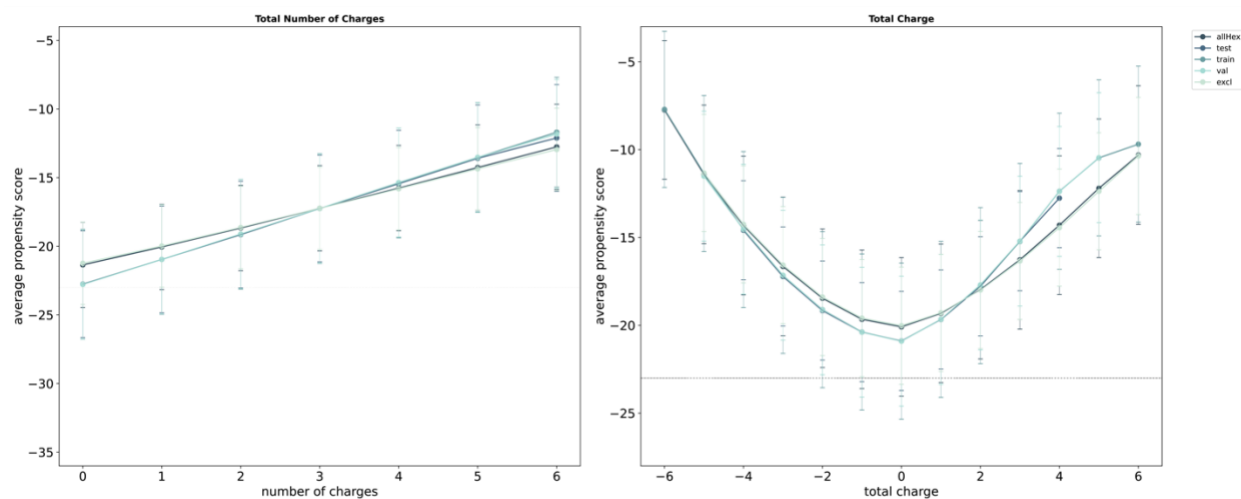


Figure 2S-4: Impact of charged residues in a hexapeptide sequence on zipper propensity scores.

As the number of charged residues or the magnitude of the total charge in the hexapeptide increases, the average propensity score increases.

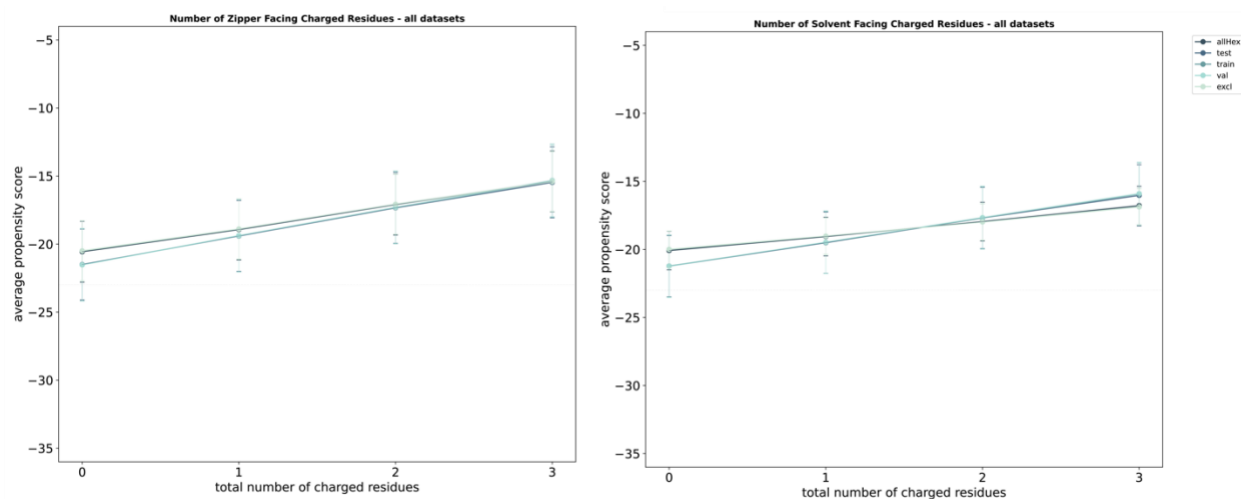


Figure 2S-5: Charged residues at the zipper interface and the solvent interface have similar effects on the hexapeptide's overall zipper propensity score.

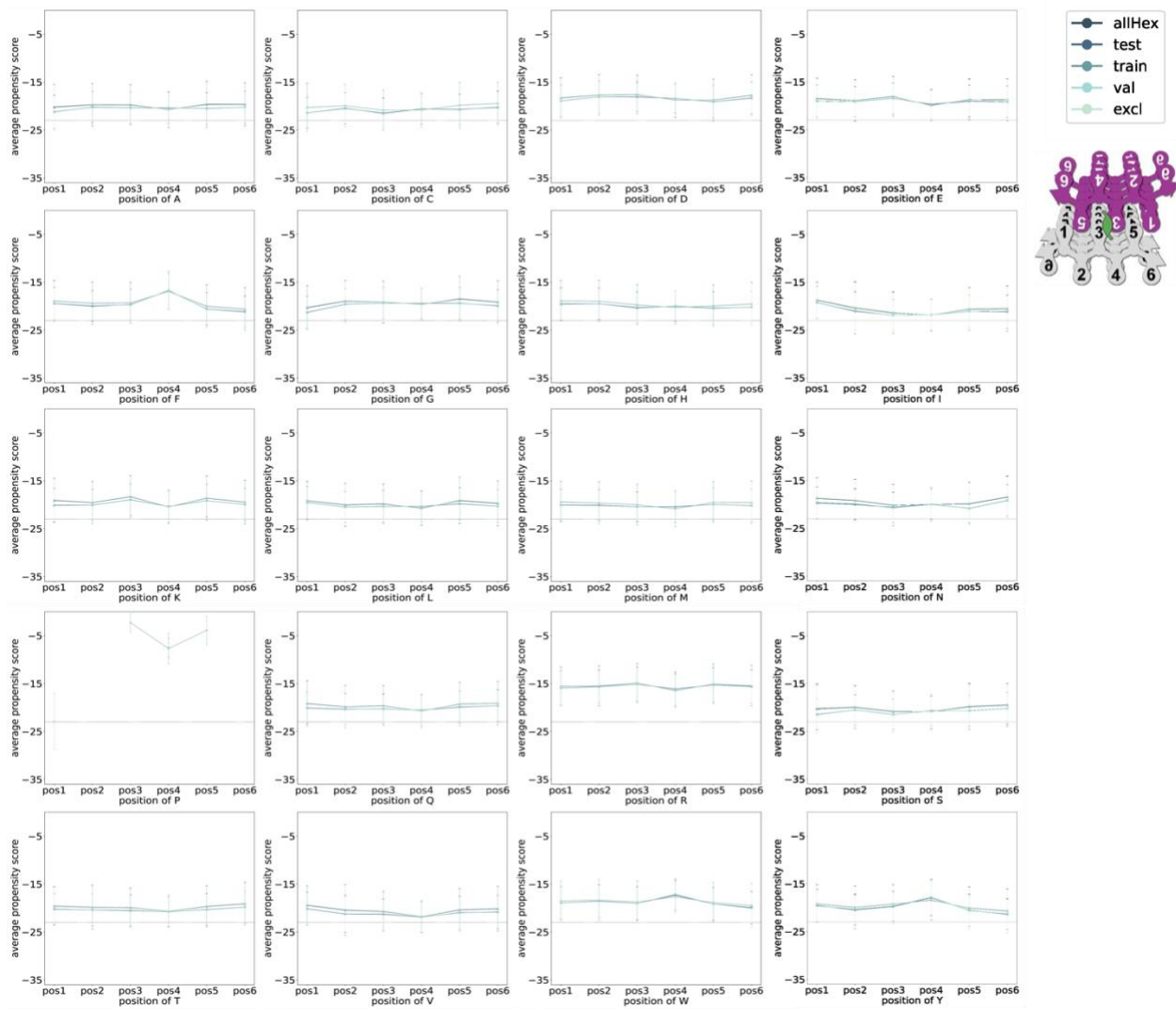


Figure 2S-6: Mean zipper propensity scores for hexapeptides based on amino acid type and position. Schematic represents numeric positions based on Sawaya et al. 2007.⁶

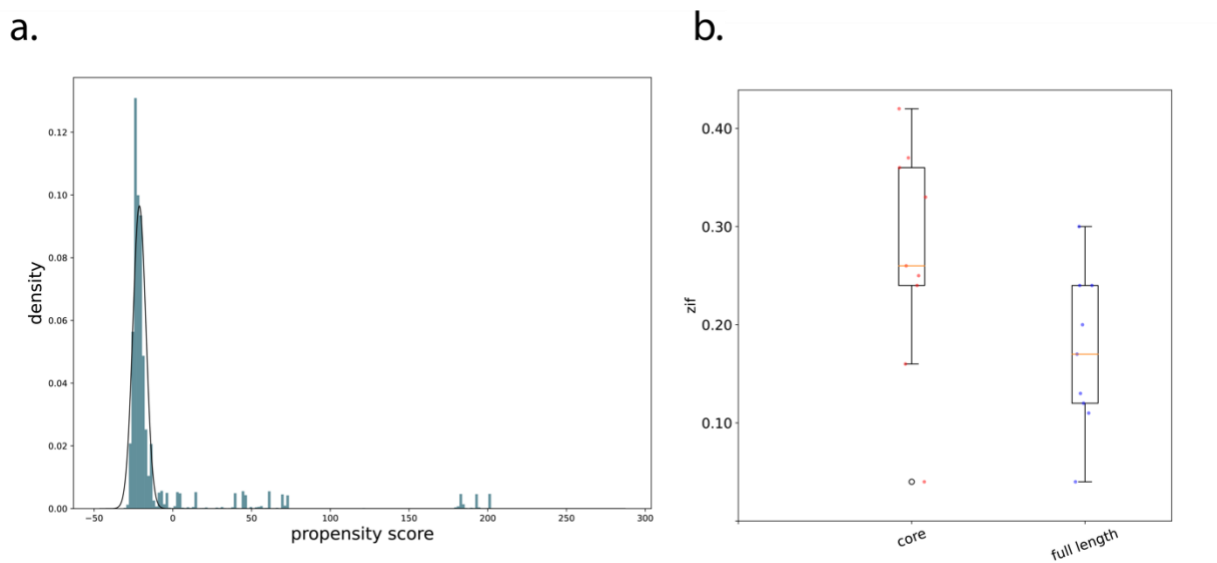


Figure 2S-7: a. Score distribution for hexapeptides present in Amyloid Atlas. The distribution shape resembles that of all hexapeptides, with similar means. Black line represents the probability density function for the Amyloid Atlas dataset. The dataset mean is -21.23 ± 4.13 REU. **b.** zif comparison of the solved core vs full-length proteoform sequences for 9 notable amyloids. The cores, on average, have higher zif scores than their full-length sequences.

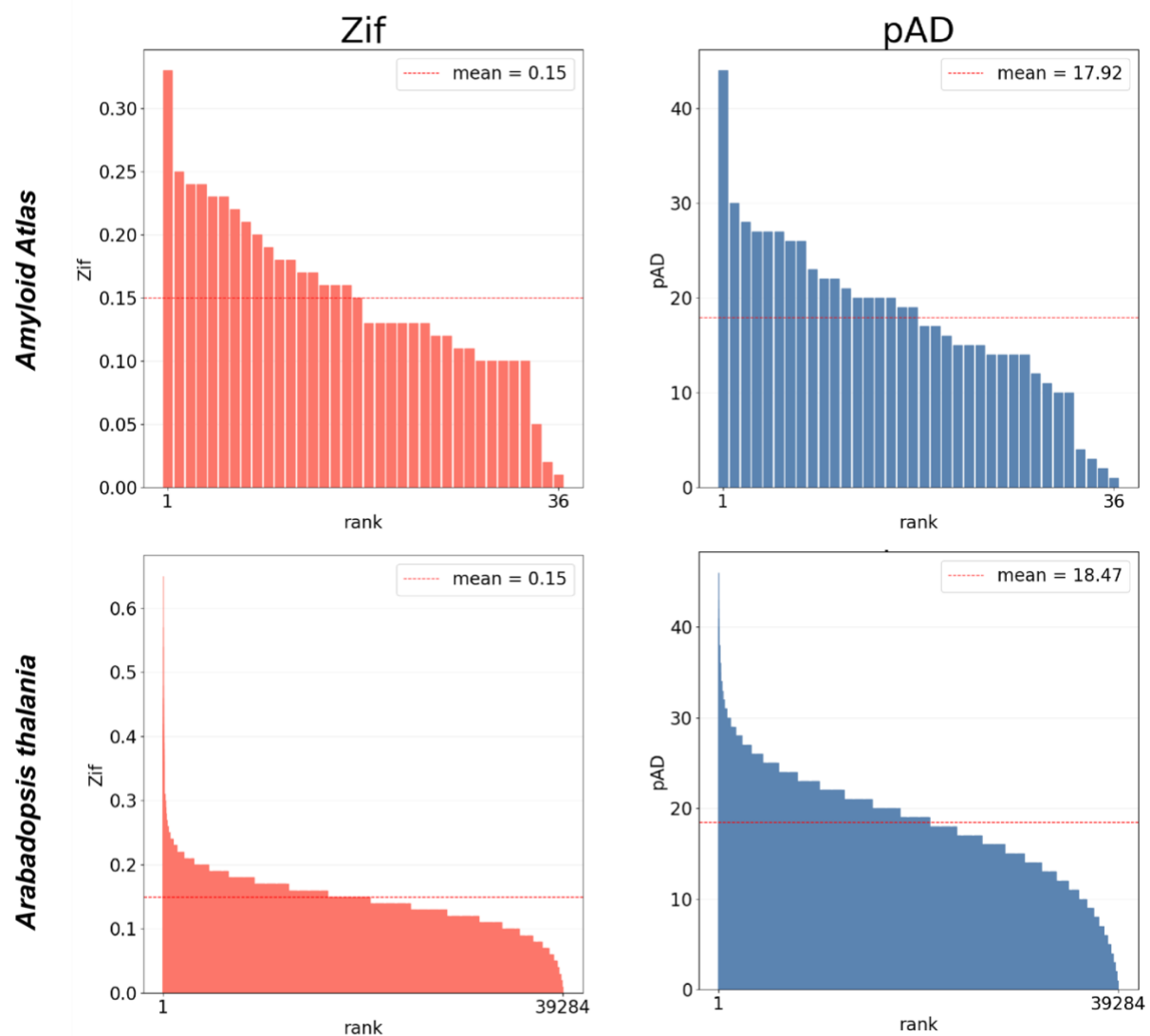


Figure 2S-8: Distribution of ranked Zif and pAD for proteins with solved amyloid fibril structures (see Amyloid Atlas). Red line denotes dataset mean.

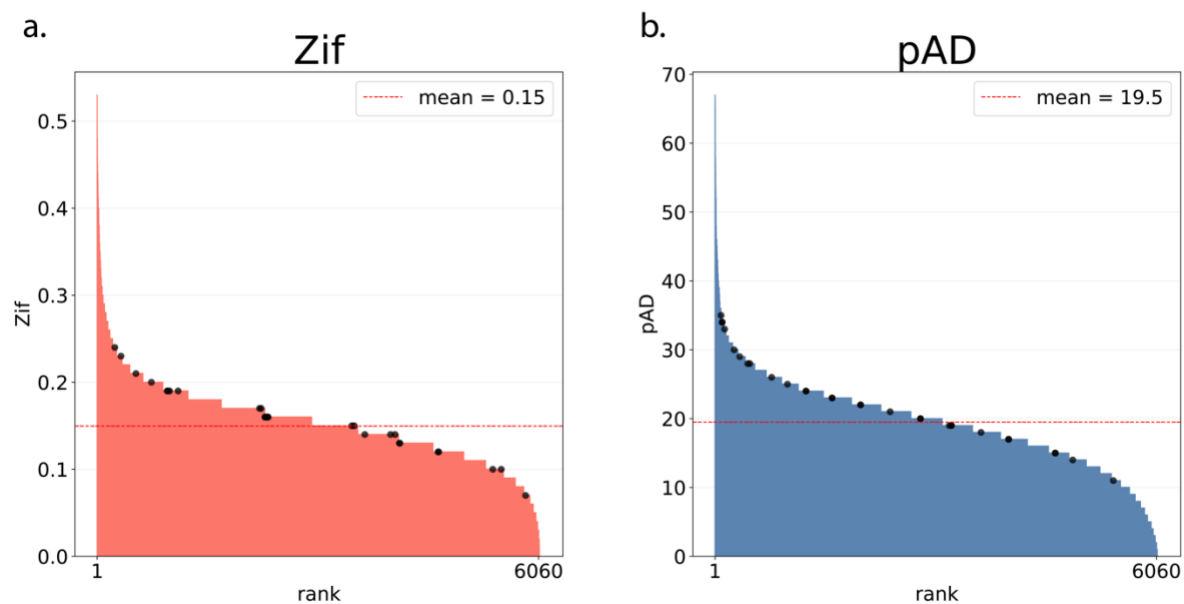


Figure 2S-9: Distribution of ranked Zif and pAD for proteins in *S. cerevisiae* proteome. Black dots represent prospective yeast prion proteins from Alberti dataset. Red line denotes dataset mean.

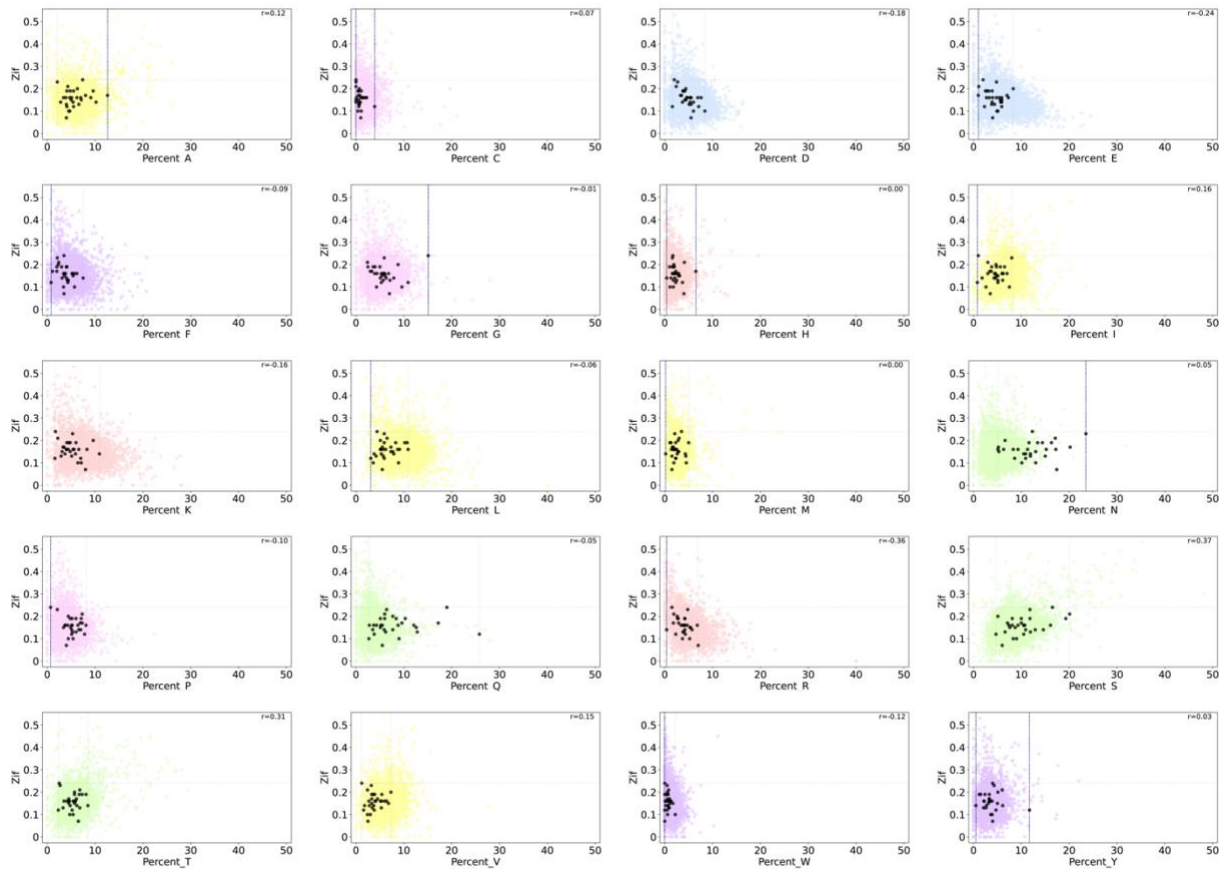


Figure 2S-10: Correlation between Zif and amino acid content for proteins in the *S. cerevisiae* proteome. Pearson's correlation coefficients shown. Bounds represent minimum and maximum values. Black dots represent predicted yeast prions from Alberti dataset.

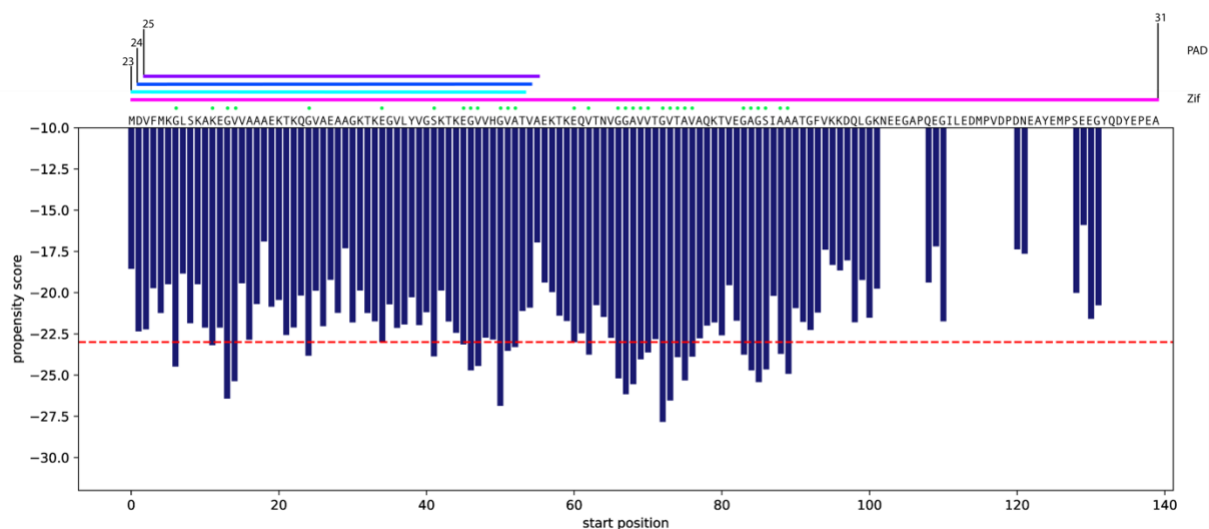


Figure 2S-11: Metric Calculation. Zif is calculated by dividing the number of zipper segments by the total number of hexapeptide segments in the protein (magenta bar) PAD is calculated via a sliding window approach. The number of zipper segments is counted in 75 residue blocks. The reported number is the segment with the highest number of zipper segments. Cyan, blue, and purple bars represent the first three windows of alpha-synuclein (with 23, 24, 25 zipper segments, respectively). Green dots represent segments predicted to form zippers.

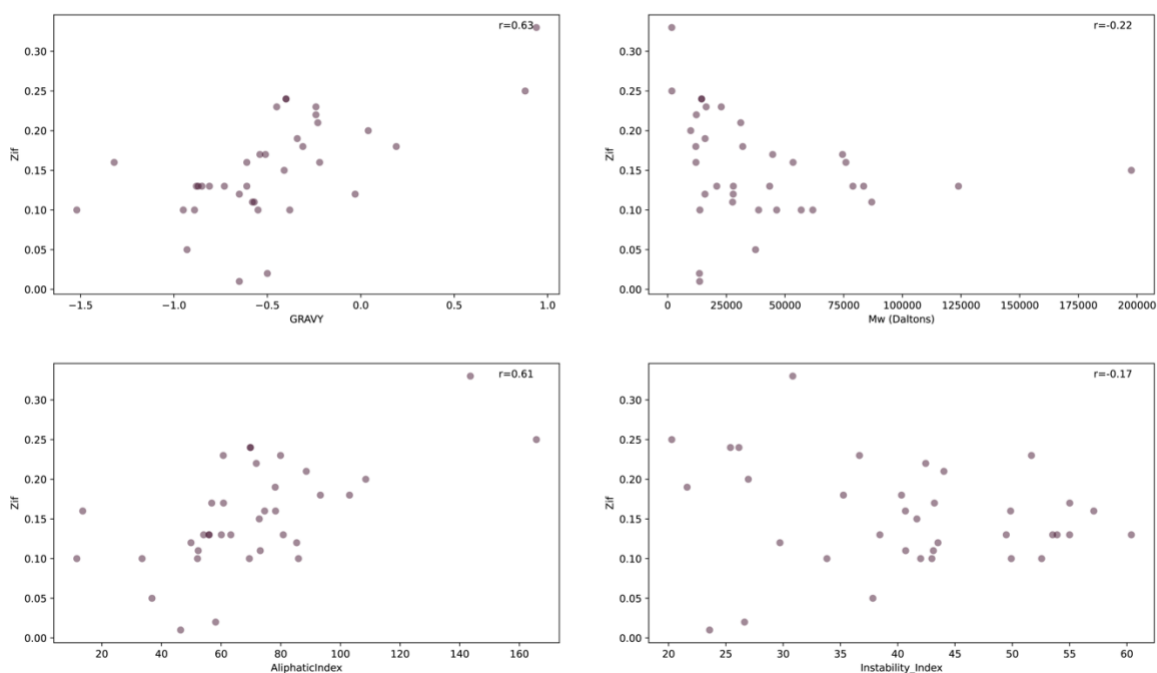


Figure 2S-12: Correlation between Zif and physical metrics for proteins with known amyloid structures (see Amyloid Atlas). Pearson's correlation coefficients shown. Bounds represent minimum and maximum values.

Table 2S- 1: Metrics for proteins in Amyloid Atlas

Accession ID	Gene	Predicted TMD	Zif	pAD
P69020	AUR31	0	0.33	4
P82042	UPE35	0	0.25	3
P37840	SNCA	0	0.24	27
O14960	LECT2	0	0.23	22
P0DOX8	IGGL	0	0.23	21
P01699	IGLV1_44	0	0.22	19
Q9NUM4	TMEM106B	1	0.21	30
P10997	IAPP	1	0.2	17
P00441	SOD1	0	0.19	14
Q03689	het_s	0	0.18	20
P01308	INS	0	0.18	14
Q9VSR3	orb2	0	0.17	26
Q13148	TARDBP	0	0.17	20
P35637	FUS	0	0.16	26
Q13546	RIPK1	0	0.16	20
P01714	IGLV3_19	1	0.16	17
P52948	NUP98	0	0.15	27
O14776	TCERG1	0	0.13	44
P10636	MAPT	0	0.13	28
P23727	PIK3R1	0	0.13	22
Q92734	TFG	0	0.13	19
P04925	Prnp	1	0.13	16
P01275	GCG	0	0.13	12
P04273	PRNP	1	0.12	15

P02766	TTR	1	0.12	11
P05067	APP	1	0.11	20
P04156	PRNP	2	0.11	15
Q92804	TAF15	0	0.1	23
O14979	HNRNPDL	0	0.1	15
P09651	HNRNPA1	0	0.1	14
Q9Y572	RIPK3	0	0.1	14
P61769	B2M	1	0.1	10
P22626	HNRNPA2B1	0	0.05	10
P0DJI8	SAA1	2	0.02	2
P05367	Saa2	3	0.01	1

CHAPTER 3

Machine learning-aided design of type-1 homozippers with fibril-surface binding potential

Introduction

Amyloidoses are a set of proteopathies associated with the accumulation of insoluble protein fibrils in tissues.¹ On the macromolecular scale, protein monomers form amyloid fibrils by adopting a nearly 2-dimensional fold and stacking on top of each other along the fibril growth axis. Each fibril layer is β -strand rich and can contain one or more polypeptide chains that form protofilaments with compact cores; multiple protofilaments can form part of a single fibril, aligning along the fibril axis at defined interfaces. Amyloid fibrils grow through nucleation-dependent polymerization, similar to that of actin/microtubule assembly and protein crystallization.² The initial step of fibril growth is dependent on the formation of a nucleus, a rare event due to the thermodynamic unfavourability. Once formed, however, polymerization proceeds rapidly where individual protein monomers align along a single dimension to form fibrils that can contain millions of stacked layers.

There are instances in which fibril growth can be accelerated. This can occur when point mutations shift the predisposition of the protein to adopt an amyloid fold, as is thought to be the case with familial mutations.³ Seeding, a phenomenon where the presence of a preformed amyloid fibril induces the misfolding of monomer, is another instance in which acceleration is enhanced. This process can occur between members of the same protein or in some cases between different proteins (cross-seeding).^{4,5} The exact mechanism behind this phenomenon is not currently understood.

Typically, canonical proteins fold into a single structure at the global free-energy minima. Pathogenic amyloids, however, can adopt multiple conformations in the fibril state. These amyloid polymorphs adhere to the same structural principles, but differ in their amino acid contacts, number of protofibrils, or both.⁶ Environmental conditions can influence which polymorph predominates in solution, both *in vitro* and *in vivo*.⁶ Additionally, studies in mice suggest polymorphs can have different levels of toxicity as well as differences in seeding and propagation properties similar to that of prion strains.⁷ This emphasizes the importance of understanding the molecular mechanisms driving these observed behaviors in treatment of amyloid-related diseases.

α -synuclein (α Syn) is a 140 residue, intrinsically disordered protein that has been implicated in a set of neurodegenerative diseases known as synucleinopathies. To date, 24 unique polymorphs of α Syn have been solved.⁶ These structures exhibit conservation of β -strand pieces but differ in connectivity, as evident in the wild-type structures solved by Li et al, 2018.⁸ Here, two polymorphs, labeled rod and twister, share a conserved bent β -arch that pack the preNAC and NACore zipper interfaces, respectively. These two regions are of particular importance due to their implication in Parkinson's Disease (PD). The preNAC region contains many of the familial mutations associated with the early onset of PD while the NAC region has been demonstrated to be both essential for and capable of inducing α Syn aggregation and its associated toxic effects.^{3,9,10}

In the present work, we utilize strategies from protein design to gain insight into the molecular mechanisms behind cross-seeding with α Syn. We designed 6-residue peptides with the idea that they had the ability to form type-1 homozippers on the surface of α Syn. We investigated to what degree (or not) are the peptides (1) stabilizing/destabilizing the nucleus, (2) promoting fibril stability, or (3) promoting secondary nucleation through a set of biochemical and structural experiments.

Materials and Methods

Computational design of surface binding zippers

The solved fibril structures of α Syn (pdb 6cu7, 6cu8) were used to identify regions on the surface of either fibril polymorph that could accommodate the formation of a type-1 homozipper. Two candidate regions were selected, with three potential hexapeptide arrangements. The zipper forming residues were held fixed while the solvent facing residues were computationally mutated. The set of peptides was assembled through enumeration of all possible combinations of amino acids (excluding C, P, R, and W) at positions 2, 4, and 6. The resultant set was scored for zipper propensity via the zipperDB.

Peptide crystallization

B1_R115 (GNAQVY) peptides were purchased from GenScript and dissolved at 10 mg/mL in ultrapure water. A ‘Wizard’ crystallographic screen (Rigaku) was used for initial crystallization efforts. Diffraction quality peptide crystals grew after 24 hours in 2000 mM ammonium sulfate, 100 mM sodium citrate/citric acid at pH 5.5 and a 1:1 volume of peptide:buffer.

Peptide structure determination

Crystals of B1_R115 were harvested from hanging drops and prepared for microED analysis by depositing ~2 microliters of crystal slurry onto a Quantifoil carbon support on a copper mesh grid and blotting excess liquid using a vitrobot (Thermo). Diffraction data were collected on a Talos F200C side-entry transmission electron microscope (Thermo-Fisher Scientific) operating at 200KeV equipped with an Apollo direct electron detector. Diffraction was observed to ~0.9-1.0 Å

resolution. Data were processed in the CCP4 Suite, direct methods tools from SHELX were used to generate an initial model, which was further refined in PHENIX and REFMAC.

ThT Assay

Wild-type α Syn monomer (175 μ M) was mixed with 20 μ M thioflavin-T (ThT) and 5 molar excess of A1_R9, B1_R115 and B1_R590 peptide in a 96-well plate. The signal was monitored using the FLUOstar Omega Microplate Reader (37 °C with 600 rpm orbital shaking, ex. 448-10 nm, em. 482-10 nm; BMG Labtech).

Peptide Docking

Peptide docking was performed as follows. The crystal structure of B1_R115 was aligned to the surface of the rod polymorph (pdb 6cu7) in PyMol. Energy minimization was performed in Rosetta using FastRelax with 6 Å distance constraints between C α atoms in the peptide and the suspected binding region on the surface of the fibril.

Results and Discussion

High-throughput computation peptide design

To design peptide surface binders, we first identified regions on either the rod or twister polymorph of wild-type α Syn that could form a type-1 homozipper (pdb = 6cu7, 6cu8).⁸ Two regions were identified: region (A) ⁵⁸KTKEQVTNV⁶⁶ and (B) ⁵¹GVATVA⁵⁶ (Figure 1a). Region A is exposed to solvent in both the rod and twister polymorph. Region B, however, is buried at the protofilament interface in the rod polymorph and exposed in the twister polymorph. Region A was

broken down into two hexapeptide segments (A₁: ⁵⁸KTKEQV⁶³, A₂: ⁶⁰KEQVTN⁶⁵), yielding three peptide scaffolds to design. Surface residues on the fibril were held constant, while interior facing residues were enumerated with all possible combinations of amino acids (excluding C, P, R, and W), totaling 12,288 candidate peptides. Candidate peptides were scored for zipper propensity (Figure 1b). Final peptides were chosen based on a favorable zipper propensity score as well as solubility (Figure 1b,c).

Zippers A₁ and A₂ had higher overall zipper propensity scores (Figure 1b) and lower hydrophobicity scores (Figure 1c) than zipper B. This is unsurprising as the B peptide has a baseline hydrophobicity score of 5.6 due to the homozipper interface (amino acids G, A, V) whereas zipper A has a baseline hydrophobicity score of -7.4 with the zipper interface forming with amino acids K, K, and Q. As noted in chapter 2, there is a correlation between more favorable zipper scores and an increase in the hydrophobicity of the peptide sequence. Three peptide segments were chosen for biochemical and structural characterization due to their favorable zipper scores and calculated hydrophobicity: KTKIQY (A1_R9), GNAQVY (B1_R115), and GNAKVY (B1_R590).

A preliminary structure solution of peptide B1_R115

We performed crystallization studies on the peptides to assess the ability of the designed peptides to form steric zippers. Needle-like crystals grew in B1_R115 (GNAQVY) condition. Subsequently, high-resolution diffraction data was collected and a 1.0 Å structure was solved via direct methods. B1_R115 crystallized as a parallel class-1 steric zipper. The interface between fibrils forms favorable hydrophobic interactions with residues Gly1, Ala3, and Val5, mimicking the interface between fibrils in the rod polymorph.

Kinetic effects of designed peptides on fibrillization

We first wanted to understand the effect the presence of the peptide had on the fibrillization kinetics of wild-type (WT) α Syn. We performed a thioflavin-T (ThT) assay with 175 μ M wt α Syn alone, as well as in the presence of 5x μ M excess of each peptide (Figure 3a). At a 5:1 peptide: α Syn stoichiometric ratio, the inflection point is observed at earlier times indicating that fibrillization is occurring more rapidly in the presence of the peptide. Using the $t_{1/2}$ as a point of comparison, B1_R590 had the largest effect on the kinetics and decreased the lag time 8-fold, whereas B1_R115 and A1_R9 decreased the lag time 5-fold and 2.6-fold, respectively. There was no observed fluorescence for the peptide alone. Additional kinetic experiments were performed at 1x and 0.2x μ M excess of each peptide (Figure 4). Increased fibrillization kinetics was observed for 1x μ M excess of each peptide but decreased fibrillization kinetics was observed for 0.2x μ M excess.

Several potential mechanisms could contribute to the observed reduction in the lag phase of fibrillization induced by the presence of molar excess of the peptide. Firstly, the peptide may promote fibril stability, enhancing overall thermodynamic favorability. Secondly, the peptide could stabilize the amyloid nucleus, lowering the energetic barrier associated with the critical nucleation event. Thirdly, the peptide may act as a catalyst for secondary nucleation processes. However, the extent to which each of these mechanisms contributes remains unknown. To elucidate the underlying mechanistic details, a systematic investigation aimed at delineating the individual contributions of these potential pathways will need to be pursued.

Morphological screening of fibrils grown in the presence of peptide

As an initial screen to identify the potential implications of the observed change in kinetics upon addition of peptide, we looked at the fibrils formed via negative-stain cryogenic electron

microscopy (cryo-EM) (Figure 3b). Wild-type fibrils alone exhibit a mix between slow and fast twisting polymorphs, presumably the rod and twister polymorphs observed by Li et al., 2018.⁸ Upon addition of both B1_R115 and B1_R590, however, there is an apparent shift in the morphology of fibrils formed where there is an observed increase in the number of fast twisting fibrils. The fibrils present in the A1_R9 condition, however, are more similar to what is observed for wild-type.

If peptide binding is occurring as designed, the observed polymorph shift is expected. Peptide B was designed to bind at the protofilament interface of the rod polymorph (Figure 1a). This interaction would prevent the interface between protofilaments from forming, potentially favoring the formation of the twister polymorph. Peptide A, however, is solvent-exposed in both the rod and twister polymorphs (Figure 1a), allowing it to bind compatibility to either polymorph. It should be emphasized that negative-stain cryo-EM can be used as an initial screen to identify potential trends, but more thorough methods will need to be perused to confirm and quantify the presence and distribution of polymorphs in the sample.

Peptide Docking

To understand if the peptide can stably interact with the target region on the surface of the fibril, we performed peptide docking experiments with B1_R115 and the rod polymorph of α Syn. The solved peptide crystal structure was aligned to the surface of the fibril. Distance constraints were applied between the peptide and the corresponding binding region on the fibril. The docked peptide and fibril were then subjected to an energy minimization cycle in Rosetta.

The resultant structure indeed exhibited a decrease in energy (~ -800 REU), suggesting a potentially stable interaction; however, caution is warranted in this interpretation. Several factors inherent to modeling in Rosetta can lead to false positives where energetically favorable conformations *in silico* may not accurately represent physiological interactions. Therefore, experimental validation is crucial to confirm the predicted interactions.

Concluding Remarks

Our research strategy involved the deliberate selection of peptides with potential binding affinity to specific fibril interfaces. We hypothesized that this approach might modify the properties of the fibrils that are formed, either through alterations in their structure and/or growth kinetics. We did not anticipate the way in which they would be altered, but what was observed is that many of the peptides that were supposed to interact with the surface, also accelerated growth. Now we suspect that it could be due to stabilization of the fibrils or the formation of additional nucleation sites. However, these hypotheses require further validation through additional experimental studies.

References

1. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).
2. Jarrett, J. T. & Lansbury, P. T. Seeding “one-dimensional crystallization” of amyloid: A pathogenic mechanism in Alzheimer’s disease and scrapie? *Cell* **73**, 1055–1058 (1993).
3. Spillantini, M. G. & Goedert, M. Neurodegeneration and the ordered assembly of α -synuclein. *Cell Tissue Res.* **373**, 137–148 (2018).
4. Andreetto, E. *et al.* Identification of Hot Regions of the A β –IAPP Interaction Interface as High-Affinity Binding Sites in both Cross- and Self-Association. *Angew. Chem. Int. Ed.* **49**, 3081–3085 (2010).
5. Sampson, T. R. *et al.* A gut bacterial amyloid promotes α -synuclein aggregation and motor impairment in mice. *eLife* **9**, e53111 (2020).
6. Sawaya, M. R., Hughes, M. P., Rodriguez, J. A., Riek, R. & Eisenberg, D. S. The expanding amyloid family: Structure, stability, function, and pathogenesis. *Cell* **184**, 4857–4873 (2021).
7. Bousset, L. *et al.* Structural and functional characterization of two alpha-synuclein strains. *Nat. Commun.* **4**, 2575 (2013).
8. Li, B. *et al.* Cryo-EM of full-length α -synuclein reveals fibril polymorphs with a common structural kernel. *Nat. Commun.* **9**, 3609 (2018).
9. Giasson, B. I., Murray, I. V. J., Trojanowski, J. Q. & Lee, V. M.-Y. A Hydrophobic Stretch of 12 Amino Acid Residues in the Middle of α -Synuclein Is Essential for Filament Assembly*. *J. Biol. Chem.* **276**, 2380–2386 (2001).
10. Du, H.-N. *et al.* A Peptide Motif Consisting of Glycine, Alanine, and Valine Is Required for the Fibrillization and Cytotoxicity of Human α -Synuclein. *Biochemistry* **42**, 8870–8878 (2003).

Figures

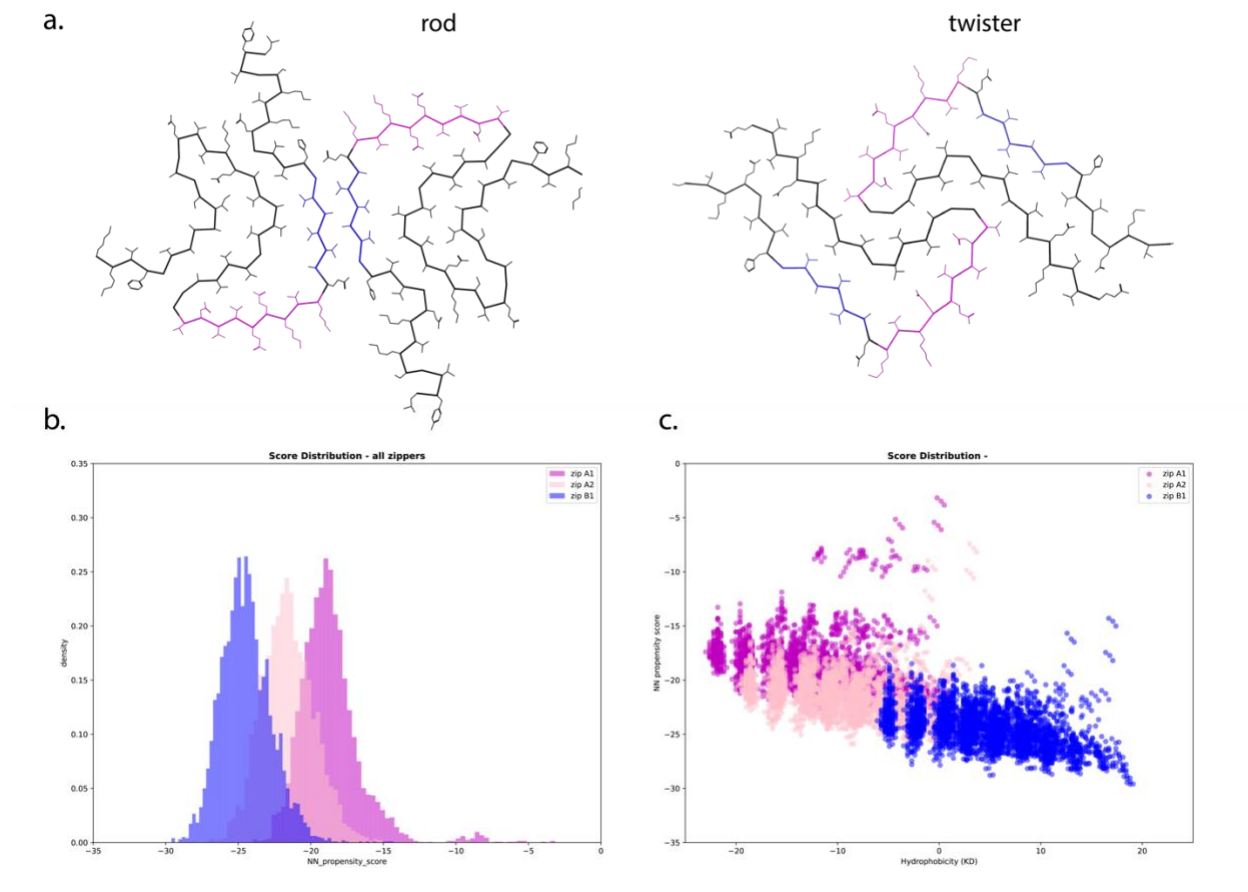
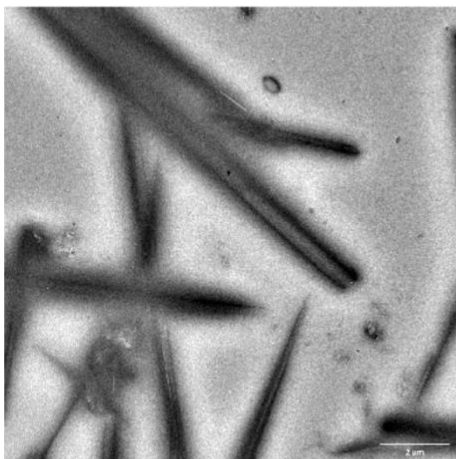


Figure 3-1: Computation design and identification of candidate surface binders. **(a)** Wild-type α -synuclein rod (6cu7) and twister (6cu8) polymorph structures were used to identify surface exposed regions to which potential surface binder peptides could form favorable type-1 homozippers; region A is shown in magenta, region B is shown in blue. **(b)** Histograms of zipper propensity scores for all designed candidate peptides; region A1 shown in magenta, region A2 shown in pink, region B shown in blue. **(c)** Hydrophobicity vs. zipper propensity scores for regions A1, A2, and B1.

a.



b.

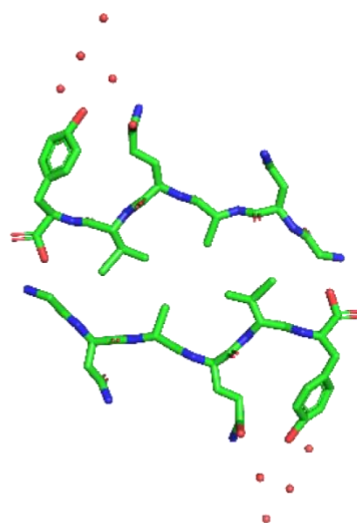


Figure 3-2: Crystallization of GNAQVY peptide. **(a)** Electron micrograph of microcrystals formed from GNAQVY peptide. **(b)** Crystal structure of designed peptide GNAQVY.

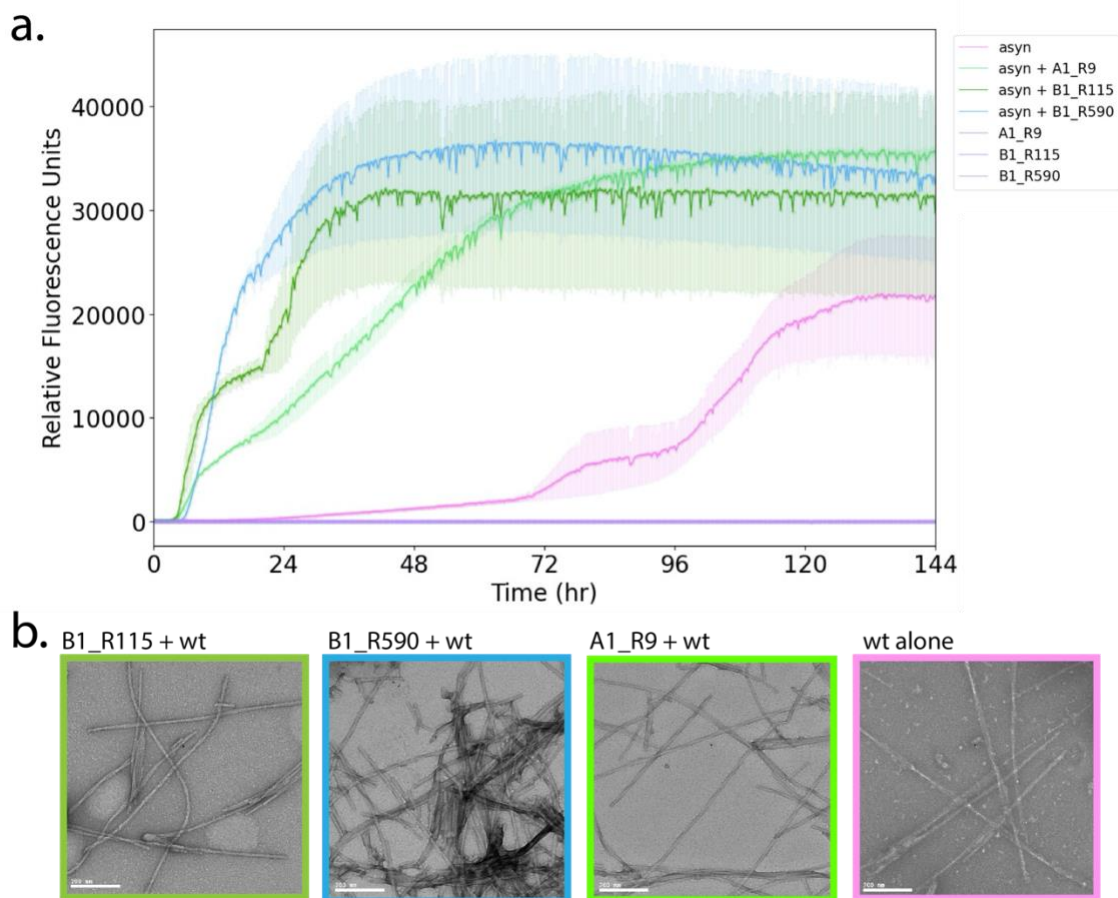


Figure 3-3: (a) Kinetic profiles shows a decrease in lag phase of α -synuclein aggregation in the presence of designed peptides. No fluorescence is observed for peptides alone. **(b)** Negative-stain EM micrographs of α -synuclein fibrils grown with and without peptides.

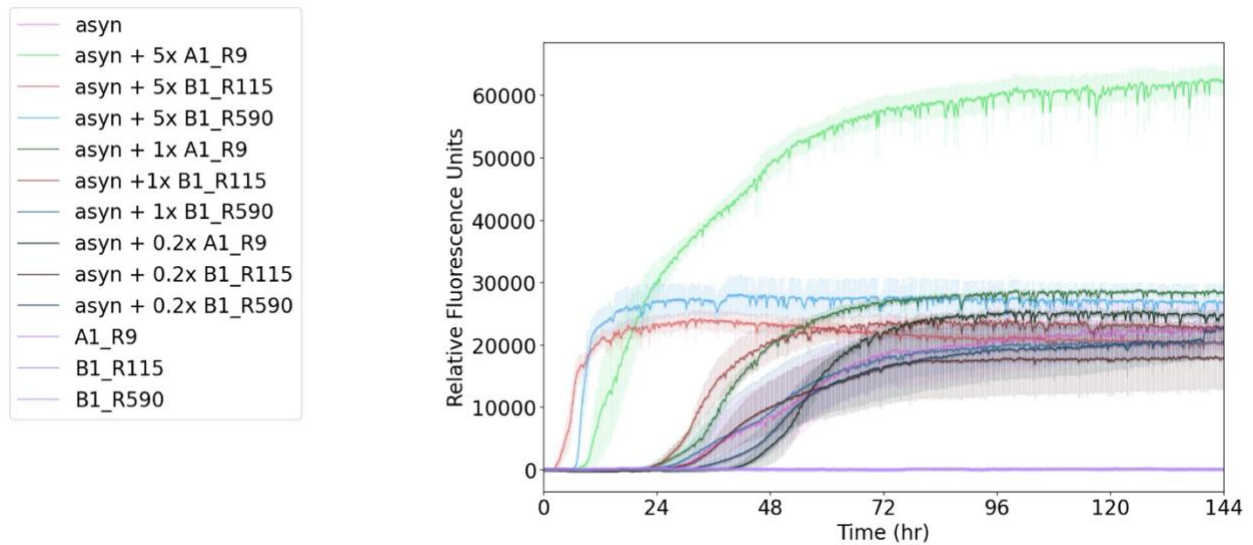


Figure 3-4: Dose dependance on fibrillization kinetics of α -synuclein aggregation in the presence of 5x, 1x, and 0.2x μ M excess of each peptide. Fibrillization kinetics is enhanced for 5x and 1x μ M excess and decreased in 0.2x conditions.

Supplementary Data

Table 3S-1: Preliminary structure solution for B1_R115 peptide

Sequence	GNAQVY
Space group	No. 4 (P 1 21 1)
Cell length constants (Angstrom)	a=15.56, b=4.73, c= 23.83
Cell angle constants (Degrees)	α =90.00, β =103.397, γ =90.00
Completeness	96.5%
Resolution (Angstrom)	0.90
R factor at outermost shell	39.2%
CC1/2 (all, outermost shell)	99.1 (67.5)
R factor overall	12.7%
R1 (SHELXL)	18.78%
R-work (Refmac)	31.12%
R-free (Refmac)	32.06%

Appendix Chapter 1

This chapter is based on the publication titled “Cryo-EM Structure of a Human LECT2 Amyloid Fibril Reveals a Network of Polar Ladders at its Core”.

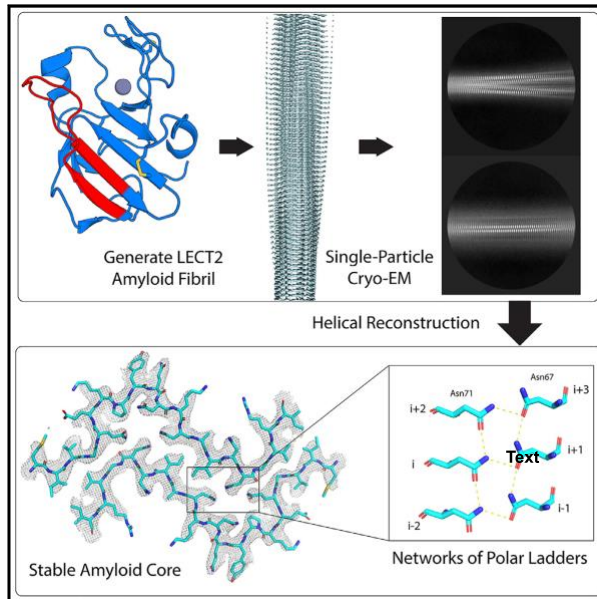
Richards, L. S., Flores, M. D., **Zink, S.**, Schibrowsky, N. A., Sawaya, M. R., &

Rodriguez, J. A. (2023). Cryo-EM Structure of a Human LECT2 Amyloid Fibril Reveals a Network of Polar Ladders at its Core. *Structure*, 31(11) 1386-1393.

Structure

Cryo-EM structure of a human LECT2 amyloid fibril reveals a network of polar ladders at its core

Graphical abstract



Authors

Logan S. Richards, Maria D. Flores, Samantha Zink, Natalie A. Schibrowsky, Michael R. Sawaya, Jose A. Rodriguez

Correspondence

jrodriguez@mbi.ucla.edu

In brief

ALECT2 systemic amyloidosis is an amyloid disease characterized by the deposition of amyloid fibrils in the kidneys, leading to renal failure. Richards et al. use cryo-EM to determine the high-resolution amyloid structure of the human protein LECT2, providing structural insight upon which future treatments for ALECT2 amyloidosis can be developed.

Highlights

- Determines a high-resolution amyloid structure of the human protein LECT2 by cryo-EM
- Identifies key regions of interaction for ALECT2 amyloid stability and aggregation
- Gives basis for the rational design of aggregation inhibitors against ALECT2



Richards et al., 2023, *Structure* 31, 1386–1393
November 2, 2023 © 2023 The Author(s). Published by Elsevier Ltd.
<https://doi.org/10.1016/j.str.2023.08.007>





Article

Cryo-EM structure of a human LECT2 amyloid fibril reveals a network of polar ladders at its core

Logan S. Richards,¹ Maria D. Flores,¹ Samantha Zink,¹ Natalie A. Schibrowsky,¹ Michael R. Sawaya,¹ and Jose A. Rodríguez^{1,2,*}

¹Department of Chemistry and Biochemistry, UCLA-DOE Institute for Genomics and Proteomics, STROBE, NSF Science and Technology Center, University of California, Los Angeles (UCLA), Los Angeles, CA 90095, USA

²Lead contact

*Correspondence: jrodriguez@mbi.ucla.edu

<https://doi.org/10.1016/j.str.2023.08.007>

SUMMARY

ALECT2 systemic amyloidosis is associated with deposition of the leukocyte cell-derived chemotaxin-2 (LECT2) protein in the form of fibrils. In ALECT2 amyloidosis, ALECT2 fibrils deposit in the glomerulus, resulting in renal failure. Patients lack effective treatment options outside of renal transplant or dialysis. The structure of globular LECT2 has been determined but structures of ALECT2 amyloid fibrils remain unknown. Using single-particle cryo-EM, we find that recombinant human LECT2 forms robust twisting fibrils with canonical amyloid features. ALECT2 fibrils contain two mating protofilaments spanning residues 55–75 of the LECT2 sequence. The geometry of the ALECT2 fibril displays features in line with other pathogenic amyloids. Its core is tightly packed and stabilized by both hydrophobic contacts and hydrogen-bonded uncharged polar residues. The robustness of ALECT2 fibril cores is illustrated by their resistance to denaturants and proteases. This ALECT2 fibril structure presents a potential new target for treatments against ALECT2 systemic amyloidosis.

INTRODUCTION

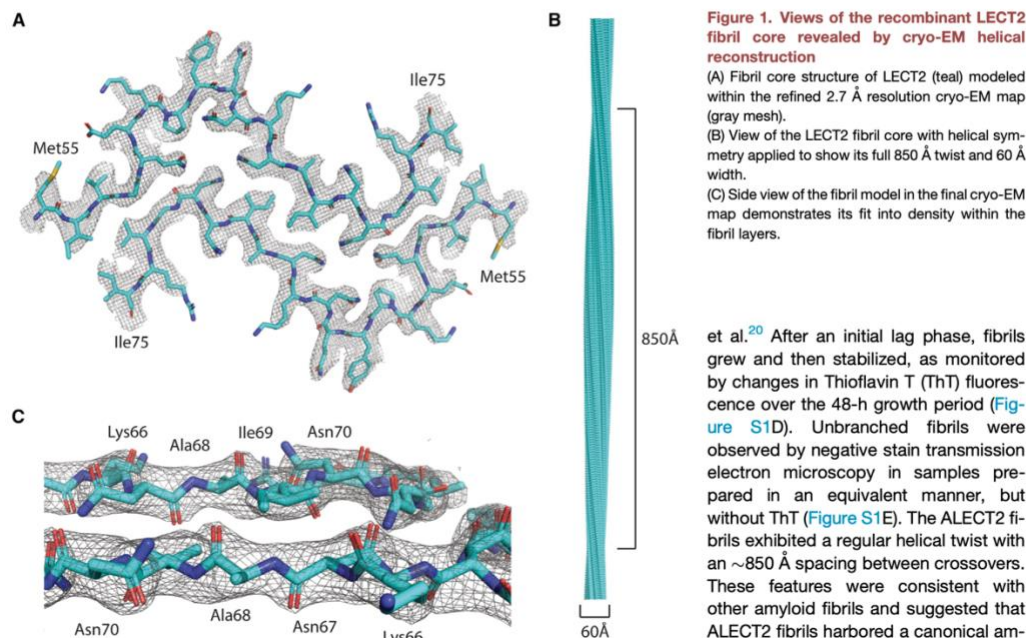
Amyloid diseases are linked to the formation and persistence of large, multimeric structures in various tissues. Amyloid fibrils are characterized by a cross-beta scaffold in which identical protein molecules mate tightly as beta-strands to form a long, unbranched fibril.^{1,2} Before amyloid-forming proteins assemble into fibrils, they can display a globular fold, requiring partial or total unfolding to convert into an amyloid state.^{3,4} Debilitating neurodegenerative disorders such as Alzheimer's and Parkinson's disease are well known to involve amyloidogenesis,^{5,6} while many other examples of amyloid diseases involve the systemic deposition of persistent amyloid fibrils throughout various organs including the heart, liver, kidneys, skin, digestive tract, and nervous system.⁷ The persistence of amyloid aggregates is a unifying feature of these diseases. However, the factors contributing to the misfolding, retention, and toxicity of each disease-associated fibril are distinct and likely dependent on their structures.⁸ This underscores the drive to characterize amyloid fibril structures.

In 2008, Benson et al. identified a new form of systemic amyloid disease associated with impaired renal function, renal failure, and nephrotic syndrome.⁹ Biochemical characterization of the fibrils observed in the glomeruli of the kidneys identified the protein leukocyte cell-derived chemotaxin-2 (LECT2) as a major component of the fibrils, inspiring the disease name, ALECT2 amyloidosis.⁹ Since its identification, ALECT2 amyloidosis has

been characterized as a common and likely underdiagnosed form of renal amyloidosis¹⁰ with high prevalence among patients of Mexican descent in the southwest United States, those of Punjabi descent, First Nations peoples in British Columbia, Egyptians, Chinese of Han ethnicity, and Native Americans.¹¹ There remains a need for diagnostic tools to distinguish ALECT2 amyloidosis from other renal amyloid diseases in patients and avoid misdiagnosis and improper treatment.¹² In contrast to some other systemic amyloid diseases, there are no available treatments for ALECT2 amyloidosis and no molecules that effectively target the amyloid state of its namesake.

Nascent LECT2 is a 151 amino acid polypeptide with an 18-residue N-terminal signal peptide that is cleaved before its secretion into the blood as a globular protein.¹³ In circulation, LECT2 performs multiple biological functions. It can act as a chemotactic factor to promote the migration of neutrophils to sites of infection.^{13,14} It can also act as a regulator of cartilage growth,¹⁵ and as a hepatokine important for metabolic homeostasis.¹⁶ Structurally, LECT2 displays an M23 metalloendopeptidase fold that coordinates Zn(II), but lacks a critical histidine residue that would enable its catalysis.¹⁷ While the clinical details of ALECT2 amyloid disease have been well characterized, the molecular details of ALECT2 amyloid aggregation remain unclear. Loss of this bound Zn(II) may play an important role in the conversion of properly folded LECT2 into an amyloid form *in vitro* as it is thought to destabilize a central beta-barrel motif within the





protein.¹⁸ Many patients who suffer from ALECT2 amyloidosis are also homozygous for the common I40V sequence polymorph of the protein.¹⁹ This mutation appears to be important, but not sufficient for inducing amyloid formation, and does not destabilize the metal-binding properties of the globular protein; accordingly, its role in amyloid conversion remains unclear.¹⁸ Amyloidogenic segments of LECT2 have been identified and assemble into amyloid-like fibrils in isolation.²⁰ However, the identity of the full-length ALECT2 amyloid core remains unclear.

Recent advances in single-particle cryo-EM using helical reconstruction have permitted the determination of near-atomic resolution structures of a wide variety of amyloid fibrils.³ These approaches have now allowed us to determine the structure of amyloid fibrils formed by recombinant full-length LECT2. This ~2.7 Å resolution cryo-EM structure of an ALECT2 fibril reveals a tightly mated amyloid core that spans residues 55 to 75 and harbors a network of polar ladders at its core. Its structure is also a first step toward a more comprehensive molecular understanding of ALECT2 amyloidosis and a starting point for the design of targeted therapeutics or diagnostic agents.

RESULTS

Assembly of recombinant human ALECT2 amyloid fibrils

Amyloid fibrils were grown from recombinantly expressed, full-length LECT2 protein encoding the I40V sequence polymorph. Purified LECT2 was allowed to assemble into fibrils over a 48-h period, while shaking at 2400 rpm at a temperature of 37°C in a modified version of the fibrillization buffer used by Tsiolaki

et al.²⁰ After an initial lag phase, fibrils grew and then stabilized, as monitored by changes in Thioflavin T (ThT) fluorescence over the 48-h growth period (Figure S1D). Unbranched fibrils were observed by negative stain transmission electron microscopy in samples prepared in an equivalent manner, but without ThT (Figure S1E). The ALECT2 fibrils exhibited a regular helical twist with an ~850 Å spacing between crossovers. These features were consistent with other amyloid fibrils and suggested that ALECT2 fibrils harbored a canonical amyloid core supported by steric zipper motifs. Indeed, certain regions of the LECT2

ALECT2 amyloid fibrils are urea and protease resistant

Amyloid aggregates are often highly stable and can resist chemical or proteolytic denaturation. To evaluate the stability of ALECT2 fibrils, a nephelometric assay was performed that monitored light scattering induced by fibrils in solution. Fibril dissolution was correlated with decreased nephelometry signal and assessed under conditions where fibrils were exposed to water, fibrillization buffer alone, fibrillization buffer with 3M urea, or 40 nM Proteinase K in 150 mM NaCl, 50 mM MOPS pH 6.5 buffer (Figure S4A). LECT2 fibrils persisted in solution when kept in fibrillization buffer or water over a 24-h period. In contrast, their exposure to 3 M urea induced an immediate, slight decrease in their abundance, followed by a stable lower signal over the 24-h period. Incubation with Proteinase K induced a larger drop in signal over the first 2 h of incubation following a stable, lower signal phase during the remaining incubation period. These initial drops in signal could have resulted from the dissolution of larger fibril clumps or the removal of more susceptible fibrils, but in all cases, there remained a stable subset of fibrils in solution (Figures S4B–S4E). The presence of a species that was resistant to protease digestion correlated with the appearance of a ~7 kDa band as observed by SDS-PAGE

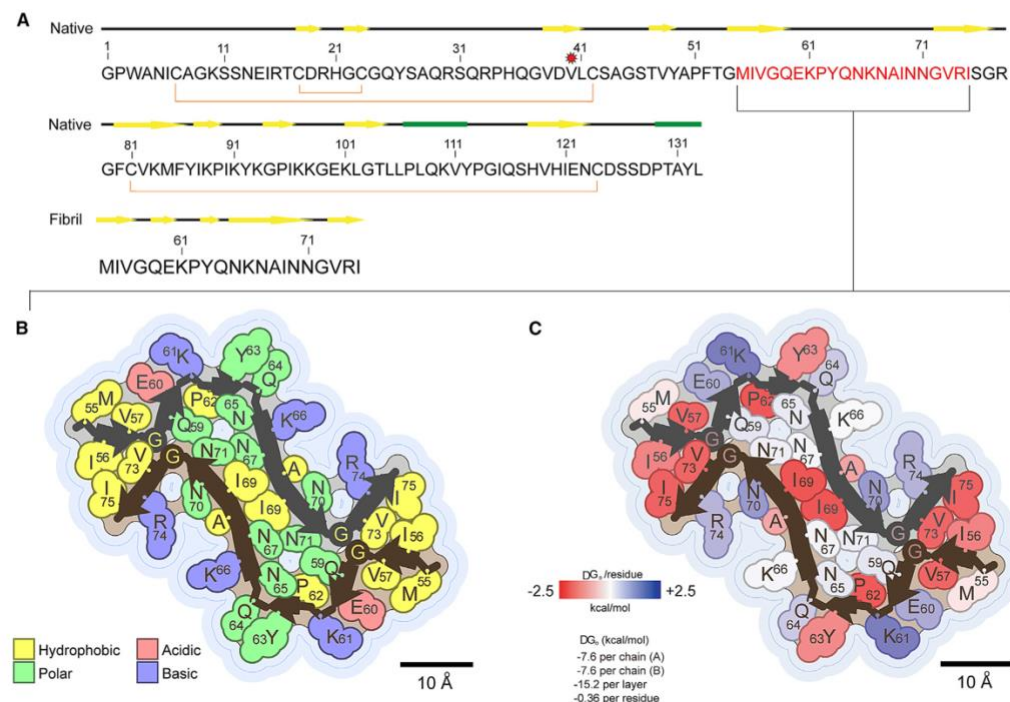


Figure 2. LECT2 fibril core sequence and structure overview

(A) Sequence of LECT2 showing native secondary structure elements as yellow arrows for beta-strands and green rectangles for alpha helices with disulfides shown as orange brackets and the I40V polymorphism noted by a red star. The sequence of the fibril core is highlighted in red. (B) Depiction of the fibril core showing the distribution of polar and nonpolar residues. (C) Depiction of the fibril core showing stabilizing residues and the calculated solvation energy of the fibril core structure. Depictions of the fibril core were generated using the Amyloid Illustrator web service (<https://srv.mbl.ucla.edu/AmyloidAtlas/Illustrator/>). Scale bars, 10 Å.

(Figure S9A). The identity of this fragment was investigated 60 min after exposing fibrils to Proteinase K, by bottom-up mass spectrometry. That analysis identified segments covering 25% of the LECT2 sequence (Table S1), including a fragment of the ALECT2 fibril core spanning residues N67-R74 and portions of its lysine-rich C-terminal region (Figure S9A).

Structural determination of an ALECT2 amyloid fibril core

After biochemical characterization, ALECT2 fibrils were prepared for high-resolution single-particle cryo-EM. The conditions for grid preparation were optimized to retain regularly twisting fibrils and discourage their adsorption to the air-water interface (Figures S4A and S5). Cryo-EM of these fibrils yielded 2D class averages that displayed fibril characteristics consistent with negative stain images (Figure S3A). Frozen hydrated fibrils exhibited one of two consistent helical crossovers. Fibrils with a 400 Å helical crossover, which we termed the fast-twisting polymorph, accounted for only about 10% of fibrils present in the sample (Figure S3C). The remaining 90% of fibrils exhibited an

850 Å helical crossover and were termed the slow-twisting polymorph (Figure S3D). The two polymorphs share very similar features apart from their different crossover distances. However, given its high abundance, only the slow-twisting polymorph proved amenable to high-resolution 2D classification and 3D structure determination. Subdivision of the fibril into two symmetry-related protofilaments could be inferred by the appearance of mirror symmetry in 2D class averages. Fibril images gave the impression of being mirrored across the fibril axis when it aligned parallel to the image plane. In this view, very near the fibril axis, the density of one protofilament appeared staggered relative to the other, resembling the teeth of a zipper (Figures S3E and S3F). These observations indicated that the two protofilaments might be related by a symmetry that approximated a 2₁-screw axis. X-ray fibril diffraction from the ALECT2 fibrils showed a strong 4.69 Å reflection, corresponding to the true spacing between ALECT2 layers along a protofilament (Figures S2A and S2B). This spacing was confirmed by analysis of Fourier transforms from high-resolution images of the fibrils (Figure S3B). The 2D classes were sufficient to generate a *de novo* 3D initial model in

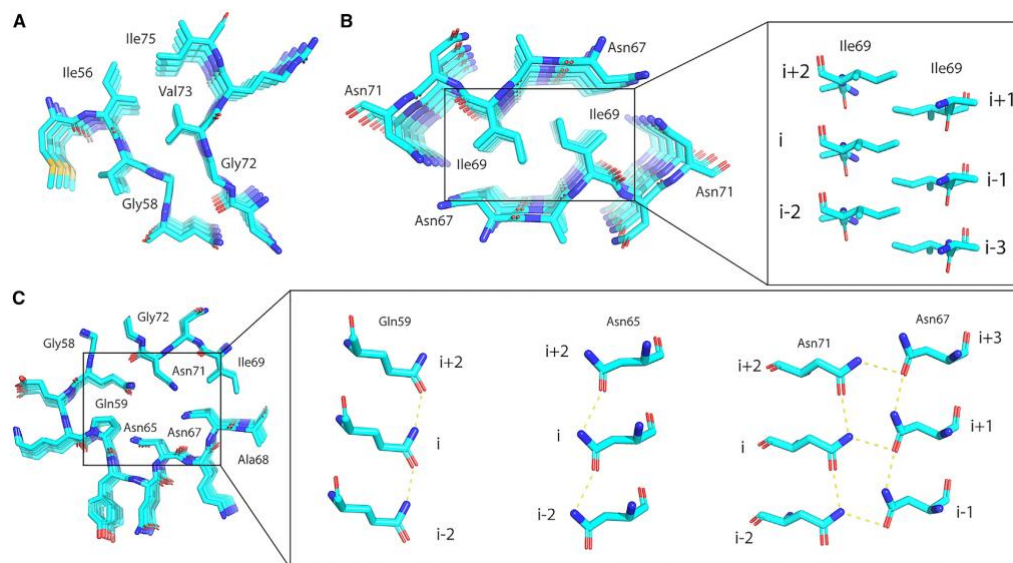


Figure 3. Stabilizing contacts at the core of the LECT2 fibril

(A) The outer steric zipper interface formed by residues Gly58, Gly72, Val73, Ile56, and Ile75.

(B) The central steric zipper interface formed by residue Ile69 is flanked by a paired polar ladder. Inset shows the pseudo- 2_1 stacking of fibril layers.

(C) Hydrophilic pocket within the LECT2 fibril core contains polar ladders formed by uncharged polar residues Gln59, Asn65, and Asn67 with Asn71, spaced 4.69 Å apart along the fibril axis (inset).

RELION and this map was used as an initial model for 3D classification (Figures S3G and S3H). Atomic force microscopy confirmed both the 850 Å helical crossover and the left-handed twist of the ALECT2 fibrils used for 3D reconstruction (Figure S11).

A three-dimensional reconstruction from the ALECT2 fibril images displayed an ordered amyloid core with an estimated resolution of ~ 2.7 Å, based on the 0.143 Fourier shell correlation criterion (Figure S6B). The map was of sufficient quality to allow for building of an unambiguous atomic model. There were no breaks in connectivity of the tube of density attributed to each molecule and there was clear separation between layers of density corresponding to individual LECT2 molecules, whether stacked along the fibril axis or mating with the opposite protofilament (Figures 1A and 1B). The resolution of the map allowed for the unambiguous assignment of side chains and peptide backbone oxygen atoms, allowing *de novo* assignment of the residues at its core (Figure 1C).^{21,22} The modeled portion of the amyloid core exhibited a clear pseudo- 2_1 -screw symmetry with the repeating unit being a 21-residue sequence stretching from methionine 55 to isoleucine 75 (Figures 2A and 2B). To assure that the assignment of this sequence was correct, Rosetta was used to calculate configurations for all possible 21-residue segments of LECT2 constrained to match the observed density. This was achieved by threading each sequence onto the peptide backbone of the model to test which threading segment produces the most stable structure. The analysis confirmed that the M55-I75 sequence was the most energetically favorable sequence assignment for the ge-

ometry of the fibril core (Figure S7A). Cryo-EM data collection parameters and fibril core structural information are summarized in Table 1.

Stability-promoting molecular features in the ALECT2 fibril core

At the core of the ALECT2 fibril, a segment from each of two protofilaments mate in a largely dry interface. The two protofilaments are nearly 2-fold symmetric in the fibril core, allowing hydrophobic zippers to coalesce a network of uncharged polar residues stabilized by polar ladders. Each hydrophobic pocket within the fibril core contains tightly packed steric zipper structures composed of residues I56, V73, and I75; residue I69 anchors the very center of the fibril core near the helical axis (Figures 3A and 3B). A pair of glycine residues, G58 and G72, spaced only 4 Å apart separates the hydrophobic patches at the edges of the fibril core from the network of polar residues within it (Figure 3A). These polar residues include Q59, N65, N67, and N71, all of which are incorporated into stabilizing polar ladders (Figure 3C). Residues N67 and N71 form a polar clasp that holds together not only layers above and below but also stitches together the molecules from opposite protofilaments (Figure 3C). This polar clasp also seals off the central pair of I69 residues and pins together the two largest beta-strands in the fibril structure. That combination of features results in an overall energetically favorable structure with a calculated standard free energy of solvation^{2,23} of ΔG_o of -15.2 kcal/mol per

Structure 31, 1386–1393, November 2, 2023



Table 1. Cryo-EM data collection and structure information

Data Collection	
Magnification	110,000x
Defocus range (μm)	−0.8 to −1.8
Voltage (kV)	300
Microscope	Titan Krios G3i
Camera	BioQuantum K3 camera
Frame exposure time (s)	2.5
No. of movie frames	40
Total electron dose (e [−] /Å ²)	52
Pixel size (Å)	0.799
Reconstruction	
Box size (pixel)	384
Inter-box distance (Å)	30
No. of segments extracted	1,622,942
No. of segments after Class2D	308,578
No. of segments after Class3D	24,770
Map resolution (Å)	2.715
FSC threshold	0.143
Map sharpening B-factor (Å ²)	−69.7
Helical rise (Å)	2.345
Helical twist (°)	179.49
Symmetry imposed	C1
Atomic Model	
Initial model used	<i>De novo</i>
No. of unique non-hydrogen atoms	166
R.m.s.d. bonds (Å)	0.006
R.m.s.d. angles (°)	1.055
Molprobt score	1.88
Favored rotamers (%)	94.74
Ramachandran favored (%)	5.26
Ramachandran outliers (%)	0
CB deviations >0.25 Å (%)	0
Bad bonds (%)	0
Bad angles (%)	0

layer and −0.36 kcal/mol per residue (Figure 2C). The solvation energy and small size of the ALECT2 amyloid core are similar to that of the human RIPK1-RIPK3 hetero-amyloid, which is hypothesized to form stable aggregates *in vivo* and signal tumor necrosis factor-induced necroptosis in cells.²⁴ These metrics could be further influenced by other stabilizing interactions, such as disulfide bridges or metal coordination sites, that in the present structure are absent from the fibril core but should exist within the fibril fuzzy coat. That coat remains unresolved in the current fibril structure.

DISCUSSION

In search of amyloid folds adopted by the LECT2 protein, we find that recombinant LECT2 polypeptides encoding the I40V sequence polymorphism readily form amyloid fibrils that display a canonical beta-sheet-rich architecture with stacked layers of

the LECT2 protein separated by a rise of 4.69 Å per layer and obeying a regular, left-handed, helical twist. The amyloid core resolved within these fibrils is composed of two identical protofilaments that each contain 21 of the 133 residues in full-length LECT2: from methionine 55 to isoleucine 75. Interestingly, this sequence has not been predicted by previous work to form amyloid fibrils. However, other studies identified the regions immediately preceding and following it as possible amyloidogenic segments,²⁰ and the newly identified core segment is predicted to form steric zipper structures by the amyloid-prediction algorithm in ZipperDB²⁵ (Figure S1A).

While a significant fraction of the LECT2 sequence is predicted to be amyloidogenic, much of the protein's globular fold is stabilized by three disulfide bonds, two in the N-terminal region, and one in C-terminal region (Figure 2A). These bonds would likely prevent many of the predicted amyloidogenic regions from assembling into fibrils. However, the core segment identified here is not involved in any of the native disulfide bonds and represents a portion of the central beta-barrel motif of LECT2.¹⁷ Specifically, the ALECT2 fibril core encompasses most of beta-strands 3 and 4 and all of loop 2 in the LECT2 globular fold (Figure S8A). That loop is solvent-facing and sits adjacent to the Zn(II)-binding pocket.¹⁷ This same region of the protein has previously been hypothesized to be susceptible to reductions in pH that could facilitate the loss of the bound zinc coordinated to the first and second beta-strands and may destabilize the beta-barrel structure, exposing amyloidogenic segments.¹⁸

A comparison of native and fibrillar ALECT2 structures reveals that the beta-strand conformation of N- and C-terminal segments of the fibrillar fold are retained from beta-strands 3 and 4 of the globular fold. However, the central segment of the fibril core represents a large ordering of secondary structure, namely rearrangement of loop 2 into three beta-strand segments, the largest of which stretches from N65 to N71 (Figure 2B). The outer surface of the ALECT2 fibril is decorated with larger amino acids including basic residues K61, K66, and R74 as well as one acidic residue, E60, and a pi-stacked Y63 (Figure 2B). These residues are also in solvent-exposed positions in the globular fold of LECT2. Residues I56, V73, and I75, which form the outer steric zipper segments of the fibril, are all buried within the central beta-barrel of globular LECT2, and I69, the central hydrophobic residue, is natively buried as part of loop 2 abutted to alpha-helix 1 (Figure S8B). In contrast, the polar-ladder-forming residues Q59, N65, N67, and N71 undergo a large conformational and environmental change from the globular fold. All four of these residues are solvent-facing in loop 2 of the globular structure of LECT2, and all are packed into a relatively dry interface in the fibril structure (Figure S8B). Polar ladder conformations like these revealed in this structure have long been hypothesized to stabilize amyloid assemblies^{26,27} and were also favored during energy minimization of the ALECT2 fibril fold in Rosetta (Figure S7B). The structural transition to a polar ladder within a dry environment confers increased stability by joining chains along the fibril axis while fully satisfying hydrogen bonds within the fibril core. This stabilizing effect is further amplified when the two polar ladders in separate protofilaments interdigitate into a zipper conformation, as is observed between residues N67 and N71 in the ALECT2 fibril.

The structural conversion of the loop 2 region of globular LECT2 plays a central role in generating the ALECT2 amyloid

core observed here, and the dimeric nature of the fibril draws some comparisons to its globular counterpart. In addition to the dimer resolved in the crystal structure of LECT2, previous studies have observed that stable, SDS-resistant, dimers and higher-order oligomers of LECT2 readily form with or without the presence of zinc.²⁸ One hypothesis for this dimerization is that LECT2 could undergo domain swapping. This arrangement has been observed for amyloid proteins previously, such as with β_2 -microglobulin, which has been shown to form domain-swapped amyloid fibrils with disulfide linkages.²⁹ Interestingly, the amyloid core arrangement of loop 2 spans the gap between monomers of LECT2 in the crystallographic dimer well. A model of a domain-swapped LECT2 dimer through the loop 2 region using the amyloid core structure resolved here shows that this region could facilitate the domain-swapped conformation with minimal disruption to the rest of the globular LECT2 fold (Figure S10A). Further, this conformation maintains the general amyloid core arrangement and could adopt the fibrillar fold following destabilization of the rest of the dimeric interface (Figure S10B).

It remains to be seen whether recombinant LECT2 amyloid fibrils match those present in the kidneys of ALECT2 systemic amyloidosis patients; however, structures of the former and the segments at their core provide a first glimpse at stabilizing interactions within ALECT2 fibrils. Further, the exact role of the I40V polymorphism in amyloid aggregation remains unclear as the residue was not resolved as part of the ALECT2 amyloid core. The segment at the core of this recombinant ALECT2 fibril is derived from part of the central beta-barrel structure of globular LECT2, supporting the theory that loss of zinc may destabilize this portion of the protein, exposing the amyloidogenic segments resolved here and allowing for aggregation. This information opens the possibility for rational design of aggregation inhibitors which could either stabilize this region of globular LECT2 or inhibit the extension of fibrils of ALECT2 by targeting residues at its core.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**
 - Microbe strains
- **METHOD DETAILS**
 - Protein expression and purification
 - Thioflavin T fluorescence
 - Cryo-EM sample preparation and data collection
 - Cryo-EM data processing
 - Model building
 - X-ray fibril diffraction
 - Fibril stability assays

- In-gel digestion and peptide mass fingerprinting of ALECT2 using GeLC-MS/MS
- Atomic force microscopy of ALECT2 fibrils
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.str.2023.08.007>.

ACKNOWLEDGMENTS

This work is supported by DOE Grant DE-FC02-02ER63421, NSF Grant DMR-1548924, the NIH-NIGMS Grant R35 GM128867. L.S.R. is supported by the USPHS National Research Service Award 5T32GM008496. J.A.R. is also supported as a Sloan Fellow, a Pew Scholar, and a Packard Fellow. Part of this work was performed at the Stanford-SLAC Cryo-EM Center (S2C2). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the following S2C2 personnel for their invaluable support and assistance: Dr. Patrick Mitchell, Dr. Yee-Ting Lee, and Professor Wah Chiu who are supported by the National Institutes of Health Common Fund Transformative High-Resolution Cryo-Electron Microscopy program (U24 GM129541). In addition, we thank the staff of the UCLA Molecular Instrumentation Center (MIC), supported by the UCLA Division of Physical Sciences, and the staff of the UCLA CNSI Nano & Pico Characterization Laboratory.

AUTHOR CONTRIBUTIONS

L.S.R. generated protein and performed all biochemical analyses, biophysical characterization, and structure determination by cryo-EM. M.D.F. assisted in cryo-EM sample screening and processing in RELION. S.Z. performed computational structure modelling in Rosetta. N.A.S. assisted in characterization of the fibrils by mass spectrometry. M.R.S. assisted in model building, performing X-ray fibril diffraction, and structural analysis. J.A.R. oversaw and provided input on all aspects of the work. All authors helped to write and provided critical feedback on the article.

DECLARATION OF INTERESTS

J.A.R. is an equity stakeholder of MedStruc Inc.

Received: March 26, 2023

Revised: June 29, 2023

Accepted: August 7, 2023

Published: August 31, 2023

REFERENCES

1. Eisenberg, D., and Jucker, M. (2012). The Amyloid State of Proteins in Human Diseases. *Cell* 148, 1188–1203. <https://doi.org/10.1016/j.cell.2012.02.022>.
2. Sawaya, M.R., Hughes, M.P., Rodriguez, J.A., Riek, R., and Eisenberg, D.S. (2021). The expanding amyloid family: Structure, stability, function, and pathogenesis. *Cell* 184, 4857–4873. <https://doi.org/10.1016/j.cell.2021.08.013>.
3. Fitzpatrick, A.W.P., Falcon, B., He, S., Murzin, A.G., Murshudov, G., Garringer, H.J., Crowther, R.A., Ghetti, B., Goedert, M., and Scheres, S.H.W. (2017). Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature* 547, 185–190. <https://doi.org/10.1038/nature23002>.
4. Schmidt, M., Wiese, S., Adak, V., Engler, J., Agarwal, S., Fritz, G., Westermark, P., Zacharias, M., and Fändrich, M. (2019). Cryo-EM structure of a transthyretin-derived amyloid fibril from a patient with hereditary ATTR amyloidosis. *Nat. Commun.* 10, 5008. <https://doi.org/10.1038/s41467-019-13038-z>.
5. Selkoe, D.J. (1991). The molecular pathology of Alzheimer's disease. *Neuron* 6, 487–498. [https://doi.org/10.1016/0896-6273\(91\)90052-2](https://doi.org/10.1016/0896-6273(91)90052-2).

Structure 31, 1386–1393, November 2, 2023

6. Spillantini, M.G., Crowther, R.A., Jakes, R., Hasegawa, M., and Goedert, M. (1998). α -Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. *Proc. Natl. Acad. Sci. USA* 95, 6469–6473. <https://doi.org/10.1073/pnas.95.11.6469>.
7. Muchtart, E., Dispenzieri, A., Magen, H., Grogan, M., Mauermann, M., McPhail, E.D., Kurtin, P.J., Leung, N., Buadi, F.K., Dingli, D., et al. (2021). Systemic amyloidosis from A (AA) to T (ATTR): a review. *J. Intern. Med.* 289, 268–292. <https://doi.org/10.1111/joim.13169>.
8. Shi, Y., Zhang, W., Yang, Y., Murzin, A.G., Falcon, B., Kotecha, A., van Beers, M., Tarutani, A., Kametani, F., Garringer, H.J., et al. (2021). Structure-based classification of tauopathies. *Nature* 598, 359–363. <https://doi.org/10.1038/s41586-021-03911-7>.
9. Benson, M.D., James, S., Scott, K., Liepnieks, J.J., and Klueve-Beckerman, B. (2008). Leukocyte chemotactic factor 2: A novel renal amyloid protein. *Kidney Int.* 74, 218–222. <https://doi.org/10.1038/ki.2008.152>.
10. Larsen, C.P., Ismail, W., Kurtin, P.J., Vrana, J.A., Dasari, S., and Nasr, S.H. (2016). Leukocyte chemotactic factor 2 amyloidosis (ALCT2) is a common form of renal amyloidosis among Egyptians. *Mod. Pathol.* 29, 416–420. <https://doi.org/10.1038/modpathol.2016.29>.
11. Picken, M.M. (2020). The Pathology of Amyloidosis in Classification: A Review. *AHA* 143, 322–334. <https://doi.org/10.1159/000506696>.
12. Mann, B.K., Bhandohal, J.S., Cobos, E., Chitturi, C., and Eppanapally, S. (2022). LECT-2 amyloidosis: what do we know? *J. Invest. Med.* 70, 348–353. <https://doi.org/10.1136/jim-2021-002149>.
13. Yamagoe, S., Yamakawa, Y., Matsuo, Y., Minowada, J., Mizuno, S., and Suzuki, K. (1996). Purification and primary amino acid sequence of a novel neutrophil chemotactic factor LECT2. *Immunol. Lett.* 52, 9–13. [https://doi.org/10.1016/0165-2478\(96\)02572-2](https://doi.org/10.1016/0165-2478(96)02572-2).
14. Yamagoe, S., Kameoka, Y., Hashimoto, K., Mizuno, S., and Suzuki, K. (1998). Molecular Cloning, Structural Characterization, and Chromosomal Mapping of the Human LECT2 Gene. *Genomics* 48, 324–329. <https://doi.org/10.1006/geno.1997.5198>.
15. Hiraki, Y., Inoue, H., Kondo, J., Kamizono, A., Yoshitake, Y., Shukunami, C., and Suzuki, F. (1996). A Novel Growth-Promoting Factor Derived from Fetal Bovine Cartilage, Chondromodulin II: PURIFICATION AND AMINO ACID SEQUENCE. *J. Biol. Chem.* 271, 22657–22662. <https://doi.org/10.1074/jbc.271.37.22657>.
16. Willis, S.A., Sargeant, J.A., Yates, T., Takamura, T., Takayama, H., Gupta, V., Brittain, E., Crawford, J., Parry, S.A., Thackray, A.E., et al. (2020). Acute Hyperenergetic, High-Fat Feeding Increases Circulating FGF21, LECT2, and Fetuin-A in Healthy Men. *J. Nutr.* 150, 1076–1085. <https://doi.org/10.1093/jn/nxz333>.
17. Zheng, H., Miyakawa, T., Sawano, Y., Asano, A., Okumura, A., Yamagoe, S., and Tanokura, M. (2016). Crystal Structure of Human Leukocyte Cell-Derived Chemotaxin 2 (LECT2) Reveals a Mechanistic Basis of Functional Evolution in a Mammalian Protein with an M23 Metalloendopeptidase Fold. *J. Biol. Chem.* 291, 17133–17142. <https://doi.org/10.1074/jbc.M116.720375>.
18. Ha, J.-H., Tu, H.-C., Wilkens, S., and Loh, S.N. (2021). Loss of bound zinc facilitates amyloid fibril formation of leukocyte-cell-derived chemotaxin 2 (LECT2). *J. Biol. Chem.* 296, 100446. <https://doi.org/10.1016/j.jbc.2021.100446>.
19. Larsen, C.P., Kossmann, R.J., Beggs, M.L., Solomon, A., and Walker, P.D. (2014). Clinical, morphologic, and genetic features of renal leukocyte chemotactic factor 2 amyloidosis. *Kidney Int.* 86, 378–382. <https://doi.org/10.1038/ki.2014.11>.
20. Tsiolaki, P.L., Nasi, G.I., Baltoumas, F.A., Fishman, S., Tu, H.-C., and Iconomidou, V.A. (2019). Delving into the amyloidogenic core of human leukocyte chemotactic factor 2. *J. Struct. Biol.* 207, 260–269. <https://doi.org/10.1016/j.jsb.2019.06.001>.
21. Falcon, B., Zivanov, J., Zhang, W., Murzin, A.G., Garringer, H.J., Vidal, R., Crowther, R.A., Newell, K.L., Ghetti, B., Goedert, M., and Scheres, S.H.W. (2019). Novel tau filament fold in chronic traumatic encephalopathy encloses hydrophobic molecules. *Nature* 568, 420–423. <https://doi.org/10.1038/s41586-019-1026-5>.
22. Scheres, S.H.W. (2020). Amyloid structure determination in RELION-3.1. *Acta Crystallogr. D Struct. Biol.* 76, 94–101. <https://doi.org/10.1107/S2059798319016577>.
23. Eisenberg, D., and McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature* 319, 199–203. <https://doi.org/10.1038/319199a0>.
24. Mompeán, M., Li, W., Li, J., Laage, S., Siemer, A.B., Bozkurt, G., Wu, H., and McDermott, A.E. (2018). The Structure of the Necrosome RIPK1-RIPK3 Core, a Human Hetero-Amyloid Signaling Complex. *Cell* 173, 1244–1253.e10. <https://doi.org/10.1016/j.cell.2018.03.032>.
25. Goldschmidt, L., Teng, P.K., Riek, R., and Eisenberg, D. (2010). Identifying the amyloids, proteins capable of forming amyloid-like fibrils. *Proc. Natl. Acad. Sci. USA* 107, 3487–3492. <https://doi.org/10.1073/pnas.0915166107>.
26. Perutz, M.F., Johnson, T., Suzuki, M., and Finch, J.T. (1994). Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci. USA* 91, 5355–5358. <https://doi.org/10.1073/pnas.91.12.5355>.
27. Nelson, R., Sawaya, M.R., Balbirnie, M., Madsen, A.O., Riekel, C., Grothe, R., and Eisenberg, D. (2005). Structure of the cross- β spine of amyloid-like fibrils. *Nature* 435, 773–778. <https://doi.org/10.1038/nature03680>.
28. Okumura, A., Suzuki, T., Miyatake, H., Okabe, T., Hashimoto, Y., Miyakawa, T., Zheng, H., Unoki-Kubota, H., Ohno, H., Dohmae, N., et al. (2013). Leukocyte cell-derived chemotaxin 2 is a zinc-binding protein. *FEBS Lett.* 587, 404–409. <https://doi.org/10.1016/j.febslet.2013.01.025>.
29. Liu, C., Sawaya, M.R., and Eisenberg, D. (2011). β 2-microglobulin forms three-dimensional domain-swapped amyloid fibrils with disulfide linkages. *Nat. Struct. Mol. Biol.* 18, 49–55. <https://doi.org/10.1038/nsmb.1948>.
30. Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. *Acta Crystallogr. D* 66, 486–501. <https://doi.org/10.1107/S0907444910007493>.
31. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* 30, 70–82. <https://doi.org/10.1002/pro.3943>.
32. Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apfelbaum, A., Hagel, P., Sitsel, O., Ralsch, T., Prumbaum, D., et al. (2019). SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun. Biol.* 2, 218–313. <https://doi.org/10.1038/s42003-019-0437-z>.
33. Scheres, S.H.W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* 180, 519–530. <https://doi.org/10.1016/j.jsb.2012.09.006>.
34. Liebschner, D., Afonine, P.V., Baker, M.L., Bunkóczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A.J., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* 75, 861–877. <https://doi.org/10.1107/S2059798319011471>.
35. Chaudhury, S., Lyskov, S., and Gray, J.J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* 26, 689–691. <https://doi.org/10.1093/bioinformatics/btq007>.
36. Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567. [https://doi.org/10.1002/\(SICI\)1522-2683\(19991201\)20:18<3551::AID-ELPS3551>3.0.CO;2-2](https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2).
37. Wagner, T., Lusnig, L., Pospich, S., Stabrin, M., Schönfeld, F., and Raunser, S. (2020). Two particle-picking procedures for filamentous proteins: SPHIRE-crYOLO filament mode and SPHIRE-STRIPER.



- Acta Crystallogr. D 76, 613–620. <https://doi.org/10.1107/S2059798320007342>.
38. He, S., and Scheres, S.H.W. (2017). Helical reconstruction in RELION. *J. Struct. Biol.* 198, 163–176. <https://doi.org/10.1016/j.jsb.2017.02.003>.
 39. Zivanov, J., Nakane, T., and Scheres, S.H.W. (2019). A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCrJ* 6, 5–17. <https://doi.org/10.1107/S205225251801463X>.
 40. Zivanov, J., Nakane, T., and Scheres, S.H.W. (2020). Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1. *IUCrJ* 7, 253–267. <https://doi.org/10.1107/S2052252520000081>.
 41. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theor. Comput.* 13, 3031–3048. <https://doi.org/10.1021/acs.jctc.7b00125>.



STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
BL21 Star™ (DE3) Chemically Competent <i>E. coli</i>	ThermoFisher Scientific	CAS: C601003
Chemicals, peptides, and recombinant proteins		
LECT2 protein	This work.	N/A
Proteinase K protein	Sigma-Aldrich	CAS: 39450-01-6
Thioflavin T	ThermoFisher Scientific	CAS: 2390-54-7
Deposited data		
Recombinant LECT2 Fibril	This work.	PDB ID: 8G2V; EMBD: EMD-29682
Oligonucleotides		
T7 promoter, forward primer (TAATACGACTCACTATAGGG)	Genewiz	N/A
T7 terminator, reverse primer (GCTAGTTATTGCTCAGCGG)	Genewiz	N/A
Recombinant DNA		
Codon-optimized synthetic gene encoding residues 19–151 of the human LECT2 protein (GenBank AAC17734.1)	Genewiz	N/A
Software and algorithms		
Pymol Molecular Graphics	System Schrodinger, LLC	https://pymol.org/2/
Coot	Emsley et al. ³⁰	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/
UCSF Chimera	UCSF RBV ³¹	https://www.cgl.ucsf.edu/chimera/
CrYOLO	Wagner et al. ³²	https://cryolo.readthedocs.io/en/stable/index.html
RELION	Sjors H.W. Scheres ³³	https://relion.readthedocs.io/en/release-4.0/
Phenix	Liebschner et al. ³⁴	https://phenix-online.org/
PyRosetta	Chaudhury et al. ³⁵	https://www.pyrosetta.org/
MASCOT	Perkins et al. ³⁶	https://proteomicsresource.washington.edu/mascot/
Other		
Quantifoil (R1.2/1.3) 300M Au Cryo-EM Grids	Ted Pella inc.	CAS: 658-300-AU-100

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jose A. Rodriguez (jrodriguez@mbi.ucla.edu), upon request.

Materials availability

Plasmids generated in this study are available from the [lead contact](#) upon reasonable request. This study did not generate other new or unique reagents.

Data and code availability

All data and code are available in either the main text or the Supplementary Material. The reconstructed cryo-EM map was deposited in the Electron Microscopy Data Bank with the accession code EMD-29682, while coordinates for the refined atomic model were deposited in the Protein Data Bank under the accession code PDB: 8G2V. Code generated for PyRosetta sequence threading has been uploaded with this work in the file [Data S1](#). Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Microbe strains

Protein expression was done using BL21 Star™ (DE3) Chemically Competent E. coli cells from ThermoFisher Scientific (CAS: C601003).

METHOD DETAILS

Protein expression and purification

Full-length, mature human sequence LECT2 protein encoding the I40V sequence polymorphism and an N-terminal 6-His tag was transformed into and expressed in BL21-gold E. coli cells. Two 1L cultures grown at 37°C to an O.D. of 0.75 were induced with 1mM IPTG and allowed to shake overnight at 23°C. Cells were then collected by centrifugation at 10,000xg for 10 minutes, resuspended in 1X PBS with 6M guanidine, and lysed on an Avestin Emulsiflex C3 (ATA Scientific). Insoluble cell material was removed by centrifugation at 9000 rpm for 50 minutes and the supernatant was collected, filtered using a 0.8µm size cutoff, and run over a His-trap column and washed with 3 column volumes of buffer. Protein was eluted with an elution buffer of 6M guanidine, 1X PBS, and 500 mM imidazole. Eluate fractions were evaluated for protein purity by gel electrophoresis then pooled and concentrated to 40 mg/ml (Figure S1B). Concentrated stocks were flash frozen for storage at -80°C in 1X PBS with 6M guanidine and 20mM TCEP.

Thioflavin T fluorescence

To generate a Thioflavin T (ThT) fibril growth kinetics curve, LECT2 protein was prepared to 2 mg/ml in fibrillization buffer as described above. Equimolar amounts of Thioflavin T, also dissolved in fibrillization buffer, was added to this solution. This preparation along with a control of only fibrillization buffer and ThT were set up in triplicates of 100 µL in a 96 well plate. The samples were then subjected to continuous shaking on a Varioskan LUX (ThermoFisher Scientific) as the ThT fluorescence was measured every 15 minutes for 48 hours using an excitation wavelength of 445 nm and measuring emission at 482 nm.

Cryo-EM sample preparation and data collection

To grow fibrils for cryo-EM analysis, stock LECT2 protein was diluted to 2 mg/ml (120 µM) into a fibrillization buffer containing 50 mM MOPS 6.5, 30 mM TCEP, 10 mM EDTA, and 150 mM NaCl. A volume of 100 µL of protein solution per well was then agitated in a 96 well plate on an acoustic shaker running at 60 Hz (2400 rpm) at 37°C for 48 hours. The solution was collected and spun down at 10,000xg for 5 minutes to remove large aggregates not amenable to single particle analysis. The solution was then diluted 6-fold in buffer and glycerol was added to the remaining supernatant to a concentration of 0.25%. A volume of 1.5 µL of this solution was added to each side of a gold Quantifoil R 1.2/1.3 cryo-EM grid (Ted Pella Inc.) within a Vitrobot System (ThermoFisher Scientific) with humidity set to 100% and a temperature of 4°C. The grid was then blotted for 1.5 seconds with the blot force set to -1, plunge frozen in liquid ethane, and stored for data collection in liquid nitrogen. High-resolution cryo-EM data was collected at the Stanford-SLAC Cryo-EM Center (S²C²) over the course of 48 hours using EPU for automated data collection. A total of 13830 movies were obtained taking three shots per hole on the grid. Movies were collected on a Titan Krios G3i microscope (ThermoFisher Scientific) equipped with a BioQuantum K3 camera (Gatan, Inc.) using a pixel size of 0.79 Å/pixel, a total dose of 52 e-/Å² over 40 frames, and a defocus range between -0.8 and -1.8 µm.

Cryo-EM data processing

Collected images were input into MotionCor2 for drift correction and CTFFIND4 was used to calculate Contrast Transfer Functions (CTF). Fibrils were automatically picked using filament mode in crYOLO^{32,37} after being trained on a manually selected pool of fibril segments picked from 100 movies. This yielded 1,622,942 segments using an overlap of 90% between neighboring segments and a box size of 384 pixels. The segments were then transferred to RELION 3.1^{33,38,39} for all subsequence 2D and 3D classification. The segments were split into six equally sized groups and subjected to iterative rounds of reference-free two-dimensional (2D) classification using T = 8 and K = 100 to remove poorly aligned classes. After narrowing each split of segments, they were combined again (T = 8 and K = 100) resulting in 308,578 segments contributing to well-defined 2D classes to be used in 3D classification. These classes were used to generate a *de novo* 3D initial model which was used as a reference for further 3D classification. Initial classification (K = 4, T = 4) did not impose pseudo-2₁-screw symmetry and used an initial helical twist of -1.06° and a helical rise of 4.8 Å. The best class from this classification was selected out and further subclassified by removing any segments with a CTF estimate poorer than 4 Å resolution. The resulting 51,543 segments were used for further 3D classification (K = 3, T = 4) allowing local optimization of helical twist and rise. The best resulting class showed clear strand separation along the fibril axis and the pseudo-2₁-screw symmetry became apparent without it having been imposed. The model and data from these 24,770 segments were used for high-resolution gold-standard 3D auto-refinement. The resulting model underwent iterative Bayesian polishing³⁹ and CTF refinement⁴⁰ to further improve the map (Figure S6A). A final refined helical twist of 179.49° was derived from the auto-refine map and a helical rise of 2.345 Å was imposed on the post-processed map based on X-ray fibril diffraction data showing 4.69 Å strand separation in the fibril samples. The final resolution was ultimately calculated to be 2.715 Å from gold-standard Fourier shell correlations at 0.143 between two independently refined half-maps (Figure S6B).



Model building

A poly-alanine model was initially built into the post-processed map in Coot.³⁰ Side chains were modified manually to test the fit of different registrations of the LECT2 sequence in the density. A close approach between tubes of density could be interpreted only as the G58-G72 interaction between symmetric strands, as the closeness of the approach would sterically exclude all other side chains. This feature helped to identify the likely segment of ALECT2 and the entire 21 residue chain from Met55-Ile75 was built out from there. Final atomic refinements and statistical calculations for Table 1 were performed in Phenix.³⁴ While the map seemed to agree clearly with this assignment, we decided to confirm its validity by testing the fit and favorability of all possible 21-residue segments from the LECT2 sequence within the map using Rosetta. Threading was performed with a custom python script utilizing the PyRosetta software package.³⁵ In a sliding window approach, each sequence window was threaded onto a poly-alanine backbone, placed in the density and Fast Relaxed in PyRosetta. The pipeline used the REF2015 score function⁴¹ with an increased electrostatics weight ($fa_elec = 1.5$) in combination with a density score term ($elec_dens_fast = 25$). Each sequence was energy minimized in triplicate and the resulting scores were averaged. Symmetry was applied to pose objects to increase speed of computation. The Fast Relaxed structure was then optimized and refined in Coot and Phenix.³⁴ The code generated for PyRosetta sequence threading has been uploaded with this work in the file [Data S1](#).

X-ray fibril diffraction

ALECT2 amyloid fibrils were grown as described above for cryo-EM sample preparation. A 3 μ L droplet of the mature fibril solution was pipetted between two glass rods held about a millimeter apart and allowed to evaporate. This process was repeated multiple times until a visible fibril bundle was observed bridging the glass rods. This bundle was placed on an in-house Rigaku FRE+ rotating anode X-ray beamline and exposed for 5 minutes onto a Rigaku HTC detector with varimax confocal optics. Proteinase K and ice ring diffraction patterns were used to calibrate the detector distance and accurately measure the 4.69 Å reflection observed from the fibril bundle.

Fibril stability assays

ALECT2 amyloid fibrils were grown as described above for cryo-EM sample preparation and then pelleted at 14,300xg for 10 minutes. The fibril pellets were then resuspending to a protein concentration of 40 μ M in water, fibrillization buffer, 3M urea with fibrillization buffer, and fibrillization buffer with 1:1000 Proteinase K (Sigma-Aldrich) to a final concentration of 40 nM. These solutions were immediately pipetted in triplicate into a 96 well tray alongside buffer control wells for each and placed on a NEPHELOstar Plus plate reader (BMG Labtech). The tray was stirred every ten minutes to homogenize the mixture and light scattering of particles in solution was measured immediately after each agitation. For analysis of protease resistant LECT2 fragments, fibrils at 120 μ M were incubated with 10 μ M Proteinase K for various timepoints and then evaluated by SDS-PAGE (Figure S9A). The protein gel band from the one-hour timepoint was extracted, digested with Trypsin, and analyzed by gel liquid chromatography tandem mass spectrometry.

In-gel digestion and peptide mass fingerprinting of ALECT2 using GeLC-MS/MS

Gel Liquid Chromatography tandem mass spectrometry spectra collected on Proteinase K digested ALECT2 fibrils were acquired on a ThermoFisher Q-Exactive Plus (UCLA Molecular Instrumentation Center, Los Angeles, CA, USA). LECT2 fibrils at 120 μ M were incubated with 10 μ M Proteinase K for one hour. These samples were removed at various timepoints, boiled for 10 minutes at 98°C to halt digestion, and loaded on a 4–12% Bis-Tris SDS-PAGE gel. The gel band from the one-hour timepoint was excised and digested with 200 ng trypsin at 37°C overnight. The digested products were then extracted from the gel bands in 50% acetonitrile/49.9% H₂O/0.1% trifluoroacetic acid (TFA) followed by desalting with C18 StageTips. Extracted peptides were then injected on an EASY-Spray HPLC column (25 cm $\times$ 75 μ m ID, PepMap RSLC C18, 2 μ m, ThermoScientific) and tandem mass spectra were acquired with a quadrupole orbitrap mass spectrometer (Q-Exactive Plus Thermo Fisher Scientific) interfaced to a nanoelectrospray ionization source. The raw MS/MS data were converted into MGF format by Thermo Proteome Discoverer (VER. 1.4, Thermo Scientific) and analyzed by a MASCOT³⁶ sequence database search.

Atomic force microscopy of ALECT2 fibrils

ALECT2 fibrils grown for 48 hours as described above were centrifuged for 5 minutes at 10,000 rpm and the supernatant was collected. A 15 μ L aliquot of the supernatant was deposited onto freshly cleaved mica, incubated for 10 min, rinsed with 200 μ L of water, and dried by nitrogen. Images were taken on a Bruker Dimension Icon microscope (UCLA CNSI Nano & Pico Characterization Laboratory) in PeakForce Tapping mode with a ScanAsyst-Air-HPI probe. Images were collected at scan angles of 0 and 90 degrees to rule out directionality bias in viewing twist handedness. Images were flattened and analyzed using NanoScope Analysis 2.0 software (Bruker).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical calculations for cryo-EM reconstruction and structural analysis were done in RELION and Phenix and are reported in Table 1. Means and standard deviations for Thioflavin T and Nephelometric assays were computed in Microsoft Excel. Statistical details can be found in the relevant figure legends.

Structure, Volume 31

Supplemental Information

Cryo-EM structure of a human LECT2 amyloid

fibril reveals a network

of polar ladders at its core

Logan S. Richards, Maria D. Flores, Samantha Zink, Natalie A. Schibrowsky, Michael R. Sawaya, and Jose A. Rodriguez

Supplementary Information:

Figure S1. Production of recombinant LECT2 amyloid fibrils. Related to Star Methods. A) Predicted amyloid forming six-residue segments (kcal/mol < -23.0) are shown for the LECT2 sequence, calculated using ZipperDB. B) A His-tagged, full-length LECT2 polypeptide was expressed and purified using a His-Trap column under the illustrated gradient. C) Purity of the fractions was evaluated by gel electrophoresis. The purest fractions: 14-17, 19 (red) were pooled and concentrated. D) The recombinant protein was used to form fibrils and Thioflavin T fluorescence was monitored over 48 hours of growth. Red line represents the mean of samples done in triplicate, error bars represent standard deviation. E) The resulting fibrils from Thioflavin T analysis were imaged by negative stain TEM.

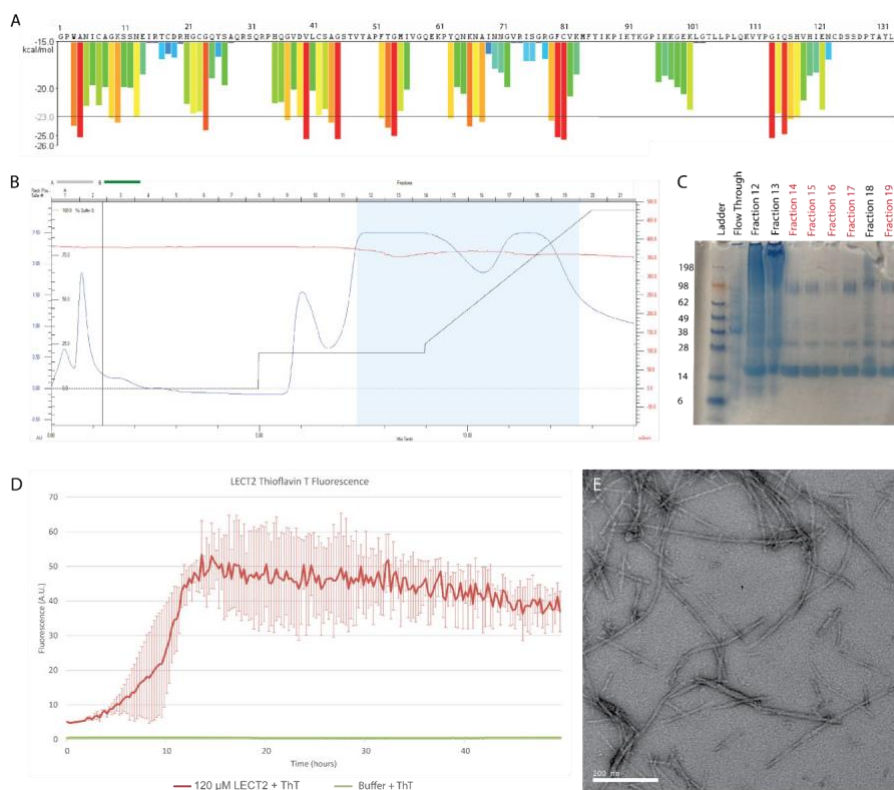


Figure S2: X-ray Diffraction from LECT2 fibrils. Related to Star Methods. A) X-ray diffraction exposure of bundles recombinant LECT2 fibrils. B) Radial profile generated from the diffraction image indicates a strong diffraction ring at 4.69Å.

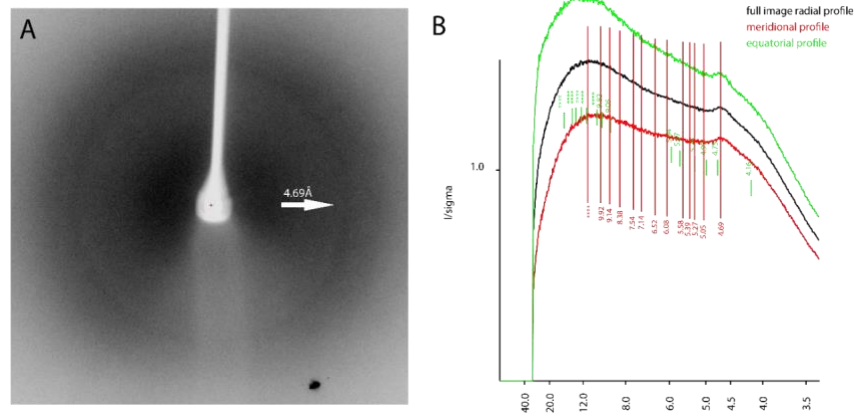


Figure S3: Vitrified LECT2 fibrils, CTF analysis, and 2D classification. Related to Figure 1 and STAR Methods. A) Example image of the vitrified fibrils used for cryo-EM data collection. B) Example CTF image showing the signature fibril diffraction profile at 4.69Å. C) Representative 2D class of the fast-twisting fibril species from a 1024 pixel box. D) Representative 2D class of the slow-twisting fibril species from a 1024 pixel box. E) A slow-twisting fibril image from 2D class after classification using a 384 pixel box in Relion. F) Composite image of 384 pixel box size 2D classes from the slow-twisting fibril polymorph stitched together to show its full twist. G) *De novo* 3D model generated from 2D classes in Relion viewed from the side and H) as a cross section.

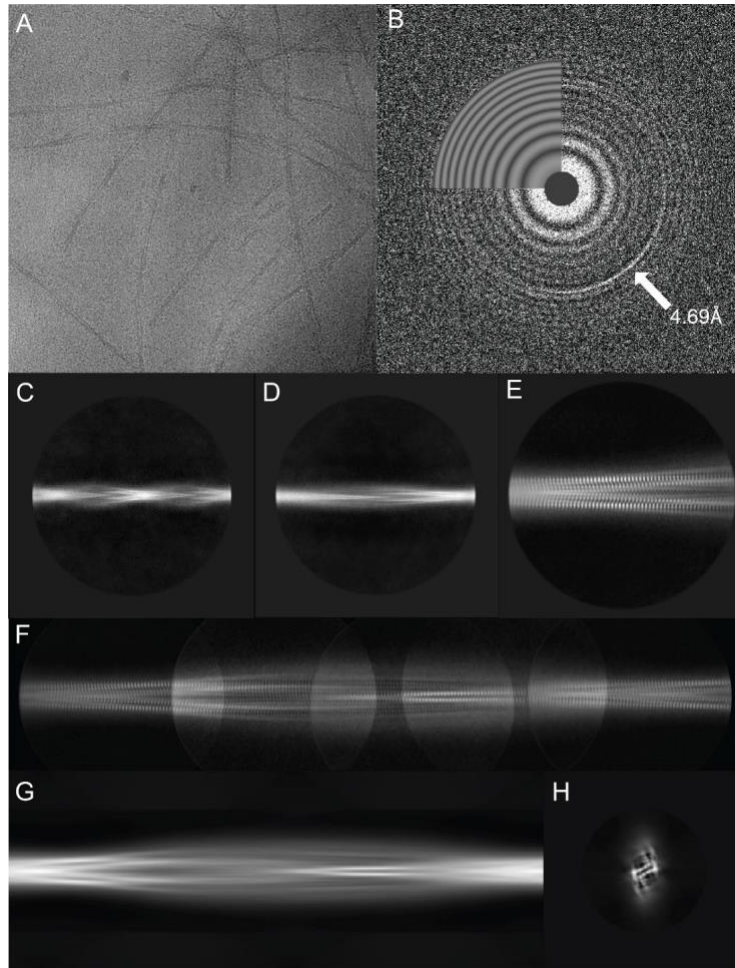


Figure S4: LECT2 fibril stability. Related to Star Methods. A) Nephelometer turbidity readings for the incubated fibril samples over 24 hours. B-E) Negative stain TEM images of the fibrils remaining after 24 hours incubation in fibrillization buffer (B), water (C), 3M urea (D), and 40nM Proteinase K (E). Average of samples done in triplicate shown, error bars represent standard deviation.

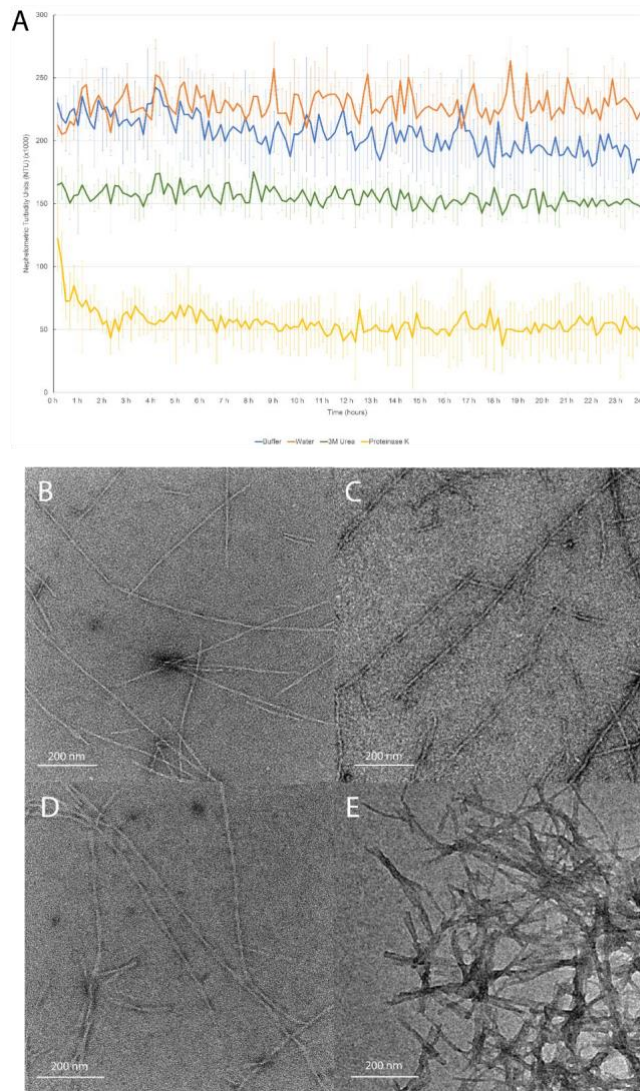


Figure S5: Cryo-EM grid optimization. Related to Star Methods. Recombinant fibrils were plunge frozen in buffer at different dilutions and Vitrobot blot settings for optimal particle dispersion, representative micrographs are shown. A) No dilution results in clumping and overcrowding. B) Addition of 0.25% glycerol improves ice distribution, but 2-fold dilution does not reduce crowding. C) 6-fold dilution improves fibril distribution in grid holes. D) Reducing the blot force improves ice distribution and maintains fibril distribution.

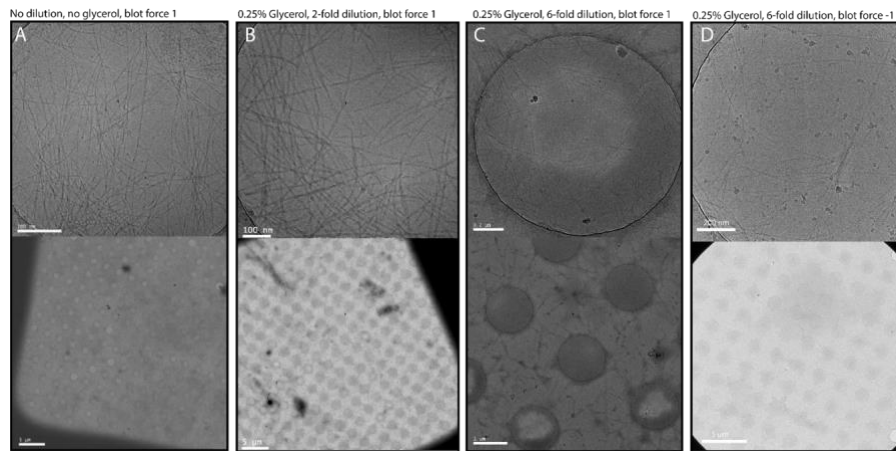


Figure S6: Cryo-EM data processing and model comparisons. Related to Star Methods. A) Overview schematic representation of helical reconstruction of fibril species performed in RELION showing maps of fibril sides and cross sections. B) FSC curves between independently refined half-maps which were phase randomized (red), unmasked and corrected (green), masked (navy blue), or masked and corrected (black). Final resolution of 2.7Å was calculated using a cutoff of FSC=0.143.

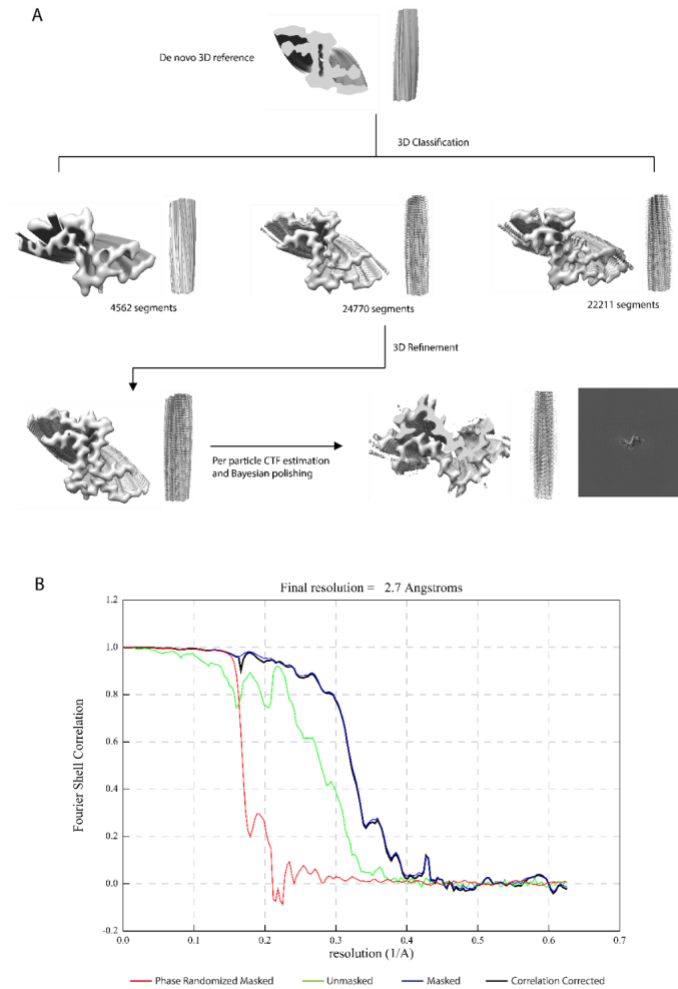


Figure S7: Identification of sequence matching LECT2 fibril structure by threading and energy minimization in Rosetta. Related to Figure 3 and Star Methods. A) Threading analysis indicates that the sequence beginning at Met55 is the most energetically favorable among possible assignments for the cryo-EM map (red circle). B) Energy-minimized, fast relaxed model from Rosetta predicts overall fibril core morphology as well as the stabilizing polar ladders in the hydrophilic pockets (inset).

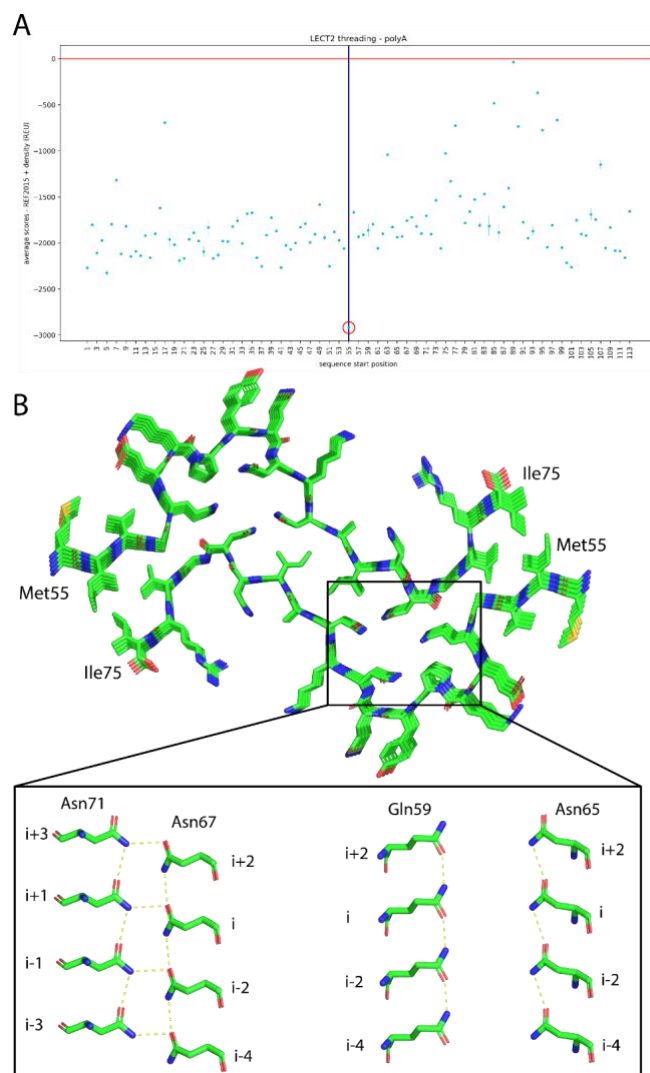


Figure S8: LECT2 globular fold and solvent accessibility of residues 55-75. Related to Star Methods. A) Views of LECT2 in its globular form (PDB:5B0H)¹⁷. Residues 55-75 (colored in red) are identified as forming the amyloid core of LECT2 in its fibril form. B) Solvent accessible surface area per residue of LECT2 in its globular fold calculated using the Ochanomizu University program (ver. 1.2).

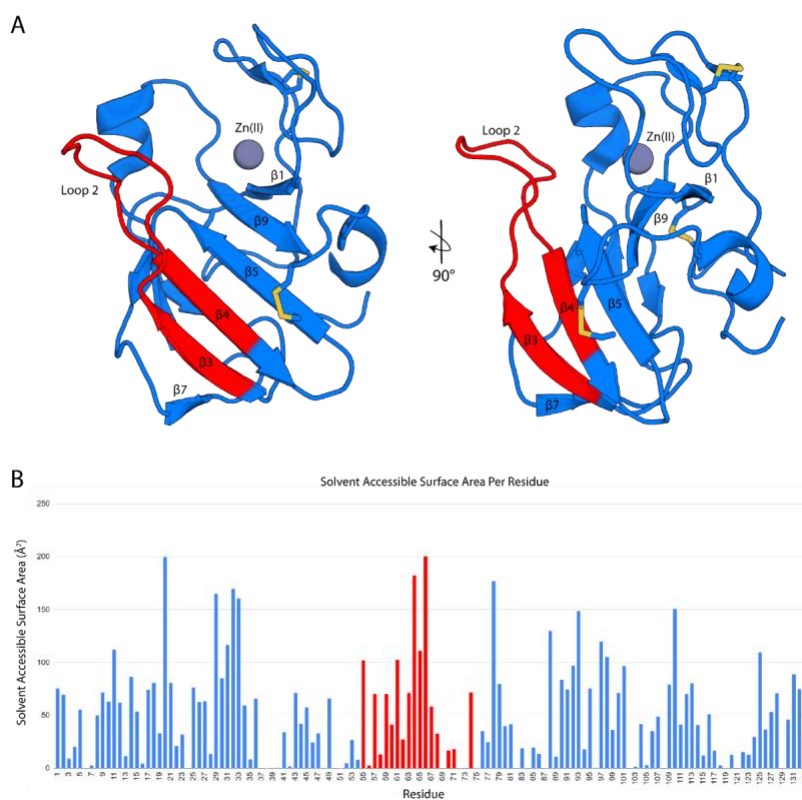


Figure S9: Mass spectrometry of Proteinase K digested fibrils. Related to Star Methods. A) SDS-PAGE shows stable protein species present over one hour of Proteinase K digestion. Boxed band was excised and peptides were extracted for GeLC-MS/MS (inset) with sequence segments detected colored in green.

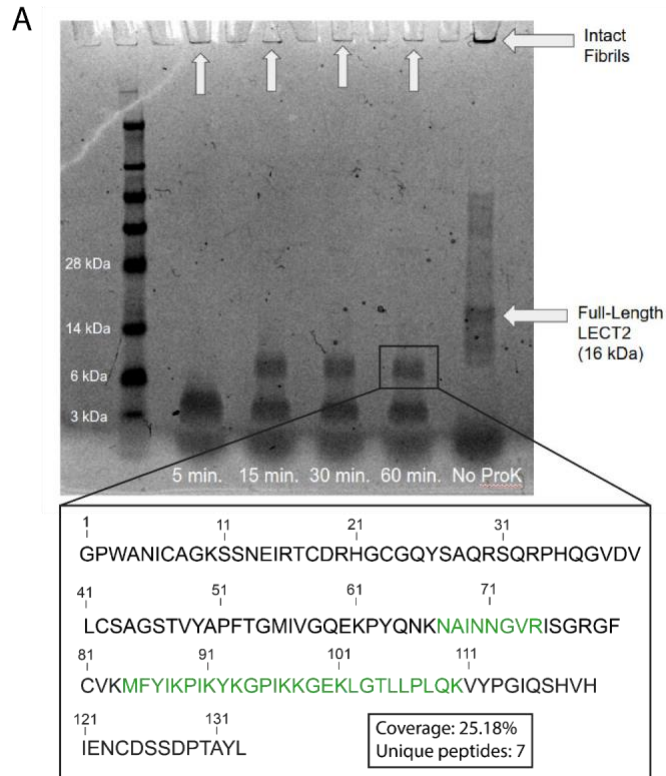


Table S1: Summary of Proteinase K digested LECT2 peptides identified by GeLC-MS/MS. Related to Star Methods.

Annotated Sequence	Modifications	MH+ [Da]	Theo. MH+ [Da]	Charge	m/z [Da]	XCorr
[K].NAINNGVR.[I]		857.4591	857.4588	2	429.2332	1.62
[K].mFYIKPIK.[Y]	M1(Oxidation)	1055.596	1055.596	3	352.5368	1.09
[K].mFYIKPIK.[Y]	M1(Oxidation)	1055.596	1055.596	2	528.3015	1.85
[K].mFYIKPIKYK.[G]	M1(Oxidation)	1346.753	1346.754	3	449.5891	1.43
[K].mFYIKPIK.[Y]	M1(Oxidation)	1055.599	1055.596	3	352.538	1.26
[K].mFYIKPIK.[Y]	M1(Oxidation)	1055.596	1055.596	2	528.3016	0.61
[K].KGEKLGTLPLQK.[V]		1424.884	1424.884	3	475.6328	0.45
[K].mFYIKPIK.[Y]	M1(Oxidation)	1055.596	1055.596	3	352.537	0.79
[K].mFYIKPIKYKGPIK.[K]	M1(Oxidation)	1742.006	1742.007	4	436.257	0.95
[K].GEKLGTLPLQK.[V]		1296.789	1296.789	3	432.9346	1.48
[K].LGTLLPLQK.[V]		982.6294	982.6295	2	491.8184	0.5
[K].LGTLLPLQK.[V]		982.6304	982.6295	2	491.8188	1.56
[K].LGTLLPLQK.[V]		982.6304	982.6295	2	491.8189	1.3

*Table includes only peptides mapped to the human LECT2 protein sequence used in these experiments.

Figure S10: Modelled domain-swapping for globular LECT2 dimer. Related to Figure 2. A) Modelled conformation of the LECT2 dimer (cyan, pink) domain-swapped through loop 2 using the ALECT2 fibril core structure (blue, red). B) ALECT2 fibril structure pictured above the domain-swapped model to highlight conformational similarities.

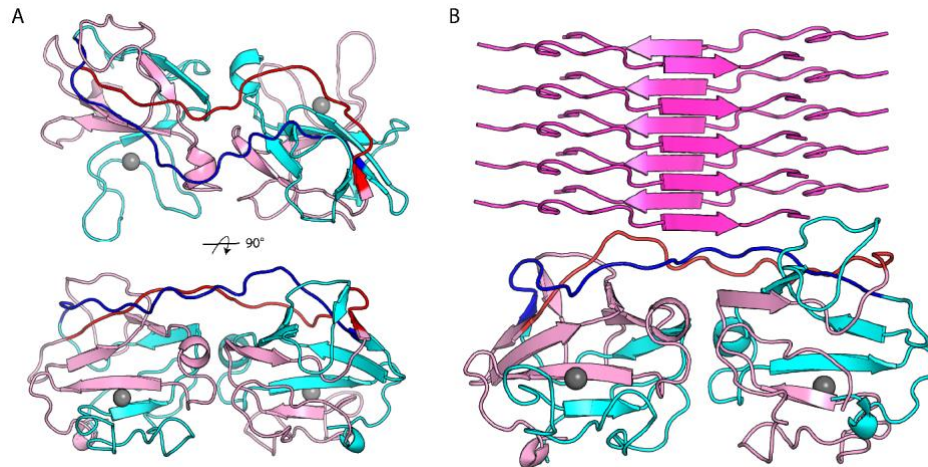
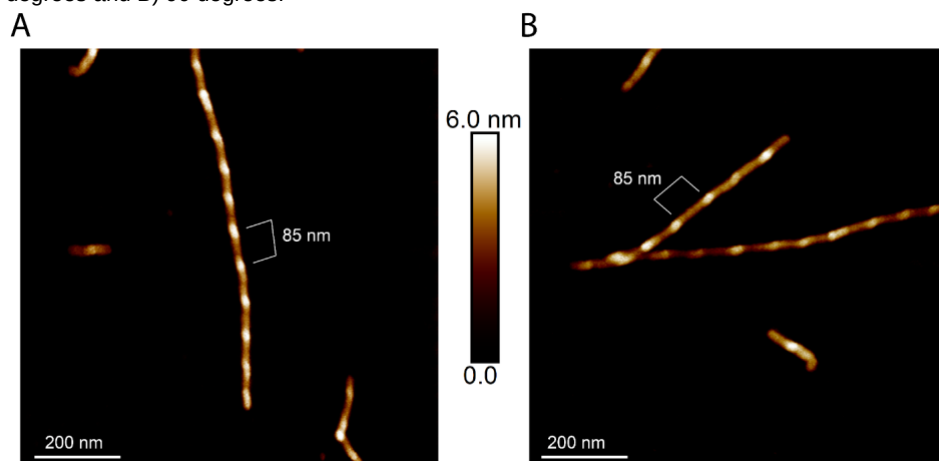


Figure S11: Atomic force microscopy of ALECT fibrils. Related to Star Methods. AFM images of ALECT2 fibrils show left-handed helical twisting and a crossover of 85nm at scan angles of A) 0 degrees and B) 90 degrees.



Data S1. PyRosetta Sequence Threading Code, related to STAR Methods

```
import os
import numpy as np
import pandas as pd
from pathlib import Path
import csv

from pyrosetta import *
from pyrosetta.rosetta import *
from pyrosetta.teaching import *
from pyrosetta.toolbox import *

from rosetta.protocols.cryst import *
from rosetta.protocols.rosetta_scripts import *

np.set_printoptions(threshold=sys.maxsize)

init('-ignore_unrecognized_res -load_PDB_components false -ignore_zero_occupancy false')

#####
### INPUTS ###
#####
### protein name
protName = 'lect2'

### to build symmetry pose
symmInfo = 'lect2.symm'
pose_file = 'lect2_INPUT.pdb'

### start/end residues in pose numbering for base layer of symmetric pose
base_start = 85
base_end = 105

### for job distribution
job = int(sys.argv[1])

### sliding window size
window_size = 21

### full sequence txt file to be threaded
seq_file = 'seq.txt'

### output name for scores file
scores_output = 'lect2_scores.csv'

#####
### PREPROCESSING ###
#####
### pre process sequence file
seq = Path(seq_file).read_text()
seq = seq.replace("\n", "")
seq_list = list(seq)
```

```
#####
### FUNCTIONS ###
#####
scorefxn = get_fa_scorefxn()

def symmetrize_pose(pose):
    """ set up symmetric pose """
    pose_symm_data = core.conformation.symmetry.SymmData(pose.total_residue(), pose.num_jump())
    pose_symm_data.read_symmetry_data_from_file(symmInfo)
    core.pose.symmetry.make_symmetric_pose(pose, pose_symm_data)
    sym_info = pose.conformation().Symmetry_Info()

    print("AssymUnit? equivalent_res")
    for i in range(1, pose.size()+1):
        print(i, sym_info.bb_is_independent(i), sym_info.bb_follows(i))

    print("Total Subunits:", sym_info.subunits())

    return pose

def def_windows(x, wnd_sz):
    """ define sequence positions for each window """
    window_list = []

    start_pos = 0
    end_pos = start_pos + wnd_sz

    num_windows = len(x) - wnd_sz + 1

    for j in range(0, num_windows):
        window_tmp = []
        for i in range(start_pos, end_pos):
            window_tmp.append(x[i])
        start_pos += 1
        end_pos += 1
        window_list.append(window_tmp)

    return window_list

def thread(pose, x, base_start, base_end):
    """ perform threading
        x = sequence window list """

    j = 0
    for i in range(base_start, base_end+1):
        mutate_residue(pose, i, x[job][j])
        j += 1

def FRwDensity(pose, rpt_num):
    """ set up and perform FastRelax with cryoEM density map """

    ## set up density FastRelax
```

```

    setup_dens = XmlObjects.static_get_mover('<LoadDensityMap name = "loaddens"
mapfile="postprocess.mrc"/>')
    setup_dens.apply(pose)
    setup_dens_pose = rosetta.protocols.electron_density.SetupForDensityScoringMover()
    setup_dens_pose.apply(pose)

    ## set up score function with correct weights
    score = get_score_function()
    score_dens_cart = create_score_function('ref2015_cart')
    score_dens_cart.set_weight(rosetta.core.scoring.elec_dens_fast, 25)
    score_dens_cart.set_weight(fa_elec, 1.5)

    ## set up fast relax parameters
    mmf = pyrosetta.rosetta.core.select.movemap.MoveMapFactory()
    mmf.all_bb(setting=True)
    mmf.all_bondangles(setting=True)
    mmf.all_chi(setting=True)
    mmf.all_jumps(setting=True)
    mmf.set_cartesian(setting=True)

    FR = pyrosetta.rosetta.protocols.relax.FastRelax(scorefxn_in=score_dens_cart, standard_repeats=1)
    FR.cartesian(True)
    FR.set_movemap_factory(mmf)
    FR.min_type("lbfgs_armijo_nonmonotone")

    ## apply FR, score relaxed pdb, and output file
    FR.apply(pose)
    output_name = 'lect2_FR_thread_' + str(job) + '_' + str(rpt_num) + '.pdb'
    pose.dump_pdb('./output/' + output_name)
    scores = output_name, score_dens_cart(pose), scorefxn(pose)

    return scores

#####
### RUN ###
#####

window_list = def_windows(seq_list, window_size) ## build list of sequence windows
seq_list_df = pd.DataFrame(window_list)         ## convert to df

for i in range(0, 3):
    pose = pose_from_pdb(pose_file)             ## create pose object
    pose = symmetrize_pose(pose)                 ## symmetrize the pose

    thread(pose, window_list, base_start, base_end) ## perform threading; mutate side chains on
    pose object to reflect the current sequence window
    scores = FRwDensity(pose, i)                 ## perform FastRelax on current sequence window

```

Appendix Chapter 2

This chapter is based on the preprint titled “Structure of a reversible amyloid fibril formed by the CPEB3 prion-like domain reveals a core sequence involved in translational regulation.”

Flores, M. D., Sawaya, M. R., Boyer, D. R., **Zink, S.**, Tovmasyan, S., Saucedo, A., Zee, C.-T., Cardenas, J., Fioriti, L., & Rodriguez, J. A. (2022). Structure of a reversible amyloid fibril formed by the CPEB3 prion-like domain reveals a core sequence involved in translational regulation. *BioRxiv : The Preprint Server for Biology*

Title: Structure of a reversible amyloid fibril formed by the CPEB3 prion-like domain reveals a core sequence involved in translational regulation

Maria D. Flores^{*1}, Michael R. Sawaya², David R. Boyer², Samantha Zink¹, Susanna Tovmasyan¹, Adrian Saucedo¹, Chih-Te Zee¹, Jorge Cardenas¹, Luana Fioriti^{*3,4} and Jose A. Rodriguez^{*1}

Affiliations

¹Department of Chemistry and Biochemistry; UCLA-DOE Institute for Genomics and Proteomics; STROBE, NSF Science and Technology Center; University of California, Los Angeles (UCLA); Los Angeles, CA 90095, USA.

²Department of Biological Chemistry and Department of Chemistry and Biochemistry, University of California Los Angeles (UCLA), Howard Hughes Medical Institute (HHMI), UCLA-DOE Institute for Genomics and Proteomics, Los Angeles, CA 90095, USA.

³Department of Neuroscience, Istituto di Ricerche Farmacologiche Mario Negri, Milano MI, 20156, Italy.

⁴Department of Neuroscience, Columbia University, New York, NY 10027, USA.

* Correspondence to: Jose A. Rodriguez (jrodriguez@mbi.ucla.edu), Luana Fioriti (luana.fioriti@marionegri.it), Maria D. Flores (mflores@mbi.ucla.edu).

Abstract

The cytoplasmic polyadenylation element-binding protein 3 (CPEB3) is a prion-like RNA-binding polypeptide. As a functional prion, CPEB3 is thought to modulate protein synthesis at synapses and enable consolidation of long-term memory in neurons. Here, we report that the prion-like domain 1 of CPEB3 self-assembles into labile amyloid fibrils *in vitro*. A cryoEM structure of these fibrils reveals an ordered 48-residue core, spanning L103 to F151. CPEB3 constructs lacking this amyloidogenic segment form abnormal puncta in cells when compared to wild type CPEB3, with reduced localization in dormant p-bodies and increased localization in stress granules. Removal of the amyloid core segment in CPEB3 also abolishes its ability to regulate protein synthesis in neurons. Collectively, this evidence suggests that the newly identified amyloidogenic segment within the CPEB3 prion domain is important for its regulated aggregation in cells and suggest its involvement in regulating translational activity and potentially long-term memory formation.

Introduction

Prions, or proteinaceous infectious particles¹, can form stable, long-lived pathogenic or functional assemblies in species throughout the tree of life². Prion proteins are thought to rely on an amyloid architecture to recruit soluble proteins and template their conversion of into ordered fibrillar β -sheet-rich assemblies^{2,3}. Because of their ability to form amyloids, prion-like proteins present a challenge to the proteostatic framework in neurons⁴. However, since prions can endure protein turnover, they also represent a possible substrate for memory consolidation⁵. In fact, functional prions have been proposed to enable the persistence of memory engrams over time^{6,7}.

A range of studies have provided evidence for prion-based memory consolidation mechanisms operating in neurons in *Aplysia*⁸, *Drosophila*^{9,10} and most recently, mammals^{7,11}. Neuronal isoforms of the cytoplasmic polyadenylation element binding protein (CPEB) are involved in regulating localized translation at synapses⁴. Of the mammalian CPEB orthologs, CPEB3 distinctly contains a Q-rich N-terminal domain¹⁰, resembling those found in functional yeast prions². CPEB3 is a 716-residue protein with two N-terminal prion-like domains linked by a short, low-complexity regulatory motif (Figure 1a). C-terminal to the CPEB3 prion domains are two RNA-binding domains and a zinc finger domain¹⁰ (Figure 1a). As a functional RNA binding protein, CPEB3 regulates translation of target mRNAs that are required for synaptic plasticity and the growth of new synaptic spines^{11–13}. The prion-like nature of CPEB3 is supported by work showcasing its ability to form heritable aggregates in yeast⁷. Similar aggregates appear to be required for its function in neurons⁷. The N-terminal, prion-like domain 1 (PRD1) of CPEB3 is important to these traits as it rescues aggregation phenotypes in cells and mediates long-term potentiation (LTP) as well as the persistence of memory in mice¹¹.

CPEB3 aggregates are tightly regulated and rely on mono-ubiquitination¹² and SUMOylation¹⁴, which allow for proper activation and repression of the prion-like protein, respectively. In its basal state, CPEB3 is SUMOylated and colocalizes to p-bodies where it inhibits translation of target mRNAs¹³. Upon proper signal induced activity it is then shuttled to polysomes where it increases translation of its targets¹³. These observations suggest that CPEB3's localization is directly related to its function.

A structure of fibrillar aggregates formed by the *Drosophila* CPEB ortholog, Orb2, isolated from fly brains, revealed a hydrophilic amyloid core with pronounced steric zipper motifs. The role played by this structure is consistent with the notion that CPEB can form persistent or stable amyloid assemblies, but how these assemblies form and how long they persist in cells remains unclear¹⁵. Additionally, while these studies present remarkable views of functional prion structures, they offer limited insights into the prion-like activity of mammalian CPEBs, and in particular CPEB3.^{7–9} Our current view of mammalian CPEB structure is instead informed by sequence analyses¹⁶ and NMR spectroscopy,¹⁷ which has revealed segments in PRD1 of CPEB3 that may form higher-order coiled-coiled α -helical assemblies. These efforts have contributed to the current delineation of certain CPEB domains as prion-like^{11,7}.

We set out to understand the structural basis for prion-like activity by CPEB3 and found that recombinant CPEB3 PRD1 can self-assemble into reversible amyloid fibrils in solution. The structure of these fibrils, determined by single particle cryoEM, showed an ordered core composed of residues 103 to 151 of CPEB3 arranged in a serpentine fold. The sequence composition of this amyloid core (AC) offers clues to its lability and possible sensitivity to environmental changes in pH. CPEB3 constructs lacking this AC sequence showed altered subcellular distribution, indicating a role for this sequence in CPEB3 activity. This hypothesis is confirmed by the altered translational regulation of AC-deficient CPEB3 in primary neurons. These results suggest a model whereby CPEB3 function is regulated in part by the AC sequence in PRD1, possibly through its adoption of a labile amyloid fold.

Results

CPEB3 PRD1 forms reversible amyloid fibrils *in vitro*.

Hypothesizing that CPEB3 PRD1 may form functional amyloid assemblies, we expressed and isolated a construct of human CPEB3 encoding its first 217 residues that included the entirety of its conserved PRD1 (Panel 1, Supplementary Figure 1). This region of CPEB3 has been

previously identified as essential for prion-seeding, heritability⁷ and is essential for LTP in mice¹¹. We assessed the ability of this construct (CPEB3₁₋₂₁₇) to form fibrils *in vitro* by light scattering, enhanced fluorescence of the amyloid-binding dye ThT, and negative stain electron microscopy (Supplementary Figure 2). CPEB3₁₋₂₁₇ tended to precipitate in denaturing solutions as amorphous aggregates, twisting fibrils, or flat, ribbon like sheets. In solutions containing 125 mM NaCl, 50 mM Tris-Base, 10mM K₂HPO₄ and 5 mM glutamic acid at pH between 4 and 5 (assembly buffer) CPEB3₁₋₂₁₇ produced what appeared by negative stain electron microscopy to be fibrils with a homogenous twist (Supplementary Figure 2b). Solutions containing CPEB3₁₋₂₁₇ fibrils exhibited enhanced ThT fluorescence (Supplementary Figure 2a). These fibrils were also labile, disassembling when diluted into water or exposed to low ionic strength solutions (Supplementary Figure 2c). To pursue structure determination in the face of this lability, fibrils had to be vitrified within an hour of formation in assembly buffer.

Structure of the CPEB3 amyloid fibril core

Solutions containing CPEB3₁₋₂₁₇ fibrils were interrogated by single particle cryoEM, with the goal of *de novo* structure determination (Supplementary Figure 4). CryoEM images confirmed that CPEB3₁₋₂₁₇ forms rapidly twisting amyloid fibrils in assembly buffer, with a single polymorph representing over 90% of imaged particles (Supplementary Figure 5a). The remaining 10% of species exhibited variable twist or no twist, and therefore eluded further classification and structure determination (Supplementary Figure 5b). Two-dimensional class averages confirmed canonical amyloid features corresponding to the stacking of β -sheets separated by 4.8 Å inter-strand distances, as well as a disordered flanking region, or fuzzy coat, surrounding the fibril core (Supplementary Figure 5b). Three dimensional reconstructions obtained from a subset of 40,329 segments yielded a 3.0 Å map that showed density for a single asymmetric protofilament that fit residues L103-F151 of CPEB3 (Figure 1b). We refer to this structure and its corresponding core sequence as the CPEB3 amyloid core (CPEB3_{AC}).

CPEB3_{AC} fibrils carry a helical pitch of 536 Å and twist of -3.32° (Figure 1c). The length of the fibrils arises from stacking of CPEB3_{AC} molecules in parallel and in register intermolecular β -sheets. The tight serpentine fold adopted by the CPEB3_{AC} core is enabled by five sharp kinks interspersed among the molecule's seven β -strands (Figure 1c). The fold exhibits two hydrophobic pockets and encloses three internal solvent channels. The CPEB3_{AC} core displays a total solvent accessible surface area per chain of 1653 Å². At its N-terminus, L103 contacts a surface-facing segment of a beta turn in the fibril core, making a hydrophobic pocket composed of residues P105, F107, G126 and I127 (Supplementary Figure 7b). An additional hydrophobic pocket is formed by residues F124, V130, M134 and F136 (Supplementary Figure 7b). The core is further stabilized by stacking of aromatic residues, although some appear solvent facing, including W111, F139, and F151. Polar residues T110, S112 and T116 face the largest solvent channel buried within the fibril core. A second, smaller solvent channel is lined by a group of largely by hydrophobic residues: V119, P139, V143, P145; E120 stands as a lone hydrophilic residue in this group. This same glutamic acid forms part of a minute channel within the core, which is also lined by S122 and N138. The significance of E120 bridging these two solvent channels is unclear, but may indicate a role for it in the lability of the CPEB3_{AC} core.

Structural features of labile CPEB_{AC} fibrils

Energy values calculated using coordinates of the CPEB3_{AC} fibril core show that, per layer, this core is less stable than those of other pathogenic amyloids (Supplementary Figure 7a)¹⁸. Estimates of its free energy are more in line with those of fibril structures formed by reversible, functional amyloids¹⁹⁻²¹. Solvent facing hydrophobic residues and buried solvent channels play a role in its predicted lability, but the CPEB3_{AC} is also strained by a series of sharp kinks. Some of these kinks tilt strands out of plane, warping the layers of the fibril to create a height

difference of 14.5 Å across each of its rungs (Figure 1d, Supplementary Figure 7c). While the role of warping in amyloid fibril cores remains unclear, it may help relieve the strain associated with the large twist angle observed in CPEB3_{AC} fibrils. In addition, some kinks appear to be stabilized by local interactions. For example, a kink near G115 facilitates hydrogen-bonding between side chains of S112 and T116. Four of the 28 proline residues in CPEB3₁₋₂₁₇ are resolved in the core. The remaining 24 proline residues are unresolved in the fuzzy coat of CPEB3_{AC} and may function to limit the size and stability of the ordered core. They appear as smeared density in 2D averages of CPEB3_{AC} fibrils (Supplementary Figure 4b). While some of these regions are hypothesized to form coiled-coils¹⁶, that region of the fibril remains insufficiently resolved to assign secondary structure.

The role of negatively charged residues in CPEB_{AC}

We observed four potentially charged residues in CPEB3_{AC}: two histidine residues (H141 and H142) are solvent exposed, and two acidic residues (E120 and D121) are buried within the fibril core (Supplementary Figure 7a). The two histidine residues are located too far from the acid residues for charge complementation to stabilize the fold. E120 is located in the second solvent channel and is otherwise surrounded by hydrophobic residues, while D121 faces the opposing solvent channel (Supplementary Figure 7b). As CPEB3_{AC} aggregates near a pH of 5, E120 and D121, are anticipated to remain protonated within the fibril core and are capable of forming stabilizing hydrogen bonds. These same residues would render the core susceptible to pH triggered changes, as observed with other functional amyloids^{22,23} such as pmel17^{22,23} and β-endorphin²³. To evaluate this hypothesis, we monitored the stability of CPEB3_{AC} fibrils in response to changing buffer conditions, using a ThT fluorescence assay. We found CPEB3_{AC} fibrils diminish in abundance as pH increases or as the ionic strength of the buffer decreases, denoted by a decrease of ThT signal and the absence of fibrils in negative stain images compared to fibrils that remain in their assembly buffer (Supplementary Figure 2c). This dissolution in response to changing environmental conditions is recapitulated by simulations tracking fibril stability over a period of 200 nanoseconds (Supplementary Figure 8a). Interlayer distances in a 5-layer fibrillar CPEB3_{AC} structure are disrupted as a pH increase from 3 to 7 is simulated; structural deviations are pronounced in charged residues E120 and D121 (Supplementary Figure 8b).

To assess whether E120 and D121 influence the morphology, distribution, and abundance of CPEB3 deposits in cells, we expressed a variety of GFP-tagged full-length CPEB3 constructs encoding wild type and E120/D121 variations, in HT22 and U87 neuronal cells (Figure 2a,b). In HT22 and U87 cells, wild type, GFP-fused CPEB3 formed fine, granular fluorescent puncta, as previously described (Figure 2c, Supplementary Figure 9a)^{7,13}. Forty-eight hours after transfection, the fine puncta were replaced by larger puncta that were associated with reduced cellular health, likely due to prolonged CPEB3 overexpression (Supplementary Figure 9a). To probe the influence of acidic side chains at residues ¹²⁰ED¹²¹ we expressed a CPEB3 variant encoding ¹²⁰QN¹²¹, which retains the polarity and size of the wild type residues but ablates their negative charge. We also expressed a variant encoding ¹²⁰AA¹²¹ to assess the effect of ablating polarity and reducing the size of wild type residues. Puncta formed by both constructs differed morphologically from those formed by wild type CPEB3-GFP under similar conditions (Figure 2c, Supplementary Figure 11). CPEB3-¹²⁰QN¹²¹ generally formed irregularly sized and shaped puncta, while CPEB3-¹²⁰AA¹²¹ formed fewer puncta that were intermediate in size and remained more regularly circular after 24hrs (Figure 2d, Supplementary Figure 10a, Supplementary Figure 11). The effects of these residue substitutions on the size and distribution of CPEB3_{AC} puncta implicate a functional role for E120 and D121.

CPEB3_{AC} is required for proper aggregation and localization in cells.

To more generally assess the functional relevance of CPEB3_{AC} in cells, we asked whether the presence of the AC sequence affected CPEB3 aggregation. While CPEB3 is known to form aggregates of various sizes when overexpressed in cells^{7,11,13}, a construct of CPEB3 lacking its amyloid core, CPEB3_{ΔAC} displayed a reduced capacity to form small aggregates (Supplementary Figure 9). Instead, this construct showed a delayed onset of larger aggregate formation as compared to wild type CPEB3 in both HT22 and U87 cells (Supplementary Figure 9). This prompted us to determine whether the localization of CPEB3_{ΔAC} differed from that of wild type CPEB3 in both cell lines. Under normal conditions CPEB3 is expected to localize in p-bodies¹³ that contain mRNAs and associated regulatory proteins. There, mRNAs are stored in a dormant state while they are trafficked to their destination to be translated, or can be degraded. Alternatively, CPEB3 could reside in Stress granules (SG), induced to promote cell survival by condensing translationally stalled mRNAs. We observed that wild type CPEB3-GFP aggregates primarily colocalized with the de-capping coactivator and p-body marker, Dcpa1, in unstimulated cells (Figure 3b). In contrast, in mouse HT22 hippocampal neurons, CPEB3_{ΔAC} aggregates primarily colocalized with the SG assembly endoribonuclease, G3BP, and minimally colocalized with the p-body marker, Dcpa1 (Figure 3b). Similar localization patterns were observed in human U87 glioblastoma cells (Figure 3c).

CPEB3_{AC} impacts translational regulation by CPEB3

In its basal state, CPEB3 represses translation of target mRNAs, is SUMOylated, and colocalizes to p-bodies. After neuronal stimulation CPEB3 undergoes de-SUMOylation, leaves p-bodies and is then shuttled to polysomes to promote translation of targets¹³. The increase in translational activity of CPEB3 is thought to be facilitated by monoubiquitination of PRD1 by Neuralized-1¹². The ubiquitination site of CPEB3 is currently unknown and efforts to identify and characterize all SUMOylation sites¹⁴ have not yielded a complete understanding of their role in CPEB3 activity. Since CPEB3_{AC} is situated within PRD1, we postulated it may influence translational regulation by CPEB3. To test this hypothesis, we assessed translation of CPEB3 targets, comparing the impact of wild type protein to CPEB3_{ΔAC}. Translation of Renilla luciferase fused to the 3'UTR of two established targets of CPEB3 targets, GluA2 and SUMO2 3' UTR, in HEK-293 cells, was strongly inhibited by both wild type CPEB3 and CPEB3 lacking PRD1 (ΔN) (Figure 3d). In contrast, translation of the same targets was unaffected by CPEB3_{ΔAC}; a marked change in translational regulation in the absence of the amyloid core segment. Co-expression of wild type and CPEB3_{ΔAC} in cells re-established the translational repression observed with the wild type protein alone, indicating CPEB3_{ΔAC} does not act in dominant negative fashion (Figure 3d).

Expression of wild type CPEB3 in primary neurons repressed translation of its targets in unstimulated cells (Figure 3e). To stimulate neurons, we used a protocol consisting of bath application of bicuculline and Glycine, a chemical treatment that induces long-term potentiation (LTP) of synapses (cLTP). As expected, glycine-mediated cLTP induced the translation of CPEB3 targets (Figure 3e). In contrast, expression of CPEB3_{ΔAC} under the same conditions ablated all translational regulation, rendering the protein essentially inactive in this capacity (Figure 3e). In fact, luciferase activity in the presence of CPEB3_{ΔAC} was most similar to that of cells expressing GFP alone, indicating that CPEB3_{AC} is required for regulated translation of CPEB3 targets.

Discussion

CPEB3, a member of the cytoplasmic polyadenylation family of proteins, binds mRNA transcripts to activate or repress protein synthesis^{24,25}. This process is hypothesized to be regulated by its prion-like self-assembly⁴, although the formation of prion-like structures by

CPEB3 remains hypothetical. We find that a segment in the first CPEB3 prion-like domain forms labile amyloid fibrils. The ordered core of fibrils formed by this segment, CPEB3_{AC}, is composed of a single asymmetric protofilament with parallel, in register strands. While the overall architecture of CPEB3_{AC} resembles that of a canonical amyloid structure, it contains buried polar residues and solvent channels that may decrease its stability. The ready assembly and disassembly of these fibrils would be a feature of CPEB3 assemblies in their role of translational regulation.

The amino terminal domains of many CPEB proteins are identified as prion-like in part due to their low-complexity sequences and their high number of uncharged polar residues^{4,8,26}. Orb2 contains prion-like domains that form three-fold symmetric amyloid assemblies as isolated from fly brains¹⁵; these fibrils contain only a short 30-residue segment at their core. Likewise, in CPEB3, a short 48-residue segment within its first prion domain can form amyloid fibrils. However, some important features distinguish CPEB3 from Orb2 fibrils. First, while the PRD1 of CPEB3 encodes a low complexity sequence, CPEB3_{AC} is not particularly rich in Q or N residues. Also, unlike pathogenic amyloid fibrils, and perhaps unlike Orb2 fibrils, CPEB3_{AC} is readily labile. Features that may contribute to the lability of CPEB3_{AC} include its buried charged residues (E120, D121), its four prolines, and its highly warped layers. The buried charged residues would allow CPEB3_{AC} to respond to pH changes²⁷, enabling environmental cues to possibly regulate assembly of amyloid-like CPEB3_{AC} structures in neurons alongside other post-translational modifications involved in CPEB3 regulation in cells^{12–14}.

The relevance of CPEB3_{AC} to the physiological role of CPEB3 is underscored by the fact that translational regulation by CPEB3 is ablated in cells that lack its amyloid core. The localization of CPEB3 to p-bodies is also sensitive to perturbations of CPEB3_{AC}. In the absence of its amyloid core, CPEB3 localizes to stress granules, whose morphological appearance resembles those of other phase separated compartments²⁸. The structural nature of CPEB3 within these aggregates, and whether it adopts an amyloid fold, remains unclear. CPEB3_{AC} may act as a regulatory element or may itself be subject to regulatory modifications such as SUMOylation¹⁴ and ubiquitination¹². Such modifications may enable proper subcellular localization of CPEB3 and help avoid wanton self-assembly. This hypothesis would explain why ablation of this sequence causes loss of translational activity and the formation of large puncta. While the structures of amyloid fibrils reconstituted *in vitro* are known to be distinct from those isolated from disease tissue^{29–33}, the relevance of CPEB3_{AC} to physiological CPEB3 function remains supported by: 1) the role of its AC sequence in localization to puncta in cells; 2) the fact that point mutations ablating charges in CPEB3_{AC} alter CPEB3 distribution in cells; and 3) that deletion of CPEB3_{AC} eliminates its ability to repress or activate mRNA targets, highlighting its role in translational activation by CPEB^{7,11} (Figure 4).

Our observation of CPEB3_{AC} provides additional experimental evidence for an expanding model of functional prions^{4,34} in which a labile amyloid state has evolved to perform biological functions^{15,20,26}. Future insights into the prion-like role of CPEB3 in neurons and its regulated assembly and disassembly in cells will continue to expand our fundamental understanding of functional prion proteins. Ultimately, the structure of CPEB3 assemblies and their interacting partners in the brain remain important targets in the quest to understand the molecular basis for cellular memory.

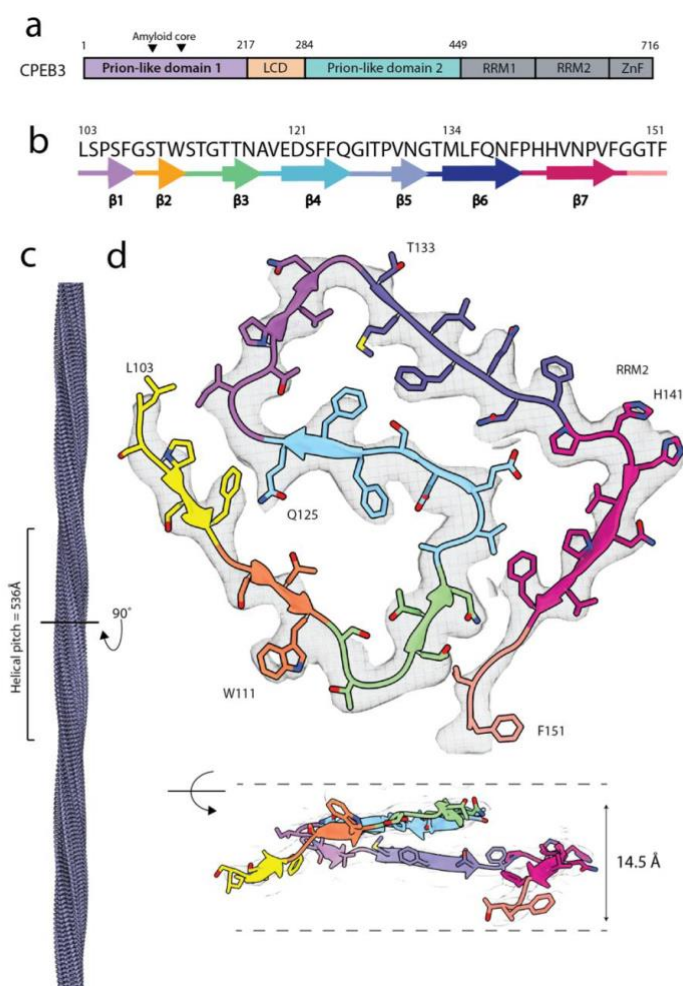


Figure 1. CryoEM structure of CPEB3's prion-like domain. **a**, Schematic summary of CPEB3 domains. **b**, Amino acid sequence alignment of fibril core corresponding to secondary structure β -strands. **c**, Model of CPEB3AC filaments with a corresponding helical pitch of 536 Å. **d**, Top: cryoEM map of cross section perpendicular to the helical axis with model rendering colored to match secondary structure schematic. Bottom: side view of map and model measuring height different within one helical layer.

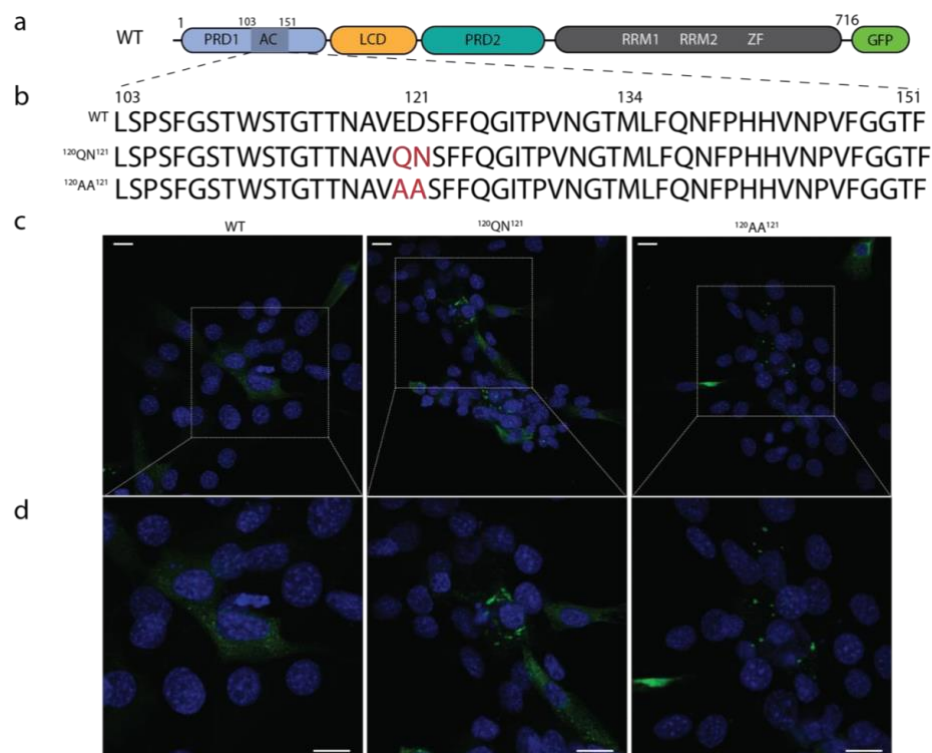


Figure 2. The role of E120 and D121 in CPEB3 puncta formation in cells. Wild type CPEB3-GFP shows fine, granular like aggregates in HT22 mouse hippocampal neurons after 24 hours of expression. **a**, Schematic of full-length WT construct expressed in cells. **b**, Sequence alignment of AC region in FL constructs with corresponding mutated residues in red. **c**, CPEB3-GFP WT and mutants (green) and nuclear staining (blue) in HT22 neurons. **d**, insets of panel c. Scale bar = 10 μ m.

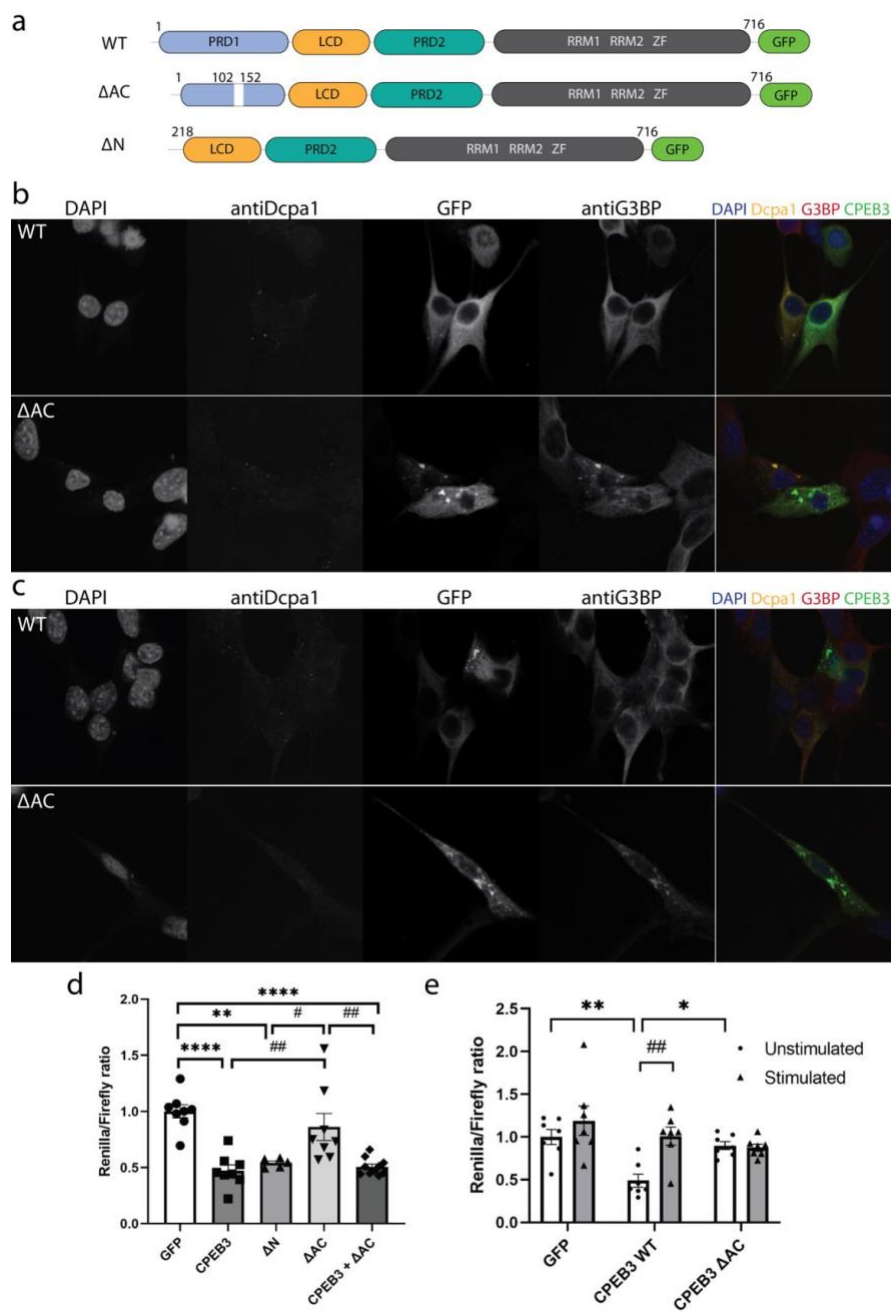


Figure 3. *CPEB3_{AC} regulates subcellular location of GFP-tagged CPEB3 in cells and affects its translational regulation.* **a**, Schematic representation of full-length constructs used in localization and translational assays. **b**, HT22 cells expressing WT (top) and AC (bottom) (green) and subsequent co-localization with antiDcp1 (yellow) and antiG3BP (red) with nuclear stain (blue). **c**, Identical experiment to Figure 1b using U87 cells. **d**, HEK 293T cells inhibitory renilla luciferase assays (n =5-11 replicates per construct.) Graphs show the quantification of renilla/Firefly ratio, mean $\pm$ s.e.m. One-way ANOVA, followed by Tukey's multiple comparisons post hoc test, ** p<0.01, ****p<0.0001 vs GFP; # p< 0.05, ## p<0.01, vs CPEB3 Δ AC. **e**, Renilla luciferase assays performed in primary neurons. (n=7 per condition). Graphs show the quantification of renilla/Firefly ratio, mean $\pm$ s.e.m. Two-way ANOVA, followed by Tukey's multiple comparisons post hoc test, * p< 0.05, ** p<0.01, vs CPEB3 unstimulated; ##p< 0.01 vs CPEB3 stimulated.

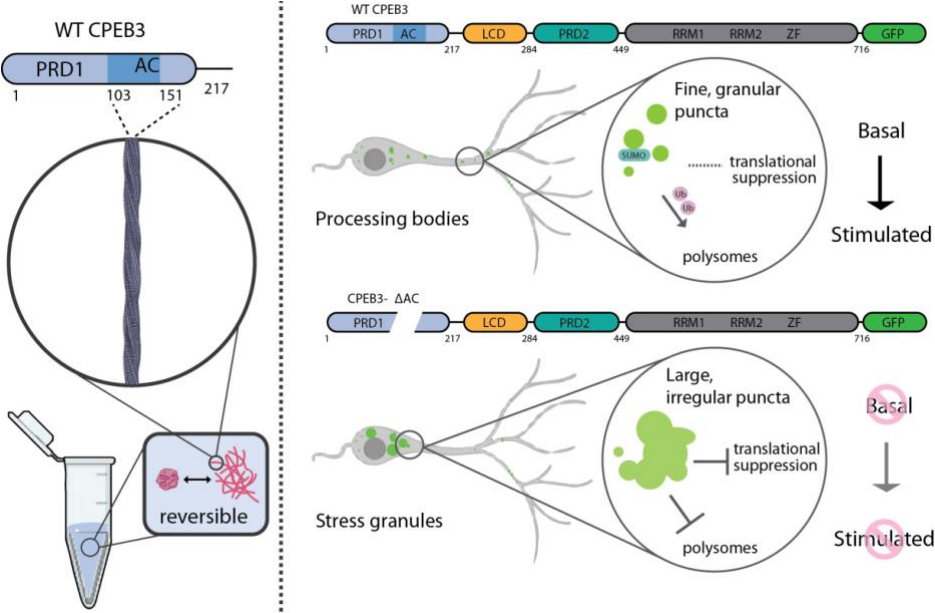
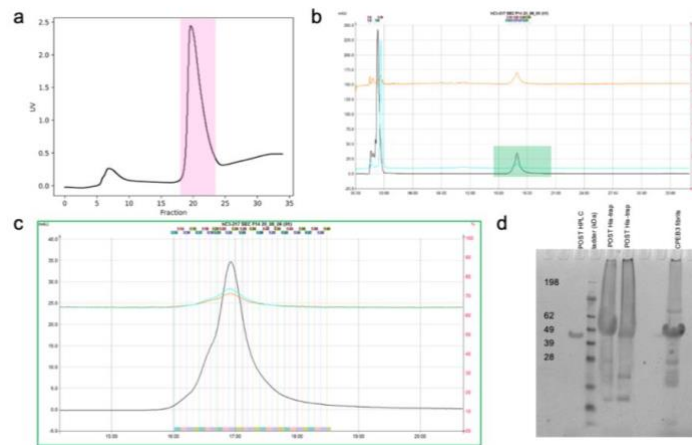


Figure 4. *Model of CPEB3 AC and its role in CPEB3 aggregation and regulation.* (Left panel) Amyloid fibrils formed in-vitro are reversible. Cryo-EM identifies a core sequence within the CPEB3 prion-like domain. (Right panel) WT CPEB3 displays fine, granular puncta that co-localize in p-bodies in the basal state and are moved to polysomes upon stimulation. CPEB3 Δ AC aggregates are larger and co-localize with SG markers, displaying no functional activity upon stimulation.

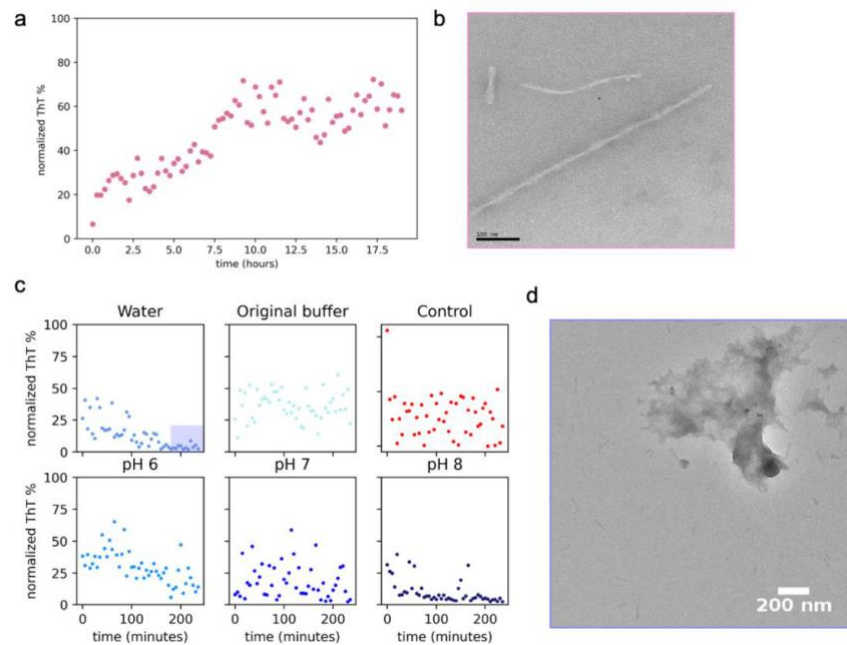
Table 1. Cryo-EM data collection, refinement and validation statistics

	CPEB3 _{AC}
--	---------------------

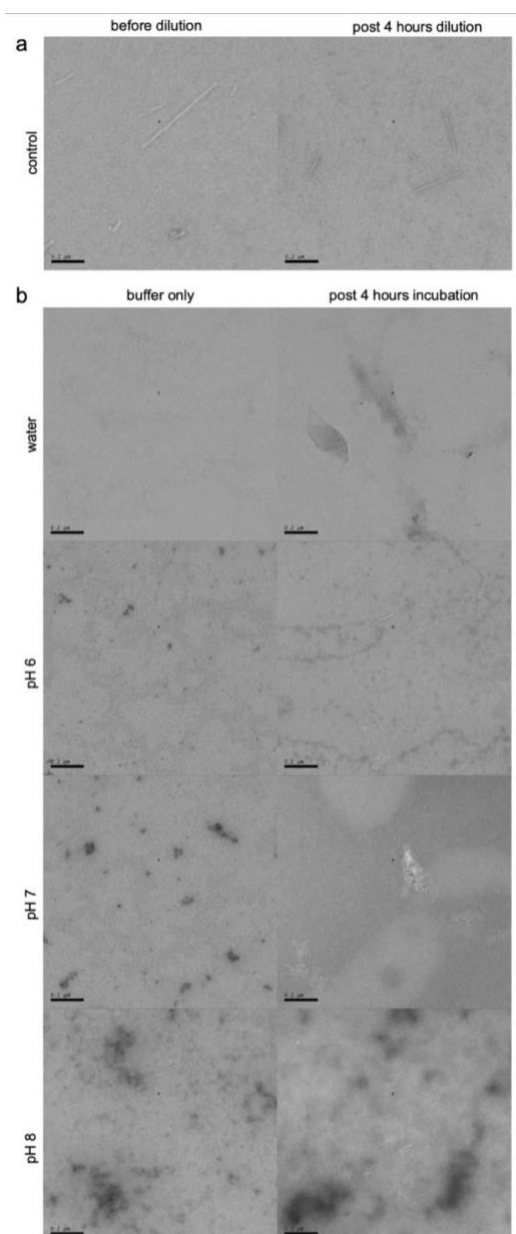
Data collection and processing	
Magnification	105,000
Voltage (kV)	300
Electron exposure (e-/Å ²)	52
Defocus range (µm)	1-3
Pixel size (Å)	0.86
Symmetry imposed	C1
Helical rise (Å)	4.8
Helical twist (°)	-3.32
Initial particle images (no.)	2,410,032
Final particle images (no.)	40,329
Map resolution (Å)	3.0
FSC threshold	0.143
Map resolution range (Å)	200-3.0
Refinement	
Initial model used (PDB code)	De novo
Model resolution (Å) FSC threshold	0.143
Model resolution range (Å)	200-3.0
Map sharpening <i>B</i> factor (Å ²)	- 100.35
Model composition	
Non-hydrogen atoms	374
Protein residues	49
Ligands	0
<i>B</i> factors (Å ²)	
Protein	
R.m.s. deviations	
Bond lengths (Å)	0.006
Bond angles (°)	0.753
Validation	
MolProbity score	1.79
Clashscore	9.26
Ramachandran plot	
Favored (%)	95.74
Allowed (%)	4.26



Supplemental Figure 1. Recombinant protein purification. **a**, Example His-trap elution profile. CPEB3¹⁻²¹⁷ eluted in fractions 18-23 (pink subset). **b**, Full view of purification chromatogram. Three wavelengths were collected to monitor the purification: 214nm (black), 280 nm (orange), and 254 nm (light blue). **c**, Zoomed purification chromatogram (green) where the target protein started eluting. Three wavelengths were collected to monitor the purification: 214nm (black), 280 nm (orange), and 254 nm (light blue). **d**, SDS-PAGE shows harvested fractions after His-trap and HPLC purification. Freshly harvested fibrils also present.

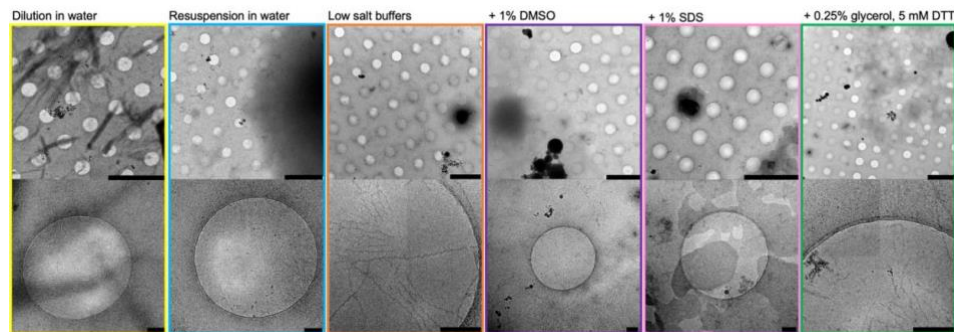


Supplemental Figure 2. Biochemical characterization of CPEB3¹⁻²¹⁷ segment. **a**, ThT curves of CPEB3¹⁻²¹⁷. Data shown are normalized triplicate conditions, n=3. **b**, Negative stain TEM images of fibrils formed by CPEB3¹⁻²¹⁷. **c**, Dissolution assays of pre-formed fibrils. Samples were harvested after formation from SF 2a, and diluted 2-fold in corresponding buffers for approximately 4 hours. Data shown n= 4 wells per condition. **d**, Negative stain TEM image of fibril sample after dilution in water for 4 hours (purple subset in SF 1d).

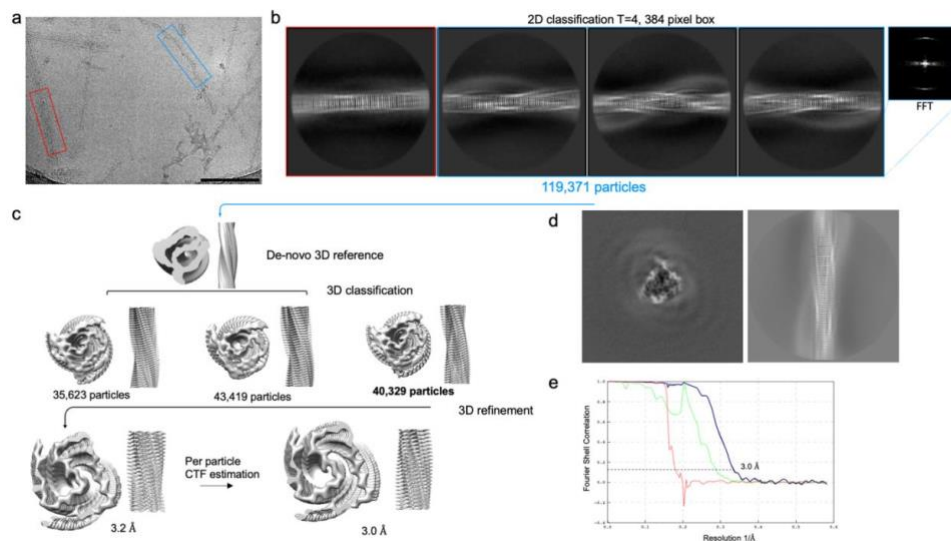


Supplemental Figure 3. Negative stain TEM micrographs of CPEB3 fibrils after dissolution assays. Samples harvested from dissolution assays in Supplemental Figure 1d were prepped for TEM analysis. **a**, Control wells were diluted in original fibrilization buffer. **b**, Representative

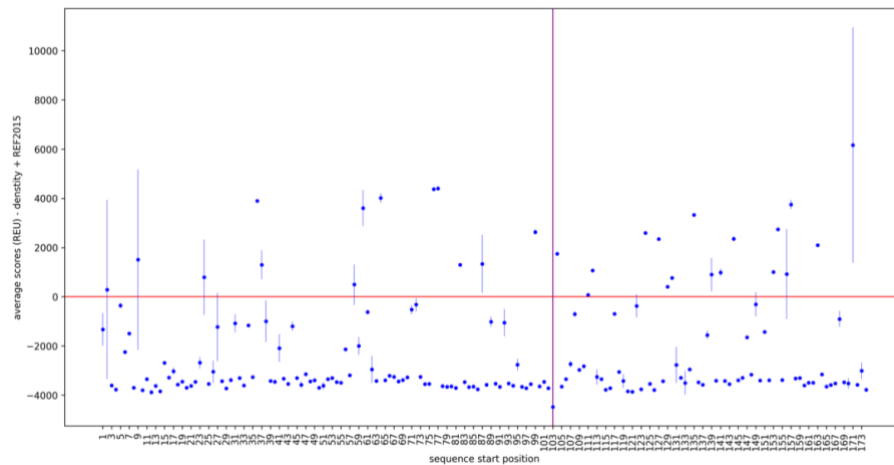
micrographs of blanks corresponding to blank, dilution buffer (left column), and protein samples after a 4-hour incubation in designated buffer (right column). Scale bar = 200 nm.



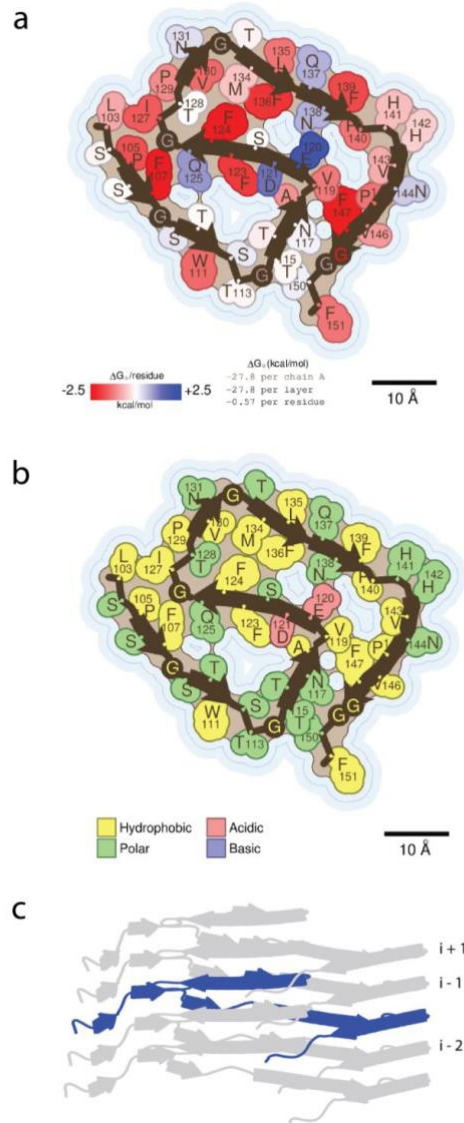
Supplemental Figure 4. CryoEM grid optimization. Reversible fibrils were plunge frozen in various buffers for optimal particle dispersion and minimum interactions at the air-water interface (AWI), representative micrographs shown. Yellow and blue: dilution and resuspension in water caused significant clumping and dissociation. Orange: low salt buffers < 100 mM NaCl cause denaturation at the air water interface. Purple and pink: addition of DMSO and SDS, respectively, reduced clumping but impacted ice and fibril integrity. Green: final freezing condition 0.25% glycerol and 5 mM DTT resulted in acceptable clumping and reduced AWI denaturation. Scale bar top row = 2 μ M, scale bar bottom row = 200 nm.



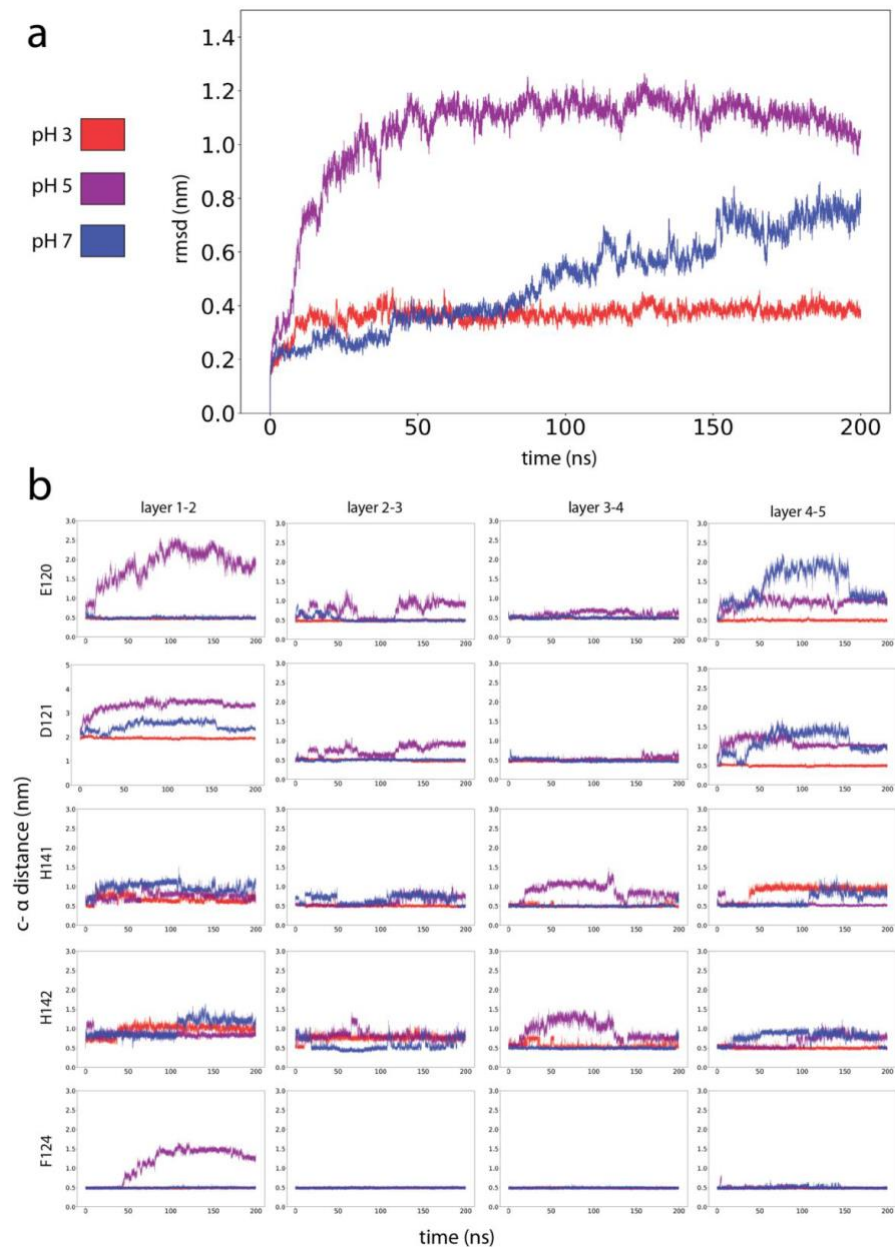
Supplemental Figure 5. CryoEM data processing and model comparisons. **a**, CryoEM micrograph of twisting fibril polymorphs (blue) and non-twisting (red) Scale bar = 50 nm. **b**, Representative 2D class averages of polymorphs pictured in SF1a (left) and 4.8 Å reflection pattern in generated Fourier transform of 2D class (right). **c**, Schematic representation of helical reconstruction of major twisting species performed in RELION. **d**, Central slice from final reconstruction (left) and projection of 3D model (right). **e**, FSC curves between independently refined half-maps masked and corrected (black), unmasked and corrected (green), masked (navy blue) and phase randomized (red). Black dotted line indicates the FSC=0.143 resolution of 3.0 Å.



Supplemental Figure 6. Average scores of threading experiment plotted against sequence start position. Average scores were calculated with Rosetta's REF2015 energy function combined with a cryo-EM density score term. Each point represents the average of three trials. Sequence start position 103 shows the most favorable sequence window and is in agreement with the modeled sequence.

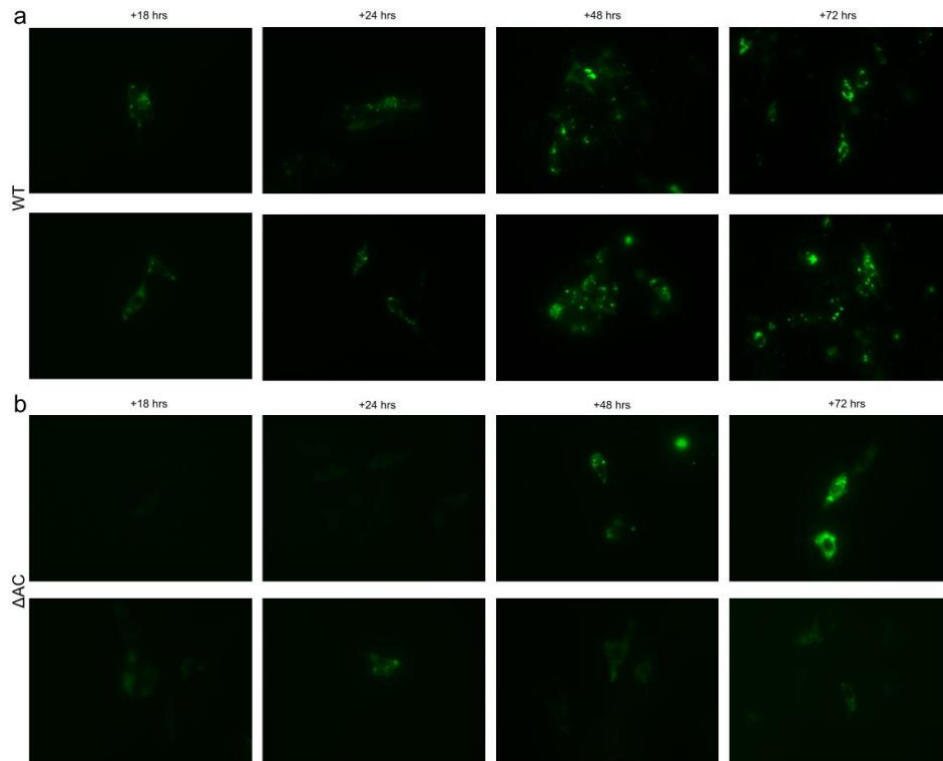


Supplemental Figure 7. Structural features of CPEB3_{AC} that may influence its stability **a**, Energy per residue stabilization, key lower left. Thin blue lines represent the closest approach that the center of a water molecule can achieve. **b**, **c**, Cartoon rendering secondary structure elements and highlighting the height differences within a single chain that interact with 3 consecutive layers.

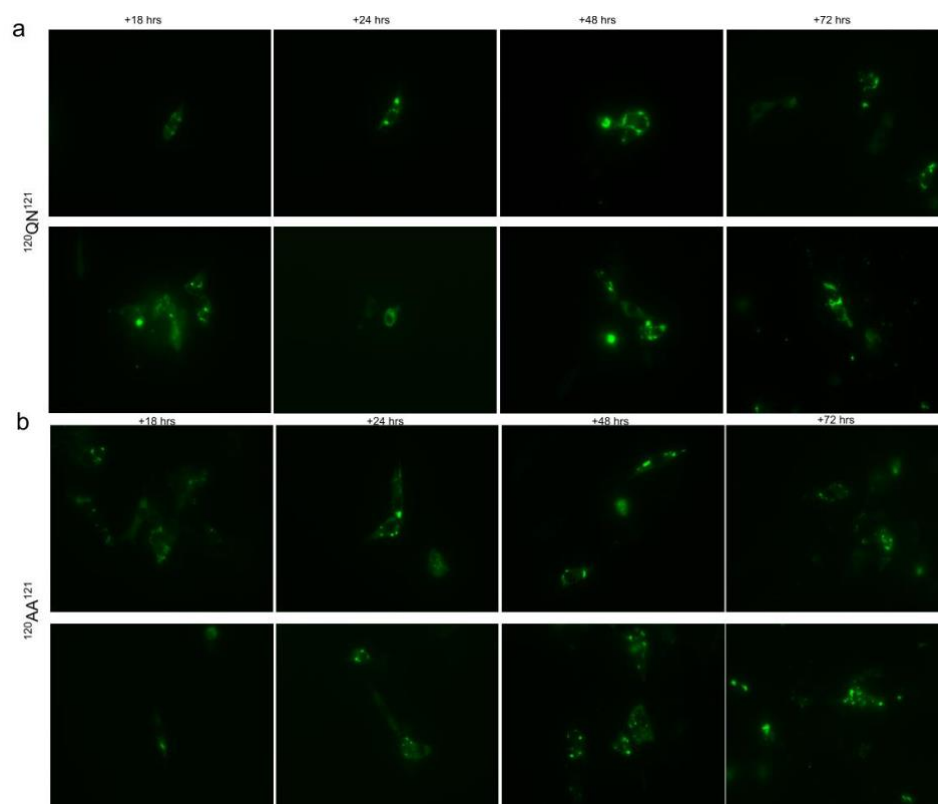


Supplemental Figure 8. Molecular dynamics simulations of pH triggered instabilities in fibril core. pH color legend where pH 3 (red), pH 5 (purple) and pH 7 (blue). **a**, Averaged RMSD for

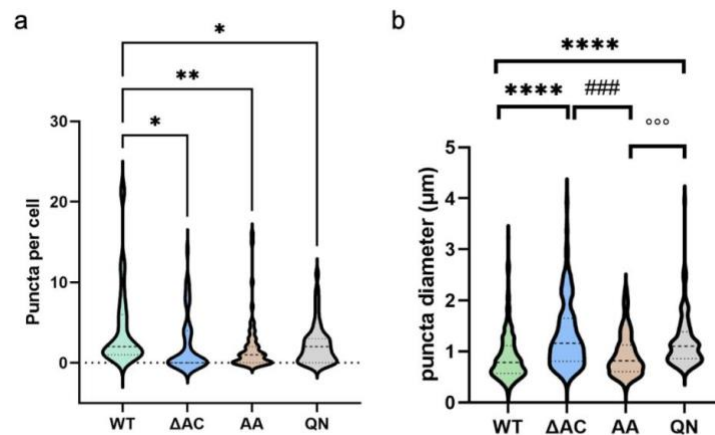
all layers across 200 ns **b**, Selected charged residues from top to bottom: E120, D121, H141 and H142. Control residue F124. From left to right: neighboring layer c- α distances across 200 ns simulations.



Supplemental Figure 9. Aggregation phenotypes of CPEB3 WT and Δ AC constructs in HT22 in GFP channel (green). **a**, Representative images of WT CPEB3-GFP expression across four time points 18, 24, 48 and, 72 hrs. **b**, Representative images of CPEB3 Δ AC-GFP expression across four time points 18, 24, 48 and, 72 hrs with earlier time points displaying significantly diffuse phenotypes. All time points indicate time after incubation with lipofectamine and corresponding vector.



Supplemental Figure 10. Irregular aggregation phenotypes of CPEB3 mutant constructs in HT22 in GFP channel (green). **a**, Representative images of CPEB3^{120QN121}-GFP across four time points 18, 24, 48 and, 72 hrs. **b**, Representative images of CPEB3^{120AA121}-GFP expression across four time points 18, 24, 48 and, 72 hrs. As in Supplemental Figure 4, all time points indicate time after incubation with lipofectamine and corresponding vector.



Supplemental Figure 11. CPEB3 mutant phenotypes. **a**, Quantification of puncta per cell after 24 hours of exogenous expression. Where WT-GFP n=60 cells, CPEB3^{ΔAC}-GFP n=52 cells, CPEB3^{120QN121}-GFP n=56 cells and CPEB3^{120AA121}-GFP n=50 cells. One-way ANOVA, followed by Tukey's multiple comparisons post hoc test, * p<0.05, ** p<0.01, vs WT. **b**, Quantification of average puncta diameter from cells surveyed in **a**. One-way ANOVA, followed by Tukey's multiple comparisons post hoc test, **** p<0.0001, vs WT, ### p<0.001, vs ΔAC, °°° p<0.001 vs QN.

References

1. Prusiner, S. B. Novel Proteinaceous Infectious Particles Cause Scrapie. *Science* (80-). **216**, 136–144 (1982).
2. Shorter, J. & Lindquist, S. Prions as adaptive conduits of memory and inheritance. *Nature Reviews Genetics* vol. 6 435–450 (2005).
3. Eisenberg, D. & Jucker, M. The amyloid state of proteins in human diseases. *Cell* **148**, 1188–1203 (2012).
4. Rayman, J. B. & Kandel, E. R. Functional Prions in the Brain. *Cold Spring Harb. Perspect. Biol.* **9**, (2017).
5. McGaugh, J. L. Memory - A century of consolidation. *Science* (80-). **287**, 248–251 (2000).
6. Majumdar, A. *et al.* Critical role of amyloid-like oligomers of Drosophila Orb2 in the persistence of memory. *Cell* **148**, 515–529 (2012).
7. Stephan, J. S. *et al.* The CPEB3 Protein Is a Functional Prion that Interacts with the Actin Cytoskeleton. *Cell Rep.* **11**, 1772–1785 (2015).
8. Si, K., Lindquist, S. & Kandel, E. R. A Neuronal Isoform of the Aplysia CPEB Has Prion-Like Properties. *Cell* **115**, 879–891 (2003).
9. Khan, M. R. *et al.* Amyloidogenic Oligomerization Transforms Drosophila Orb2 from a Translation Repressor to an Activator. *Cell* **163**, 1468–1483 (2015).
10. Keleman, K., Krüttner, S., Alenius, M. & Dickson, B. J. Function of the Drosophila CPEB protein Orb2 in long-term courtship memory. (2007) doi:10.1038/n1996.
11. Fioriti, L. *et al.* The Persistence of Hippocampal-Based Memory Requires Protein

- Synthesis Mediated by the Prion-like Protein CPEB3. *Neuron* **86**, 1433–1448 (2015).
12. Pavlopoulos, E. *et al.* Neuralized1 activates CPEB3: A function for nonproteolytic ubiquitin in synaptic plasticity and memory storage. *Cell* **147**, 1369–1383 (2011).
13. Ford, L., Ling, E., Kandel, E. R. & Fioriti, L. CPEB3 inhibits translation of mRNA targets by localizing them to P bodies. **116**, 18078–18087 (2019).
14. Drisaldi, B. *et al.* SUMOylation Is an Inhibitory Constraint that Regulates the Prion-like Aggregation and Activity of CPEB3. *Cell Rep.* **11**, 1694–1702 (2015).
15. Hervás, R. *et al.* Cryo-EM structure of a neuronal functional amyloid implicated in memory persistence in *Drosophila*. *Science* (80-.). **367**, 1230–1234 (2020).
16. Fiumara, F., Fioriti, L., Kandel, E. R. & Hendrickson, W. A. Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell* **143**, 1121–1135 (2010).
17. Reselammal, D. S. *et al.* Mapping the Fibril Core of the Prion Subdomain of the Mammalian CPEB3 that is Involved in Long Term Memory Retention. *J. Mol. Biol.* **433**, (2021).
18. Sawaya, M. R., Hughes, M. P., Rodriguez, J. A., Riek, R. & Eisenberg, D. S. The expanding amyloid family: Structure, stability, function, and pathogenesis. *Cell* **184**, 4857–4873 (2021).
19. Murray, D. T. *et al.* Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. *Cell* **171**, 615–627.e16 (2017).
20. Wasmer, C. *et al.* Amyloid fibrils of the HET-s(218-289) prion form a β solenoid with a triangular hydrophobic core. *Science* (80-.). **319**, 1523–1526 (2008).
21. Lu, J. *et al.* CryoEM structure of the low-complexity domain of hnRNPA2 and its conversion to pathogenic amyloid. doi:10.1038/s41467-020-17905-y.
22. McGlinchey, R. P. & Lee, J. C. Reversing the Amyloid Trend: Mechanism of Fibril Assembly and Dissolution of the Repeat Domain from a Human Functional Amyloid. *Isr. J. Chem.* **57**, 613–621 (2017).
23. Maji, S. K. *et al.* Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science* (80-.). **325**, 328–332 (2009).
24. Richter, J. D. CPEB: a life in translation. *Trends Biochem. Sci.* **32**, 279–285 (2007).
25. Kozlov, E., Shidlovskii, Y. V., Gilmutdinov, R., Schedl, P. & Zhukova, M. The role of CPEB family proteins in the nervous system function in the norm and pathology. *Cell Biosci* **11**, 64 (2021).
26. Raveendra, B. L. *et al.* Characterization of prion-like conformational changes of the neuronal isoform of Aplysia CPEB. *Nat. Struct. Mol. Biol.* **20**, (2013).
27. Sinning, A. *et al.* Minireview: pH and synaptic transmission. (2013) doi:10.1016/j.febslet.2013.04.045.
28. Boeynaems, S. *et al.* Protein Phase Separation: A New Phase in Cell Biology. *Trends Cell Biol.* **28**, 420–435 (2018).
29. Cao, Q., Boyer, D. R., Sawaya, M. R., Ge, P. & Eisenberg, D. S. Cryo-EM structures of four polymorphic TDP-43 amyloid cores. *Nat. Struct. Mol. Biol.* doi:10.1038/s41594-019-0248-4.
30. Arseni, D. *et al.* Structure of pathological TDP-43 filaments from ALS with FTL. *Nature* **601**, 139–143 (2022).
31. Glynn, C. *et al.* Cryo-EM structure of a human prion fibril with a hydrophobic, protease-resistant core. *Nat. Struct. Mol. Biol.* **27**, 417 (2020).

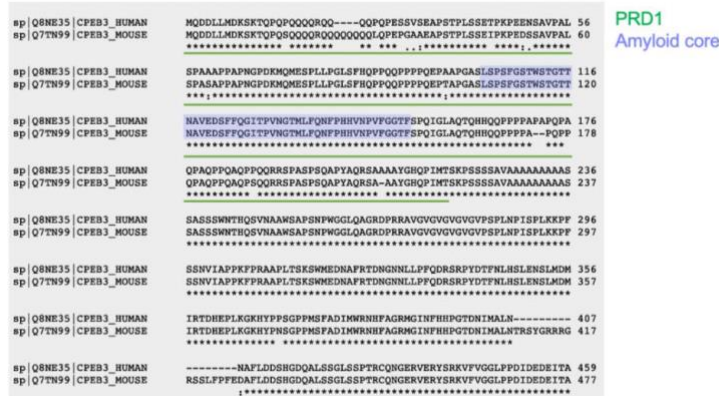
32. Manka, S. W. *et al.* 2.7 Å cryo-EM structure of ex vivo RML prion fibrils. doi:10.1101/2021.12.13.472424.
33. Kraus, A. *et al.* High-resolution structure and strain comparison of infectious mammalian prions. *Mol. Cell* **81**, 4540–4551.e6 (2021).
34. Uptain, S. M. & Lindquist, S. Prions as protein-based genetic elements. *Annu. Rev. Microbiol.* **56**, 703–741 (2002).
35. Mompeán, M. *et al.* The Structure of the Necrosome RIPK1-RIPK3 Core, a Human Hetero-Amyloid Signaling Complex. *Cell* **173**, 1244–1253.e10 (2018).
36. Scheres, S. H. W. Amyloid structure determination in RELION-3.1. *bioRxiv* 1–9 (2019) doi:10.1101/823310.
37. Wagner, T. *et al.* SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun. Biol.* **2**, 1–13 (2019).
38. Fitzpatrick, A. W. P. *et al.* Cryo-EM structures of tau filaments from Alzheimer’s disease. *Nat. 2017 5477662* **547**, 185–190 (2017).
39. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Biological Crystallography Features and development of Coot. doi:10.1107/S0907444910007493.
40. Afonine, P. V *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Res. Pap. Acta Cryst* **74**, 531–544 (2018).
41. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
42. Alford, R. F. *et al.* The Rosetta all-atom energy function for macromolecular modeling and design HHS Public Access. *J Chem Theory Comput* **13**, 3031–3048 (2017).
43. Eisenberg, David, M. D. A. Solvation energy in protein folding and binding. *Nature* **319**, 199–203 (1986).
44. Koehl, P. & Delarue, M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249–275 (1994).
45. Van Der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).
46. Vanommeslaeghe, K. *et al.* CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671 (2010).
47. Cormier, M. J. & Kazuo, H. Studies on the bioluminescence of Renilla reniformis. *Biochim. Biophys. Acta - Spec. Sect. Biophys. Subj.* **88**, 99–104 (1964).
48. Jaafari, N., Konopacki, F. A., Owen, T. F., Kantamneni, S. & Rubin, P. SUMOylation Is Required for Glycine-Induced Increases in AMPA Receptor Surface Expression (ChemLTP) in Hippocampal Neurons. *PLoS One* **8**, 52345 (2013).

Materials and Methods

Construct design

Design of recombinant CPEB3 segment included previously reported residues critical for aggregate formation and prion-like behavior in cells⁷. CPEB3 residues 1-217 were inserted into

pET—24(+) vector from TWIST biosciences. Sequences were verified via DNA sequencing (Genewiz). CPEB3-EGFP constructs were a kind gift from Luana Fioriti.



Panel 1: Sequence alignment of CPEB3 human and mouse sequences. Green light denotes residues within PRD1 and lilac highlight corresponds to residues placed in amyloid core structure.

Protein expression and purification

Recombinant CPEB3¹⁻¹⁷ was expressed in *Escherichia coli* BL21 GOLD cells. Briefly, cells were grown in LB media with 50 ug/ml kanamycin at 37° C to an OD600 of 0.6-0.8. 1 mM isopropyl B-D-1-thiogalactopyranoside (IPTG) was added for protein expression and cells were cultured an additional 6 hrs at 37° C. Cells were harvested and stored at -20° C until purification. Cell lysates were resuspended in lysis buffer containing 10 mM Tris, 1 mM EDTA at pH 8.2 and supplemented with 1% halt protease inhibitor cocktail (Thermo Scientific), 1% homemade Benzonase and subsequently lysed using 6 consecutive freeze-thaw cycles in liquid nitrogen. Cell lysates were centrifuged (12,000gs for 30min) and pellet was resuspended in 6M Gdn-HCl, 50mM Tris-Base, 25mM imidazole at pH 8 and left gently shaking on ice overnight. Lysates were centrifuged (3,000gs for 30 min) and supernatant was filtered and loaded onto an equilibrated HisTrap HP column (GE) using NGC (Biorad). The HisTrap was pre-equilibrated with 6M Gdn-HCl, 50mM Tris-Base and 25 mM imidazole. Protein was eluted with gradient protocol containing 6M Gdn-HCl, 50mM Tris-Base and 500 mM imidazole. Eluted protein was flash frozen and stored at -80° C until further purification. Further purification was executed with RP-HPLC, using an Interchim puriFlash® 4125 Preparative Liquid Chromatography System equipped with an Interchim PT-15C8-F0040 cartridge (C8,15 µm particle size). The eluted protein fractions were defrosted from -80°C and pooled. The pooled sample was then diluted to 1.8 mL with a dilution buffer consisting of 6 M Gdn-HCl, 50 µM NaH₂PO₄, 10 µM Tris, and 20 µM imidazole. Before injection the 2-mL sample loop was washed with the above dilution buffer to maintain high concentrations of guanidine to minimize protein aggregation. [A] consisted of a 99 to 1 water to acetonitrile solution with 15 mM NH₄OH added as a buffer. [B] consisted of a 40 to 60 water to acetonitrile solution with 15 mM NH₄OH. Flow rate was set at 40 mL/min. Before injection, the system was equilibrated to 100% [A]. The protein sample was loaded onto the loop and injected. Three wavelengths were used to monitor the purification: 214nm (black), 280 nm (orange), and 254 nm (light blue). The gradient was held at 100% [A] for 5 minutes then a gradual increase to 100% [B] was executed over a span of 25 minutes (Δ5% per minute). The target protein started eluting at 16 minutes.

Fibril preparation and negative stain TEM

Lyophilized protein powder was resuspended in H₂O at 1.2 mg/mL concentration, aliquoted and immediately flash frozen and stored at -20°C until further use. Various buffers containing urea and guanidine were initially screened for fibril formation to attempt to control clumping and aggregation. These aggregates were heterogenous and difficult to reproduce. 96-well screen ranging in salt concentrations and pH were used for subsequent attempts without denaturants. The optimized fibril growth condition was incubation in 125mM NaCl, 50mM Tris-Base, 10mM K₂PO₄ and 5mM glutamic acid ~pH 4.5 and left shaking overnight on the acoustic shaker. Fibrils were collected by scraping and pooling solution from wells and immediately prepared for subsequent experiments. Negative-stain transmission EM samples were prepped by applying 3 µl of fibril solution to glow-discharged 300 mesh carbon-coated formvar support films mounted on Cu grids (Ted Pella). The samples were wicked, washed briefly with H₂O, stained with 2% uranyl acetate for 90 seconds and wicked again, allowing to air dry for 3 minutes. Grids were imaged on a T12 (FEI) electron microscope.

Cryo-EM sample preparation and data collection

Freshly harvested fibril solution was diluted 5 fold with buffer containing 100mM NaCl, 10mM Tris-Base, 5mM DTT and 0.25% glycerol. 1.6 µL of diluted fibril solution was applied to each side of a glow-discharged Ultrathin carbon on Quantifoil 1.2/1.3 300 mesh on Au grid (Ted Pella) and plunge frozen into liquid ethane using a Vitrobot Mark IV (FEI) set at 100% humidity, 4°C with a blot force of -1 and blot time of 1.5 seconds. Data were collected on a Titan Krios G3i microscope with a K3 direct detection camera (Gatan) operated at an accelerating voltage of 300 keV and a 20eV slit width (BioQuantum). Automated data acquisition were performed using EPU 2.8 software (Thermo Scientific). Images were acquired with a nominal magnified pixel size of 0.86 Å/pixel, a total of 40 frames with a dose of 1.3 electrons per Å² per frame resulting in a total dose of 52 electrons per Å².

Data processing and helical reconstruction

CTF estimation were performed with CTFFIND 4.1.8. Drift-correction and dose-weighting were performed using MotionCorr2 implemented in RELION 3.1³⁶. Particles were picked with the automatic particle picking software crYOLO³⁷. All subsequent image processing and helical reconstruction were performed in RELION as previously described.^{29,38} Briefly, fibril segments were initially extracted with a 90% overlap into 512 pixel boxes binned by 2 (1.72 Å/pix) and subjected to reference-free 2D classification using a T=2 regularization parameter. Segments belonging to suboptimal 2D averages were discarded and homogenous subsets were selected for further processing. Upon identification of a rapidly twisting species, a smaller box size was chosen. All segments were then extracted with a 90% overlap into 384 pixel boxes for a total of 2,410,032 segments. Segments were split into 10 groups and several rounds of reference free 2D classification were performed to weed out suboptimal segments. Homogenous subsets from each were regrouped and reference free 2D classification were iteratively run decreasing psi_step and offset_step from 8 to 1. After the majority of suboptimal particles had been discarded an additional round of classification were performed with the tau regularization parameter set to T= 8. Only segments contributing to averages showing clear 4.8 Å signal in corresponding 2D FFTs were selected for subsequent 3D processing, resulting in 119,372 segments. Using both the 512 pixel box and 384 pixel box an initial helical pitch of 536 Å and helical twist of -3.22° were estimated. Starting with these estimates a 3D reference was reconstructed de-novo with 384 pixel 2D averages showing clear β-strand separation along the helical axis. 3D manual refinements were performed using a 30 Å low-pass filtered initial model, K=3 and with manual control of tau_fudge and healpix to reach a resolution of ~ 5 Å.

Segments contributing to a homogeneous class as defined by stable helicity and separation of β -strands resulted in a final subset of 40,329 segments. Again, a manual 3D refinement was performed on selected segments leading to a reconstruction of 3.5 Å. Several 3D auto-refinements using a 7 Å low-pass filtered map from the final manual refinement were run with optimization of helical parameters leading to an estimated helical rise of 4.91 Å and a helical twist of -3.32° . This map was then used for per-particle CTF estimation and a 3D auto-refinement was repeated. The nominal pixel size was adjusted from 0.86 Å to 0.879 Å, resulting in an expected helical rise of 4.8 Å. Post-processing of the map with an extended initial mask of 3 pixels and soft edge of 10 pixels was followed by clipping and centering the map using clip.com. Auto-sharpening using phenix.auto_sharpen at the resolution cutoff indicated by the half-map GSFSC led to a final overall resolution of 3.0 Å. The atomic model was built into the refined map using COOT³⁹. We performed automated structure refinement using phenix.real_space_refine (Phenix 1.2)⁴⁰.

Threading

Threading was performed with a custom python script utilizing the PyRosetta⁴¹ software package and a sliding window approach. For each window, the experimentally expressed sequence was threaded onto a poly-alanine backbone then energy minimized with the PyRosetta FastRelax with cryo-EM density function. FastRelax used the Ref2015⁴² energy function with a modified fa_elec weight to 1.5 and elec_dens_fast weight of 25. Each sequence window was energy minimized in triplicate and the resulting scores were averaged. Symmetry was applied to pose objects to increase speed of computation.

Energetic calculations

Solvation free energy calculations were performed as recently described¹⁸. Briefly, the solvent accessible surface area (SASA) for each atom in a residue was determined (folded state). Next, the SASA for each atom of a residue was determined in the absence of the other residues (reference state), and the difference between both states was calculated ($SASA_{Ref} - SASA_{Fold}$). This value was then multiplied by the Atomic Solvation Parameter (ASP) specific to each atom, as determined previously by Eisenberg et al⁴³. An entropic term is also included to take into consideration the degrees of freedom lost in going from a disordered to ordered state⁴⁴. The energies of all atoms were then summed to generate the solvation energy for each structure. Difference energy maps were generated by subtracting solvation free energies pairwise for each atom in the two structures being compared.

Molecular dynamics simulations

Molecular dynamics simulations were performed with the GROMACS⁴⁵ software package (version 2022.2) using the CHARMM27⁴⁶ all-atom forcefield. Simulations were carried out with a 5-layer CPEB fibril under varying charge states to simulate fibril dynamics at pHs 3, 5, and 7. In each case, the fibril was placed in a cubic box, solvated with SPC/E three-point water molecules and added counter ions. The system was energy minimized, then temperature and pressure equilibrated for 100 ps. Final simulations were executed for 200 ns.

Thioflavin T-binding and dissolution assays

Protein was resuspended in fibril formation buffer as described previous at 25 μ M concentration with equal parts ThT. Solution was pipetted into a 96-well plate with optical bottom (Sigma-Aldrich) and incubated at 37°C with continuous shaking. ThT fluorescence was measure with an

excitation filter of 440 nm and emission filter of 480nm using a Varioskan plate reader (ThermoFisher). Aggregation curves were generated from triplicate, n=3 wells. Dissolution assays were started after 18 hours of fibril formation. Because fibrils were not able to be resuspended in different buffers due to their liability, a 96-well plate was prepared by diluting each fibril condition 3-fold in corresponding buffers. Normalization of ThT intensity was performed to account for variations across experiments and reduction of signal when diluting samples for dissolution assays.

Cell culture

HT22 hippocampal neurons and U87 glioblastomas were cultured in Dulbecco's modified Eagle's medium with 10% Fetal Bovine Serum (ThermoFisher) and 1% penicillin/streptomycin (ThermoFisher) in a tissue-culture incubator at 37°C and 5% CO₂. Cells were cultured as described above and plated on 12-mm glass coverslips coated with poly-L-lysine (Sigma) and Lamin 50 uG (Sigma) in 6-well cell-culture plates. Cells were plated at 70% confluency overnight and transfected using Lipofectamine 3000 kit (ThermoFisher) for an 8 hr incubation. Cells were harvested and/or imaged at 24 hours post transfection for all experiments. Live fluorescent images were acquired on an EVOS M7000 (ThermoFisher) on a 40X objective.

Immunofluorescence and puncta analysis

Cells were fixed in 2% paraformaldehyde at 4°C for 10 minutes and permeabilized using 0.5% PBS-Triton X-100 for 10 minutes. Cells were subsequently washed with PBS-T and blocked with 2% BSA, 0.5%FSG, and PBS-T solution for 1hr at RT. Cells were incubated with 1° antibodies diluted in blocking solution for approximately 2 hours. Cells were washed with PBS-T three times before beginning incubation with 2° antibodies and washed with PBS-T. Coverslips were mounted using Prolong Glass Antifade Mountant (ThermoFisher) and allowed to dry for 72 hours. Fluorescent images were acquired on a Leica confocal SP8-STED/FLIM/FCS on a 60X water immersion objective. Puncta were manually counted and distances were manually calculated by averaging X and Y for each puncta in GFP channels. Image analyses were performed in FIJI (ImageJ).

Luciferase assays

HEK293 cells were plated at 50% confluency overnight and transfected using Lipofectamine 3000 kit (ThermoFisher) with ~0.5 µg of CPEB3 DNA, 0.5 µg of Renilla luciferase appended with GluA2 3' UTR or SUMO2 3'UTR, and 0.5 µg of Firefly luciferase⁴⁷. 18/24 hr after transfection, cells were lysed with 100 µl of buffer for dual luciferase assay (Promega). 20 µl of the lysates were used for the quantification of Firefly and Renilla Luciferase activity using a Luminometer (GloMax, Promega).

Hippocampal neurons cultured for 9–10 days in Neurobasal medium with B27 supplement (Invitrogen) at a cell density of 30 000–40 000/cm² were cotransfected (Lipofectamine LTX, LifeScience) for 3 h with ~0.5 µg of CPEB3 DNA, 0.5 µg of Renilla luciferase appended with GluA2 3' UTR, and 0.5 µg of Firefly luciferase. We used glycine to stimulate the NMDA receptors selectively, a protocol that mimics LTP induction in cultured neurons. Chem-LTP was induced as described previously⁴⁸. Briefly, neuronal cultures were transferred from Neurobasalgrowth medium to extracellular solution (ECS) containing: 150 mM NaCl, 2 mM CaCl₂, 5 mM KCl, 10 mM HEPES(pH 7.4), 30 mM glucose, 0.5mM TTX, 20mM bicuculline methiodide. The transfected neurons were stimulated with 200µM glycine for 3 min and then washed for 30 minutes before lysis in 100 µl of buffer for dual luciferase assay (Promega). 20 µl of the lysates were used for the quantification of Firefly and Renilla Luciferase activity using a

Luminometer (GloMax, Promega). The ratio between Renilla and Firefly is calculated for each experimental condition.

Statistical analyses

Number of repeats and (Avg) $\pm$ SE are stated in figure legends. Statistical significance was measured by one or two-way ANOVA followed by the Tukey's multiple comparisons test.

Acknowledgments:

We are grateful to Wong Hoi Hui of the Electron Imaging Center for NanoMachines (EICN) for help and advice during cryoEM screening and sample preparation. We thank Peng Ge (UCLA) for helpful discussions during cryoEM data processing, assistance during data acquisition and Duilio Cascio (UCLA) for assistance during model building and refinement. We thank Corey Hecksel (SLAC) for assistance during data acquisition. Part of this work was performed at the Stanford-SLAC Cryo-EM Center (S2C2), which is supported by the National Institutes of Health Common Fund Transformative High-Resolution Cryo-Electron Microscopy program (U24GM1295410). M.D.F is supported by Eugene V. Cota-Robles Fellowship, Ruth L. Kirschstein NRSA GM007185, Whitcome Pre-Doctoral Fellowship, Audree V. Fowler Fellowship and a National Science Foundation Graduate Research Fellowship. J.A.R. is supported as a Pew Scholar, a Beckman Young Investigator and a Packard Fellow. This work was performed as part of STROBE, an NSF Science and Technology Center through Grant DMR-1548924, and by DOE Grant DE-FC02-02ER63421. The work was also supported by NIH-NIGMS Grant R35 GM128867.

Author Contributions:

M.D.F designed experiments, purified recombinant constructs, screened negative stain samples, performed ThT assays, prepared cryoEM samples, performed cryoEM data collection and processing helical reconstruction, purified vectors for mammalian expression, cultured mammalian cells, performed live fluorescent imaging, immunofluorescent experiments and performed data analysis. M.R.S performed solvation energy calculations and assisted in model building and refinement. S.Z. performed computational threading and MD simulations. D.R.B assisted in cryoEM data processing helical reconstruction. S.T and A.S performed puncta quantification. T.Z. and J.C. assisted in recombinant protein expression and purification. L.F. cultured primary neurons, performed luciferase assays and analysis and guided the project. M.D.F and J.A.R wrote the manuscript with input from all authors. J.A.R supervised and guided the project.

Competing Financial Interests:

JAR is an equity stake holder of MedStruc Inc.