Structural and Thermodynamic Classification of Amyloid Polymorphs

Jack P. Connor[1], Sheena E. Radford[1*], and David J. Brockwell[1,2*]

[1]Astbury Centre for Structural Molecular Biology, School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK

[2]Lead contact

[*]Correspondence: s.e.radford@leeds.ac.uk; d.j.brockwell@leeds.ac.uk

SUMMARY

Over 500 amyloid structures have been solved to date to near-atomic resolution. This has highlighted an enormous diversity of fibril structures conforming to the canonical cross-β amyloid fold. Using α-synuclein and tau amyloid structures as models, we show that they can be hierarchically clustered into topologically distinct fold families. Despite their different topologies, fibrils display remarkably similar energy profiles, as determined by FoldX, with the same regions providing stability among different polymorphs. We found that the regions that stabilise the amyloid core pair in different ways to generate distinct topologies. The results provide a framework to classify newly solved fibril structures as belonging to an existing class or forming a new topological cross-β fold. Furthermore, the analysis facilitates comparisons between fibrils found in disease and those formed *in vitro*, including their nearest structural neighbours. The workflow has been automated, enabling users to interrogate new amyloid structures as they emerge.

KEYWORDS

Amyloid, α-Synuclein, Tau, Polymorphism, Thermodynamics, Hierarchical Cluster

INTRODUCTION

Amyloid fibrils are supra-molecular structures comprising stacked monomers folded into β-strands that are organised in a cross-β array[1]. Amyloid fibrils can comprise a single filament or, more commonly, two or more protofilaments, stabilized by interlocking sidechains at the protofilament interface(s)[2]. These supra-molecular structures are of relevance to disease, with aberrant accumulation of amyloid deposits in the brain being a hallmark of neurodegeneration[3], and localized or systemic deposition of amyloid in the viscera associated with diseases affecting the heart, kidney, pancreas and other regions[4]. In other cases, amyloid may be functional[2]. The protein involved in each amyloid disease is different. In general, amyloid deposition of amyloid-β (Aβ)[5,6] and tau[7] are found in Alzheimer's disease, tau is involved in other tauopathies[1], and α-synuclein[8] in Parkinson's disease and other synucleinopathies.

The first atomic resolution amyloid structure of the 140-residue protein α-synuclein was solved in 2016 using solid-state NMR on fibrils formed *in vitro*[9]. Two years later, the first α-synuclein amyloid structures, again generated *in vitro*, were solved using cryo-electron microscopy (cryo-EM)[10,11]. Recent advances in cryo-EM have increased the rate of amyloid structure elucidation, with the Protein Data Bank now containing 506 amyloid structures (as of September 2024)[2,12]. Of these, a striking 101 are cryo-EM-derived α-synuclein amyloid structures assembled under various solution conditions and/or containing different mutations (Figure S1)[13], while 68 cryo-EM structures have been deposited of amyloid formed from tau variants (4R and 3R+4R) (Figure S2)[13]. These include structures of α-synuclein amyloid from individuals with Parkinson's disease (PD), Dementia with Lewy Bodies (DLB)[14], Multiple System Atrophy (MSA)[15] or Juvenille Onset Synucleinopathy (JOS)[16], as well as tau amyloid purified from the brains of donors with Alzheimer's disease or other tauopathies[17–22].

Solving multiple structures of amyloid fibrils formed from the same protein is vital to understand how and why a polypeptide sequence can adopt different amyloid structures, a phenomenon known as polymorphism[2]. Improved understanding of which factors govern the formation of the adopted amyloid fold, and the potential differing biological impact of each structure, is of clinical relevance as amyloids solved from *ex vivo* samples can display disease-specific polymorphism[3]. With the growing body of solved amyloid structures and the scientific and clinical interest in understanding polymorphism, a computational approach is required to quantify the degree of variation among amyloid folds.

Here, inspired by the classification of globular protein folds into different families and hierarchies, using CATH[23,24] and SCOP[25], we set out to classify the 101 α-synuclein amyloid

structures and 68 4R and 3R+4R tau cryo-EM amyloid structures into different hierarchical classes. From this, we determined the number of different structural classes that can be formed from the same, or similar, protein sequence. These proteins were chosen for our analysis since together they comprise >33% of all amyloid structures deposited to date.

Structural classification was achieved by hierarchical clustering of the root-mean-squared deviation (RMSD) between aligned protofilament structures. This revealed eleven distinct classes for α-synuclein and eleven for tau amyloids. We then investigated how the stability of the amyloid fibrils in the different classes compare, to explore how thermodynamic stability influences polymorphism. This builds on a previous study that analysed 107 amyloid structures from a wide range of pathologies[26]. Consistent with the previous example, we show that amyloid polymorphism arises from different pairing of the same stabilizing regions in different α-synuclein/tau amyloid, resulting in fibrils that are structurally distinct, but thermodynamically approximately isostable.

Our results highlight the importance of kinetic selection in determining the pairing of stabilizing regions during amyloid formation and hence the selection of which amyloid fold is formed. They also rationalize the sensitivity of the amyloid structures to the solution conditions employed. In addition, our analysis pipeline enables the rapid comparison of newly solved structures to the expanding database of pre-existing amyloid folds for a given protein of interest, enabling comparison of new structures to previously determined amyloid folds.

RESULTS

Analysis Pipeline

To characterize the extent of polymorphism in tau and α-synuclein amyloid fibrils, a comprehensive list of the published structures to date was obtained from the Amyloid Atlas[2,13]. This excellent resource collates amyloid structures solved to near atomic resolution. Given that the number of solved amyloid structures is increasing rapidly, our analysis pipeline is designed to scrape a list of PDB codes matching the desired protein of interest from the Amyloid Atlas before running the fully automated downstream analysis. This enables the pipeline to run in a high-throughput manner and can be easily repeated as new structures are published and become available online.

A graphical summary of the analysis pipeline is shown in Figure 1. Briefly, PDB codes are obtained from Amyloid Atlas, the number of protofilaments in each structure is counted, and the number of unique folds within each PDB structure is determined. If the different protofilament monomers in a single PDB file adopt an identical structure (defined by their radius of gyration (R_g)), it is classed as having a single distinct protofilament fold. As a single fold can represent all of the chains in the given PDB, a single chain is taken for subsequent analysis. If the different protofilaments adopt distinct folds, a single chain from each distinct protofilament is taken for further analysis.

To identify distinct amyloid folds, structures are aligned and Cα-Cα RMSD values calculated for atoms in shared regions of the sequence (Methods). These RMSD values are used in Euclidean distance hierarchical clustering to group PDBs by their amyloid fold similarity (Methods). To calculate stability, the thermodynamic contribution of each residue ($\Delta G°$ per residue) is calculated using FoldX[27]. We exemplify this pipeline below with analysis of the 101 and 68 cryo-EM amyloid structures of α-synuclein and tau (4R or 3R+4R), respectively (as of 16th September 2024), but the pipeline can readily be applied to any amyloid protein for which there are sufficient deposited structures. We chose to classify only full length 4R or 3R+4R tau structures as our analysis requires comparisons between equivalent Cα atoms which can be difficult when using fragments or 3R isoform structures.

Quality Control

FoldX is more accurate at assigning free energy values for high resolution crystal structures compared to solution structures obtained using nuclear magnetic resonance (NMR). This may be due to the propensity for NMR structures to adopt improbable backbone dihedral angles and/or over-training of FoldX parameters on crystallographic structures[26,27]. Therefore, the two

ssNMR α-synuclein structures (2n0a[9] and 8fpt[28]) were removed from our analysis (details of all PDB entries used are given in Tables S1 (α-synuclein) and S2 (3R+4R and 4R tau)).

For the remaining 101 α-synuclein and 68 tau cryo-EM amyloid structures, we wanted to ensure only well-resolved, high-resolution structures were included in our analysis. Using the published Q-scores, a measurement of the resolvability of individual atoms in cryo-EM maps[29], structures with an overall low resolution (Q-scores <0.39 for α-synuclein and <0.43 for tau) were removed (Figures S3A and S4A). Six α-synuclein structures (7nci[30], 8gf7[31], 7ynl[32], 8cyr[33], 7nch[30] and 7ncj[30] with a Q-score range of 0.39 to 0.18) and five tau amyloid structures (7mkg[34], 7mkh[34], 7u0z[35], 7upg[36] and 5o3t[17] with a Q-score range of 0.41 to 0.40) were removed. For the remaining highly resolved structures, single residues/regions with poor resolvability were removed from subsequent analyses (Figures S3B-C and S4B-C). The tau amyloid structure 6tjo[22] was removed and it is not shown in Figure S4A as it had no published Q-score data. From these considerations, 95 α-synuclein and 62 tau amyloid structures were selected for downstream analysis.

To validate the use of FoldX to calculate the per residue stability of α-synuclein and tau fibrils, we compared these values to the solvation free energy, calculated as described by Eisenberg and colleagues[37,38] (Methods). Comparing the mean FoldX and the mean solvation free energy scores for each residue across all 95 unique α-synuclein and 62 tau structures showed significant agreement (Pearson correlation coefficients of 0.77 and 0.79, respectively (Figures S5 and S6)). Hence, FoldX was used in subsequent analyses.

Polymorphism

Before the degree of structural polymorphism can be characterized, the total number of folds adopted across the 95 remaining α-synuclein and 62 tau amyloid structures was determined. To achieve this, the radius of gyration (R_g) of all the monomeric sub-units in these structures were calculated:

$$Rg = \sqrt{\frac{\sum_{i=1}^{N} (C_{xi} - CoM_x)^2 + (C_{yi} - CoM_y)^2 + (C_{zi} - CoM_z)^2}{\sum_{i=1}^{N} Mw_i}}$$

where N is the total number of residues in the ordered fibril core, Mw is the molecular weight, and the 3D coordinates for the respective Cα and centre of mass are given by C and CoM respectively.

If the R_g between two protofilaments from the same PDB differed by ≥5% of the given PDB's mean R_g the two protofilaments were deemed as distinct folds, and both protofilament folds were taken for further analysis. If the R_g variation was <5% of the mean R_g, a single protofilament structure was taken to represent all monomers in the PDB entry (Figure S7). Comparing the change in R_g to the percentage of the mean R_g enables the analysis to deal with proteins of varying size. From this analysis, five amyloid structures were identified as containing two distinct protofilament folds for α-synuclein (6pes[39], 6xyo[15], 6xyp[15], 6xyq[15] and 7lc9[68]). Hence, 100 protofilament folds were used to represent the entirety of polymorphism in the 95 α-synuclein amyloid structures. For tau, no fibrils have different protofilament folds, resulting in the total remaining at 62 distinct amyloid structures.

To quantify the degree of variation, the monomers from the different structures for each amyloid were aligned using PyMol[40]. The Cα-Cα distance between shared residues in the ordered fibril cores was calculated:

$$C\alpha\text{-}C\alpha\ Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

where x, y and z represent the coordinates of the two points in 3D space.

Finally, the root-mean-squared deviation (RMSD) was calculated from Cα-Cα distances to give a single value representing the amyloid protofilament fold similarity of residues in the structured core between two structures:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} (v_{xi} - w_{xi})^2 + (v_{yi} - w_{yi})^2 + (v_{zi} - w_{zi})^2}{N}}$$

where N is the total number of residues resolved at high resolution. The two monomers being compared are denoted as v and ω, and x, y and z donate the coordinates of the respective Cα in 3D space.

Using the RMSD values as a measurement of amyloid fold similarity, a single PDB can be chosen as the reference and used to identify which amyloid structures it is most similar (or dissimilar) to (Figures S8A and S9A). Furthermore, the degree of variation amongst all the amyloid folds in the dataset can be visualized in a heatmap of RMSD values (Figures S8B and S9B). To better visualize these distinct groups, hierarchical clustering was performed on the RMSD scores. After manually assigning the cut height and testing its robustness (Figure S10) (Methods), the 100 α-synuclein monomers can be classified into eleven distinct groups corresponding to distinct amyloid polymorphs (Figure 2). For the 62 tau structures, eleven

distinct structural classes resulted (Figure S11A). Note that the regions of the sequence forming the ordered amyloid cores for each of these proteins are highly conserved (Figures S3D and S4D), despite belonging to different structural classes, consistent with recent findings from other groups using these, and other, amyloidogenic proteins[26,41].

Thermodynamic analysis

After classifying α-synuclein and tau amyloid fibrils with distinct polymorphs, we next investigated how the differences can be quantified and easily visualized. Previous work has shown that the diverse polymorphic folds adopted by a protein are all stabilized by a few short sequence segments which remain surprisingly invariant from polymorph to polymorph[26,41]. Inspired by this work, we determined the per-residue contribution to amyloid stability across the 95 α-synuclein and 62 tau well-resolved amyloid fibril structures in our dataset (Figures 3A and S11B). To quantify the similarity of the thermodynamic profiles of each structure, a pairwise correlation matrix was created which revealed a median Pearson correlation of 0.61 for α-synuclein and 0.61 for tau fibrils (Figure S12) (consistent with values previously obtained using a smaller number of sequences[26]). Hence, despite significant differences in their structure, there is a common, polymorph-independent, conservation of the regions of the protein sequence that contribute to the stability of each protein's amyloid fold.

Next, we investigated which types of amino acids are enriched in stabilizing (mean – 1 standard deviation (SD)) or destabilizing (mean + 1SD) regions of each α-synuclein or tau amyloid structure. The number of times a specific amino acid was found in either the stabilizing or destabilizing regions was counted and normalized by the total number of occurrences. This showed, in each case, that stabilizing interactions are mostly driven by hydrophobic or aromatic amino acids (Figures 3B and S11C), whereas destabilizing residues are mostly charged or polar amino acids (Figures 3C and S11D).

Given that the distinct polymorphs of α-synuclein and tau amyloid are stabilized by common regions of the protein sequence, their different polymorphs must arise by different pairings of these regions. To identify stabilizing regions (Methods), we first calculated the mean $\Delta G°$ at each residue position and the mean $\Delta G°$ across all residue positions and structures. Next, we smoothed the $\Delta G°$ per residue values by a sliding window of three residues and found the local negative and positive maxima. Between each local positive maximum, residue positions with a $\Delta G°$ lower than the mean $\Delta G°$ for all residues from each structure was used to denote stabilizing regions (Figure S13). This revealed 11 regions of the α-synuclein sequence and 17

regions of tau that positively contribute to stabilizing their fibril folds. For structures within each RMSD cluster group, the number of times stabilizing regions contained a Cα atom within <10.8Å of a Cα from a different stabilizing region was then calculated, revealing regions that stabilize the amyloid core by non-local interactions.

We used <10.8Å as our distance threshold as this is approximately the Cα-Cα spacing across β-sheets[42,43]. While β-strand propensity for each residue has been used to compare amyloid polymorphs[41], we instead adopted a thermodynamically driven approach. Although β-strands forming regions and stabilizing regions largely overlap, not all β-strand confer the same degree of stability. Therefore, by opting for a thermodynamically driven approach we can separate the presence of β-strands from their contribution to fibril stability (Figure S14).

Diagrams displaying stabilizing regions as nodes and contacts as edges enables visualization of the pairings of stabilizing regions between the different cluster groups obtained using their RMSD, which we define here as different 'structural families' of polymorphs (Figure 3D and S11E). The width of the edges connecting each pair of nodes denotes how many structures that contact occurs in and shows clear differences for each polymorph family. Focusing on the two largest groups (4 and 8) (Figure 4), both contain a highly prevalent contacts between stabilizing regions 5 and 9. However, three strong contacts in group 4 (stabilizing regions 4 to 9, 5 to 8 and 2 to 10) are absent in cluster group 8, which instead contains prevalent contacts between regions 6-9 and 8-11. Hence, different pairings of stabilizing regions can be used to classify and define each amyloid fold. Similar results for the tau amyloid are shown in Figures S11E and S15. While these network diagrams are useful for rapid side-by-side comparisons of the different pairing of stabilizing regions, they do not report the connections at the single residue level. Accordingly, a more detailed network diagram showing the contacts within <10.8Å for each individual residue in each of the amyloid fold families can also be output and analysed (examples for α-synuclein and tau are shown in Figures S16 and S17, respectively).

The fibril structure(s) that are observed during amyloid formation are exquisitely sensitive to the solution conditions employed. For example, changing pH from 5.8 to 7.0 during α-synuclein amyloid formation results in distinct polymorphs residing in groups 3 and 4 (8pk4 vs 8pjo RMSD 25.4Å)[44]. In addition, small changes in sequence (e.g. truncations, point mutations, phosphorylation at specific sites) in both α-synuclein and tau have contributed to the different amyloid structures obtained[39,45–54]. This raises the possibility that the growth conditions, and/or changes in the protein sequence, might steer amyloid formation into the different fold clusters. We also questioned whether the fibrils found in disease might be more stable, or involve different stabilizing regions/residues, than their counterparts formed *in vitro*. The different

resolutions of the fibril structures and sparse population in number of entries for some groups makes statistical analysis of significance difficult. Nonetheless, these calculation suggest that there is no gross difference in stability of ex vivo α-synuclein fibrils[14-16] compared to those formed *in vitro* (no seed) (Figure S18A). Similarly, the different *ex vivo* structures of tau also have similar stability[17–22,34,36,55–61] (Figure S18C). Lastly, we compared the overall stability for each polymorph between each RMSD cluster group (Figure S18B,D). Again, we found no large differences between cluster groups, although we note that some groups are so sparsely populated that form conclusions cannot be drawn until more examples of fibril structures in the currently sparsely populated classes are obtained.

Disease Relevant Polymorphism

(i) α-Synuclein

Using the data presented above, we grouped 95 high-resolution structures from the 101 solved cryo-EM structures of α-synuclein amyloid (Figure S1) into their respective RMSD clusters (Figure 5) showing the location of the 11 stabilizing regions defined in Figure S13A. This analysis highlights the utility of the structural organization and classification and provides insight into how sequence modifications or solution conditions can result in a protein forming amyloid in a similar or highly distinct structural class. Furthermore, for structures with a common topology, it can assign which entries are structurally most closely related.

Starting with the structures of amyloid obtained from individuals with Juvenile-Onset Synucleinopathy (JOS) (8bqv and 8bqw)[16] , the analysis shows that the protofilament folds fall into the largest group (cluster group 8) (Figure 5, red box) and hence they adopt an amyloid topology related to those solved from many *in vitro* studies. The closest related *in vitro* amyloid structure to the JOS amyloid structures is 8pk2[44] (formed at pH 7 in phosphate buffered saline (PBS)) with an average RMSD 2.3Å.

The amyloid fibrils purified from individuals with MSA (6xyo[15], 6xyp[15] and 6xyq[15]) contain two different protofilament structures within the same assembly. The first protofilament structure is located in RMSD cluster group 7 (6xyo_1, 6xyp_1 and 6xyq_1)[15] (Figure 5, green box). It has a distinct structure with an extended N-terminal region that includes stabilizing region 2 that forms a contact with stabilizing region 3 (Figure S19) (average RMSD to the nearest *in vitro* structure, 8hzb[64] of 12.2Å). The second MSA protofilament structure, found in group 8 (Figure 5, green box), extends in the C-terminal region including stabilizing region 11 forming contacts

with stabilising regions 8, 9 and 10 (Figure S19). While 6xyo_2 remains distinct from the other MSA amyloid protofilament folds, 6xyp_2 and 6xyq_2[15] show high structural similarity to some structures formed from wild-type α-synuclein in group 8 formed *in vitro*, with the closest being 7nck[30] and 9euu[65] (both with an average RMSD of 2.8Å). Interestingly, both 7nck[30] and 9euu[65] were produced by incubating wild-type α-synuclein monomers with MSA seeds. While both 7nck[30] and 9euu[65] closely resemble the MSA folds, 7nck[30] only recapitulates the fold with the shorter fibril core and is unusual in that this entry contains only a single protofilament (81 out of 95 of the well resolved α-synuclein amyloid structures contain ≥2 protofilaments). In 9euu[65], both protofilaments adopt the shorter fibril core fold of MSA polymorphs. Hence, creating polymorphs *in vitro* with structures identical (rather than topologically related) to those of *ex vivo* MSA fibrils remains a challenge, even with seeded assembly.

Similarly, 8a9l[14] is the sole member of cluster group 9 (Figure 5, blue box), solved from an *ex vivo* PD sample. This structure is unique and has yet to be reproduced *in vitro*. The closest related structure to 8a9l[14] is 7l7h[66] in group 5 (RMSD of 9.9Å) which was produced by incubating α-synuclein *in vitro* in the presence of tau monomers.

The analysis presented can be used to investigate the impact of different growth conditions on the adopted amyloid structure. For example, six structures of α-synuclein amyloid assembled in the presence of lipid have been solved to date. Three lipidic structures, 8adu[67], 8adv[67] and 8adw[67], are found in group 3 (Figure 5, purple box). The other three structures formed in the presence of lipid are in cluster group 4 (Figure 5, purple box), all 3 of which are closely related (8a4l[67] and 8ads[67]) (RMSD 0.2Å), with 8aex[67] being the most dissimilar with an average RMSD of 2.8Å.

Next, we analysed the impact of familial mutations on the adopted α-synuclein polymorph. The familial variants A53E (7uak[48]), A53T (6lrq[47] and 7wnz[46]) and H50Q (6peo[39] and 6pes_1[39] and 6pes_2[39]) (all formed *in vitro*) are found in group 8 (Figure 5, dashed boxes) and are structurally similar to the MSA type 2 and JOS *ex vivo* structures. One exception is 7wo0[46], an A53T variant, which is found in group 4 (Figure 5, green dashed box box). The E46K variant (structures 6l4s[49] and 7c1d[50]) and the G51D variant (7e0f[51]) are found in cluster group 10 (Figure 5, black/purple dashed boxes, respectively. It is also noteworthy that E46K has also been shown to form an amyloid polymorph (6ufr[45]) similar to 7ncg[30] (RMSD 2.0Å) and 8pjo[44] (RMSD 0.8 Å) formed from wild-type α-synuclein in cluster group 4 (Figure 5, black box).

(ii) Tau

The 62 high resolution structures from the 68 solved cryo-EM structures of 3R+4R and 4R tau amyloid shown in Figure S2 are grouped into their respective clusters in Figure S20. Notably, the largest cluster group (group 3) contains all the 3R+4R tau structures (Figure S11A), while 4R tau structures are distributed across multiple groups, including group 3. Akin to the analysis for α-synuclein fibrils, classification of the different tau fibril folds also enables comparisons of the effects of fibril growth conditions, mutations and fibrils purified from individuals with different tauopathies to be readily compared. For example, analysing the clustering of classes by disease showed that structures from the following tauopathies clustered together in the largest cluster group (group 3): Alzheimer's disease (5o3l[17], 8azu[69], 8bgs[19], 8bgv[19], 8uq7[20], 8fug[70], 7nrv[57], 7upe[36], 7upf[36] and 7nrx[57]), Down syndrome (9bxi[56], 9bxq[56], 9bxo[56] and 9bxr[56]), Alzheimer's disease with Down syndrome (8seh[55] and 8sei[55]), cerebral amyloid angiopathy (7mkf[34]), Age-related tauopathy (7nrq[57], 7nrs[57] and 7nrt[57]), amyotrophic lateral sclerosis with Parkinsonism dementia complex (8otj[21], 8ot6[21], 8otg[21], 8otc[21], 8oth[21] and 8ot9[21]), chronic traumatic encephalopathy (6nwp[18], 8byn[59], 6nwq[18] and 8oti[21]), and subacute sclerosing panencephalitis (8caq[58] and 8cax[58]). Each of the aforementioned structures display great structural similarity with the largest RMSD value occurring between 8bgs[19] and 8oti[21] of 5.9Å.

Interestingly, we found that a subset of diseases form structural classes that are distinct from other tauopathies, but are closely related to each other. Group 6 is solely comprised of structures from corticobasal degeneration (CBD) (6tjx[22] from *ex vivo* samples, 8orf[60] and 8org[60] from cell extracts seeded with *ex vivo* CBD fibrils) and argyrophilic grain disease (7p6d[61] and 7p6e[61]). The largest RMSD score for group 6 is 11.9Å between 6tjx[22] and 8orf[60]. Similarly, group 8 contains structures from progressive supranuclear palsy (PSP) (7p65[61]), globular glial tauopathy (7p66[61], 7p67[61] and 7p68[61]) and limbic-predominant neuronal inclusion body tauopathy (7p6a[61], 7p6b[61] and 7p6c[61]) amyloids. The largest RMSD score for group 8 is 10.4Å between the PSP structure, 7p65[61], and the globular glial tauopathy fold, 7p67[61], showing that fibrils with a related topology can differ significantly in atomic detail. In other cases, tauopathies that are clinically similar but neuropathologically distinct, such as both CBD and PSP,have different topological folds that are clustered in the structural classes of group 6 and group 8, respectively[61,71].

Looking at mutant tau structures, V337M (9eo7[54], 9eoh[54], 9eoe[54] and 9eo9[54]), R406W (9eog[54]) and D395G (9erm[53], 9ern[53] and 9ero[53]) comprise the remaining structures in group 3. Other structures of mutants result in distinct topologies that become sole members of their RMSD cluster group: S202E + T205E + S208E (8ttl[52]) in group 2, S396E + T403E + S404E (8ttn[52]) ingroup 7 and P301L (8wcp[72]) in group 10. The point mutation, P301S, is of particular interest as it forms two amyloid structures that are not only distinct from each other (RMSD for 8q92[73]

and 8q96[73] is 42.2Å) but also from the other published 4R and 3R+4R tau structures. The closest structure to 8q92[73] is 9eoh[54] with an RMSD score of 24.3Å. 8q92[73] forms its own distinct cluster (group 5) whereas 8q96[73] is in group 11, but appears visually distinct from the only other member, 6qjh[74] (RMSD 17.2Å). Hence, single residue substitutions can drive assembly to a different amyloid fold, highlighting the sensitivity of the assembly process to small sequence changes (and presumably therefore also to post-translational modifications *in vivo*).

Discussion

Classifying protein structures into different, or structurally related, families of folds has been vital to understand protein function[23–25]. In these approaches, protein folds that belong to the same class are grouped and further divided into sub-classes based on additional metrics such as evolutionary origin, sequence conservation, topology and the arrangement of secondary structure elements. We have shown here that a similar strategy can be used to interrogate amyloid structures, especially those for which sufficient different sequences have been solved to near-atomic resolution. This enables the similarity of their folds (the topological class, defined by the proximity of stabilizing elements of structure) and their relatedness within a topological class (defined by RMSD) to be clustered and classified. The high-throughput nature of our analysis pipeline enables the rapid comparison of newly solved structures to the expanding database of pre-existing topological classes for a given protein of interest. The scripts written for this analysis are available online (see Data Availability).

We demonstrate our strategy for α-synuclein, 3R+4R tau and 4R tau amyloids. The results highlight that the current α-synuclein amyloid structures can be clustered into eleven distinct classes, in which two (clusters 4 and 8) contain the vast majority of structures solved to date (81 of the 95 well-resolved structures considered here). For tau, eleven structural classes are observed, with one (group 3) containing 40 of the 62 well-resolved tau amyloid structures deposited to date. The different *ex vivo* tau amyloid structures occupy 3 of the eleven topological classes.

Another striking finding from our analysis is that all 95 amyloid structures of α-synuclein have similar per-residue stability profile (Figure 3A) and have a similar overall stability in the fully formed fibril (Figure S18A-B). This result was recapitulated for tau (Figure S18C-D). This suggests that the fibrils observed are perhaps being those that are kinetically most able to nucleate and/or elongate[44], rather than being those that are most stable thermodynamically. In contrast with the folding of globular proteins, therefore, the ultimate fold of amyloid fibrils is determined by the assembly process, rather than the amino acid sequence alone. Recent findings which show that amyloid fibrils formed early in assembly differ structurally from those observed at steady state also raise the possibility that the amyloid structures formed *in vitro* that have been solved to date may not represent the end points of an assembly reaction, but could be assembly intermediates[75,76]. Further experiments in which the influence of mechanism of assembly (the balance of primary versus secondary events), the solution conditions (titration of different ligands e.g. lipids, metal ions, metabolites and others) and changes in the sequence (truncations, mutations and relevant post-translational modifications)

are tracked over the time-course of assembly will be needed to better understand how amyloid formation is funnelled towards a specific polymorph and to address the gross differences in polymorphism in amyloid fibrils formed *in vitro* compared with those observed *ex vivo*.

Identifying growth conditions that recapitulate disease-specific polymorphs will provide further insight into the factors that govern the amyloid fold adopted, as well as allowing the reproduction of disease-relevant amyloids for use in *in vitro* and in-cell experiments. The classification of amyloid structures into different topological groups, ordered by their closest structural neighbours, enables the identification of solution conditions that result (or most closely result) in the formation of a particular amyloid fold of interest. Our analysis shows that for α-synuclein, the JOS protofilament fold can be recapitulated *in vitro,* whereas those for the MSA and PD polymorphs remain limited to *ex vivo* samples, at least to date.

Finally, we have also shown that the per-residue thermodynamic profile of α-synuclein and tau amyloid folds is consistent across polymorphs, with the same, or very similar, regions of the sequence contributing to fibril stability. The different amyloid structures thus result from the differential pairing these stabilizing regions into an amyloid fold. The thermodynamic stability of the amyloid fold hence does not define the product of assembly (at least over the timescales currently tested and structures currently available). Instead, the amyloid structure that is observed must result from the kinetic search for the amyloid fold. Factors that reduce kinetic barriers, alter solubility, or tip what must be a shallow energy landscape with deep energy wells, will determine which amyloid fold(s) result.

Resource Availability

Lead Contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Prof. David Brockwell (d.j.brockwell@leeds.ac.uk).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The presented work utilises previously published amyloid structures that are publicly available. A curated list of the PDB IDs of all structures used can be found in Supplementary Tables S1 and S2. All original code generated for this analysis is publicly available on GitHub: https://github.com/JackConnor98/classifying_amyloid_polymorphs. The data associated with this paper that is not already provided in the manuscript and Supplementary Information i is publicly available from the University of Leeds at https://doi.org/10.5518/1659. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Acknowledgments

JPC is funded by BBSRC (BB/T007222/1). SER acknowledges funding from the Royal Society (RSRP/R1/211057). We thank members of our groups, especially in our Amyloid Team, for many helpful discussions during the development of this work. We thank David Eisenberg and Michael Sawaya for supplying scripts to calculate amyloid solvation energies. For the purpose of open access, the authors have applied a CC-BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Author Contributions

All authors contributed to the conceptualization of the work. JPC performed all data curation and analysis and wrote all the software used. All authors contributed to writing and editing the

manuscript. DJB and SER acquired funding and were responsible for supervision and project administration.

Declaration of Interests

The authors declare no competing interests.

FIGURES AND LEGENDS

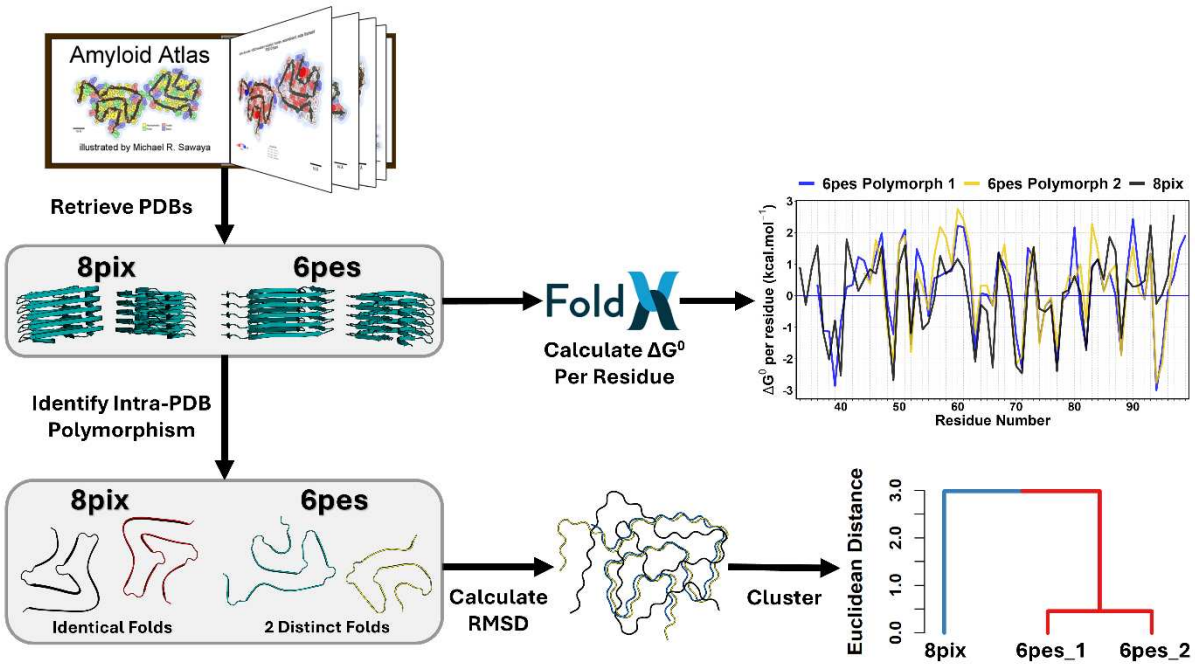

Figure 1: A schematic of the analysis pipeline. PDB entries of the amyloid of interest are retrieved from Amyloid Atlas[13]. Any intra-PDB variation between the structures of protofilaments that form a single amyloid fibril is determined. Unique protofilament structures are then aligned and the Cα-Cα distances are calculated across all well-resolved residues in their cores. The calculated RMSD scores are then used to cluster PDBs based on their amyloid fold similarity. In addition, whole fibrils with 10 layers are analysed by FoldX[27] to calculate the per-residue free energy for each structure (omitting the top and bottom layer of each fibril (Methods)).

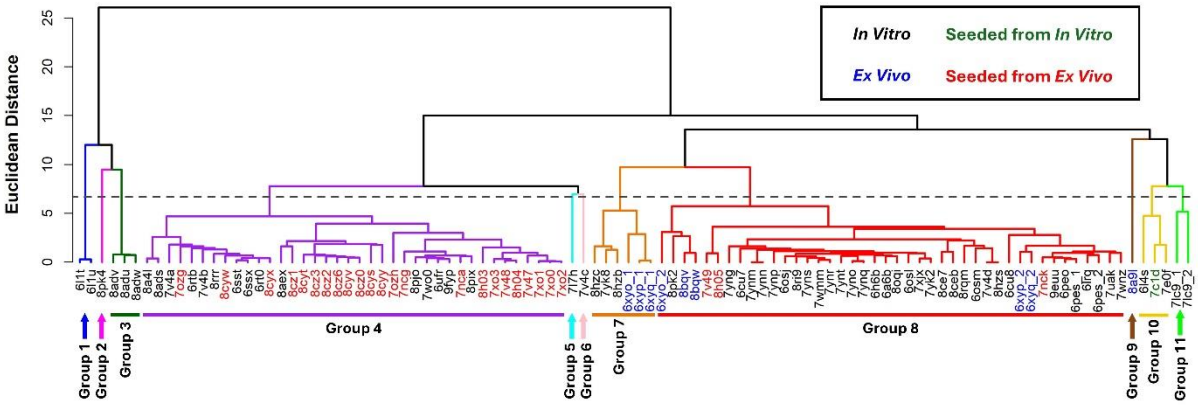

Figure 2: Euclidean distance hierarchical clustering of amyloid fold similarity between solved PDB structures of α-synuclein based on RMSD values. Eleven structural classes (Group 1-11) result with branches shown in different colours. The PDB code for each structure is given below coloured black (*in vitro*), blue (*ex vivo*), red (seeded from *ex vivo*) and green (seeded from *in vitro*).

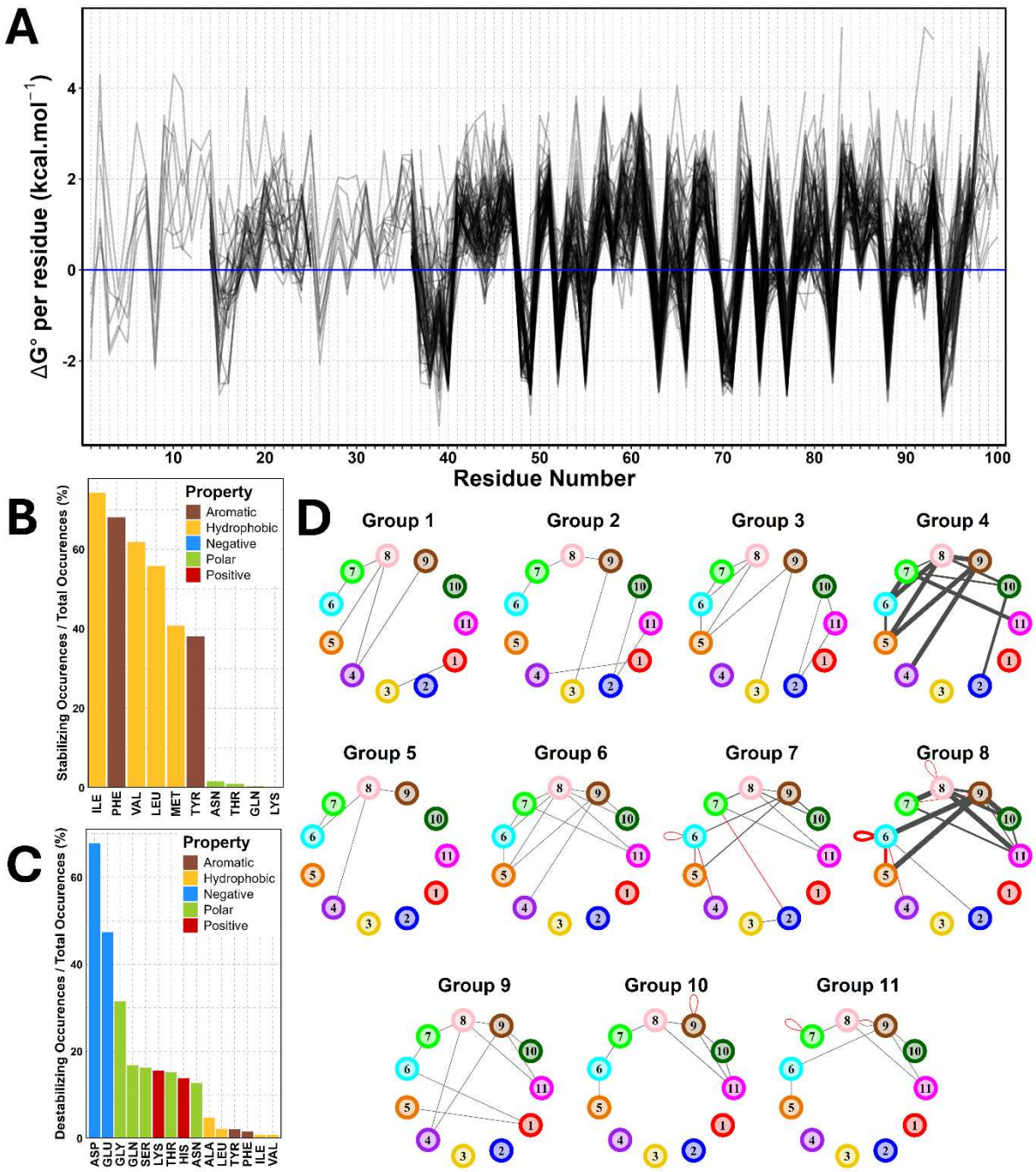

Figure 3: Thermodynamic analysis of α-synuclein using FoldX. A) Distributions ΔG° per residue for published α-synuclein structures after Q-score validation. The blue line denotes a ΔG° per residue of 0, with positive values indicating a destabilizing contribution and negative values indicating stabilizing residues. B-C) Bar charts showing the normalized number of times each amino acid was found to be either stabilizing (B) or destabilizing (C). D) Network diagrams showing the number of times stabilizing regions are found within <10.8Å of each other. Nodes represent the stabilizing regions and are coloured as in Figure S13A. The intra-protofilament (within a single layer) and inter-protofilament contacts are shown by black and red lines, respectively. The width of the lines indicates the number of times a contact

between the stabilizing regions occurs across all members of the group. The α-synuclein structures are separated into distinct polymorphs based on the cluster groups identified in Figure 2. The percentage of structures involved in each contact within a group is presented in Table S3.

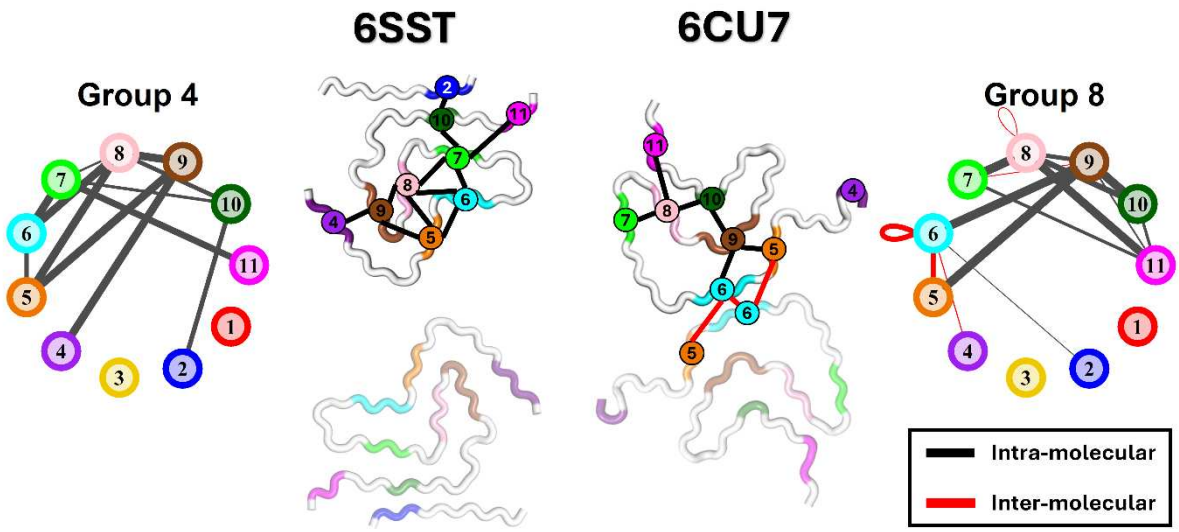

Figure 4: Visualization of the network diagrams shown in Figure 3D for example α-synuclein structures. Shown are two example structures: 6SST[62] and 6CU7[63] taken from the two largest cluster groups (Figure 2), group 4 and group 7, respectively. Both structures are formed from two protofilaments. For each structure, a single protofilament has been labelled with its stabilizing regions denoted by small, numbered circles and contacts within <10.8Å shown as solid lines. For both the network diagrams and the annotated structures, the intra-protofilament (within a single layer) and inter-protofilament contacts are shown by black and red lines, respectively. For simplicity, the second protofilament for each structure is unlabelled except for instances of inter-protofilament contacts.

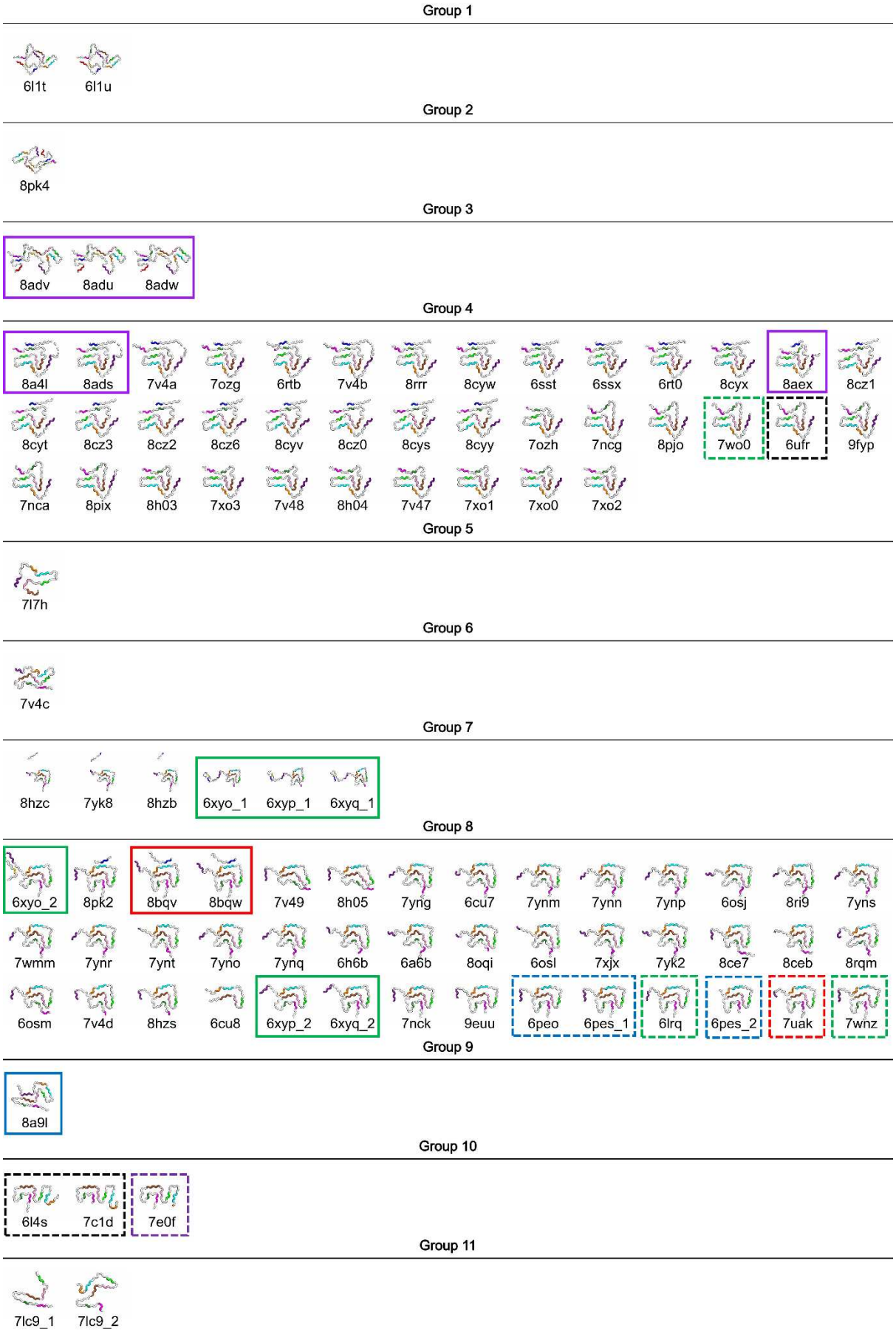
Group 1
6l1t
6l1u
Group 2
8pk4
Group 3
8adv
8adu
8adw
Group 4
8a4l
8ads
7v4a
7ozg
6rtb
7v4b
8rrr
8cyw
6sst
6ssx
6rt0
8cyx
8aex
8cz1
8cyt
8cz3
8cz2
8cz6
8cyv
8cz0
8cys
8cyy
7ozh
7ncg
8pjo
7wo0
6ufr
9fyp
7nca
8pix
8h03
7xo3
7v48
8h04
7v47
7xo1
7xo0
7xo2
Group 5
7l7h
Group 6
7v4c
Group 7
8hzc
7yk8
8hzb
6xyo_1
6xyp_1
6xyq_1
Group 8
6xyo_2
8pk2
8bqv
8bqw
7v49
8h05
7yng
6cu7
7ynm
7ynn
7ynp
6osj
8ri9
7yns
7wmm
7ynr
7ynt
7yno
7ynq
6h6b
6a6b
8oqi
6osl
7xjx
7yk2
8ce7
8ceb
8rqm
6osm
7v4d
8hzs
6cu8
6xyp_2
6xyq_2
7nck
9euu
6peo
6pes_1
6lrq
6pes_2
7uak
7wnz
Group 9
8a9l
Group 10
6l4s
7c1d
7e0f
Group 11
7lc9_1
7lc9_2

Figure 5: High-resolution cryo-EM α-synuclein structures grouped into RMSD clusters and coloured by stable regions. Solid line boxes denote structures solved from *ex vivo* samples from Parkinson's disease dementia (PDD) (blue), MAAAEKT insertion JOS (red), and MSA (green). Structures formed in the presence of lipid are indicated by solid purple outlines. Dashed boxes are structures solved from familial variants of α-synuclein: H50Q (blue), A53E (red), A53T (green), E46K (black), and G51D (purple). The order in which structures are shown in each group is based on their fold similarity, matching the order of the dendrogram in Figure 2 from left to right. Stable regions are coloured as shown in Figures 3D and S13A.

STAR METHODS

Method Details

Collating α-Synuclein Structures

The analysis pipeline employed in this work was designed to enable high-throughput analysis of amyloid structures. As new structures are being solved at an increasingly high rate, we created a web-scraping Python[77] script that retrieves the PDB codes for a given protein of interest directly from the amyloid atlas. This way, the analysis pipeline can be repeated upon the release of new structures. The structures analysed in this work incorporate all published cryo-EM structures of α-synuclein listed on Amyloid Atlas as of September 2024 and are curated in Supplementary Tables S1 and S2.

Quality Control and Validation

Structures with low resolution were removed before structural and thermodynamic analysis to account for differential quality and resolution amongst the published structures. The quality of each structure was determined using their Q-scores[29]. To retrieve the Q-scores, a web scraping script was written in Python[77] to extract the Q-scores for each structure from the Electron Microscopy Data Bank[78]. First, the average Q-score for each individual structure is determined. Structures with an overall low quality (determined as the mean minus one standard deviation from the average Q-score from all published structures) were removed from the analysis (Tables S1 and S2) (Figures S3A and S4A). As FoldX requires high-resolution structures for accurate calculations, we removed poorly resolved individual residues

from an overall well-resolved structure. As before, the mean Q-score for each residue from all structures was calculated. Residues with a Q-score lower than the mean minus one standard deviation were removed from thermodynamic analysis (Figures S3B-C and S4B-C).

RMSD clustering

The R[79] package dendextend[80] was used for hierarchical cluster analysis of amyloid fold similarity using the calculated RMSD scores. The Euclidean distance between datapoints was calculated and the dendrogram was constructed using average-linkage. To assign the cut height, scree plots were generated for both α-synuclein and tau (Figure S10). From this a cut height was assigned and manually adjusted to create groups with visibly distinct amyloid folds. The assigned cut heights were 6.7Å and 5.5Å for α-synuclein and tau, respectively.

Calculating ∆G° per Residue

The FoldX version 5 force field[27] was used to calculate the overall contribution of each residue to the total free energy (∆G°) for high-resolution α-synuclein and tau structures. To account for published structures differing in the number of layers, each structure was extended to a layer depth of 10 (i.e. each fibril contained a stack of 10 monomers). Only the internal 8 layers were used for FoldX calculations with the terminal monomers being ignored to account for the loss of stabilizing head-to-tail stacking interactions. Most published structures do not include water molecules, so all water atoms were removed prior to thermodynamic analysis. Residue 35 from the α-synuclein structure 7v4a[81] was removed from **Figure 3A** as an outlier due to incorrect modelling of the side chain (ΔG° per residue = 9.54 kcal.mol^{-1}).

A Python[77] script using the pyFoldX package[82] was used to automate FoldX calculations of ∆G° allowing for high-throughput analysis. First, the command RepairPDB is used to minimise the energy of the structure by rearranging the side chains to find a new energy minimum. Next, the script runs the command, SequenceDetail, which returns the FoldX calculated energy terms averaged for each residue. SequenceDetail calculates the total ∆G° contribution for each residue as follows:

$$\Delta G^\circ = \Delta G^\circ_{wdw} + \Delta G^\circ_{solvH} + \Delta G^\circ_{solvP} + \Delta G^\circ_{wb} + \Delta G^\circ_{hbond} + \Delta G^\circ_{el} + \Delta G^\circ_{kon} + \Delta G^\circ_{clash} + \Delta S_{mc} + \Delta S_{sc}$$

The total contribution to free energy for each residue is calculated by summing the following individual energy terms; the sum of Van der Waals contributions (ΔG°_{vdw}), the solvation energy

for apolar ($\Delta G°_{solvH}$) and polar groups ($\Delta G°_{solvP}$), water bridges ($\Delta G°_{wb}$), hydrogen bonding ($\Delta G°_{hbond}$), electrostatic interactions ($\Delta G°_{el}$), a second metric measuring the electrostatic interactions between different polypeptide chains ($\Delta G°_{kon}$), steric clashes ($\Delta G°_{clash}$), and the entropy of the main (ΔS_{mc}) and side chains (ΔS_{sc}). The total contribution to free energy does not include the backbone van der Waals clashes.

Calculating Solvation Free Energy

Calculating the solvation free energy was achieved using methods developed by the Eisenberg group[37,38]. Firstly, each PDB amyloid structure is extended to a layer depth of 10 and all hydrogen atoms are removed. Next, 1000 points are evenly distributed on the surface of each atom using the golden spiral algorithm. Using these points, the solvent accessible surface area (SASA) of each atom in the fibrillar state is calculated. The SASA for each residue in the unfolded state is calculated using isolated tripeptides. For residue i in the unfolded state, all atoms except for the main chain atoms of residues i + 1 and i - 1 are removed and SASA is recalculated. Using the folded and unfolded SASA values and the atomic solvation parameters (ASP) reported previously[37,38], we can calculate the solvation energy using the following formula:

$$\Delta G_s = \Delta\sigma \cdot \sum_{i=1}^{n} fA_i - uA_i$$

where the solvation energy (ΔG_s) is given as the sum across all atoms (n) of the dot product of the ASP ($\Delta\sigma$) and the area buried. The area buried is calculated from the difference in SASA from the folded (fA) and unfolded states (uA).

Defining Stable Regions

To define stabilizing regions within each amyloid structure, the $\Delta G°$ at each residue position was calculated using FoldX and averaged for all structures giving a single value representing the mean $\Delta G°$ per residue at each position. The data were smoothed using a sliding window average with a window size of 3. In addition, the mean $\Delta G°$ for all residue positions from all structures was calculated and used as the threshold for determining stabilizing regions.

The local minima and maxima were calculated by employing the turnpoints function within the R package pastecs[83]. Consecutive residues between two local maxima with a mean ∆G° per residue lower than the assigned threshold were defined as a stabilizing region. This process was repeated between all local maxima peaks (Figure S13). This analysis identified 11 stabilizing regions of α-synuclein comprised of the following residues: 2-5, 15-17, 26-27, 36-40, 48-50, 52-55, 63-65, 69-72, 74-78, 88-89 and 93-95. For tau, 17 such regions were observed: 275-283, 285-287, 296-301, 306-314, 317-319, 327-331, 336-340, 343-345, 350-351, 353-354, 358-364, 375-377, 391-395, 397-399, 405-411, 423-427 and 435-438.

β-strand Assignment

A Ramachandran plot was generated for each structure and, using the thresholds specified by Wilmot and Thornton, 1990[84], each residue was classed as either β-strand or non-β-strand. In addition to phi psi angles, residues were only considered to be in a β-conformation if they belonged to a continuous β-strand of ≥4 residues (Lβ ≥ 4). As previously described[41], we counted the number of times a given residue occurred in a β-conformation across all chains from all structures. Taking the number of β-conformation occurrences and the total number of occurrences, we calculated the fraction in β-conformation for each residue (Figure S14).

Programs and Packages

Work with PDB structures was conducted using Python[77] and the molecular visualisation tool PyMol[40]. Web scraping was written in Python using the BeautifulSoup4[85] library. Data analysis was performed using R[79] in RStudio[86]. Data visualisation employed the R package ggplot2[87] with the exception of network diagrams which were produced using the R package igraph[88,89].

Quantification and statistical analysis

Pearson correlation was performed to compare the outputs of ∆G° per Residue and Solvation Free Energy calculations (described in the legends of Figure S5 and S6). Pearson correlation was also performed to compare the similarity of ΔG° per residue profiles of all α-synuclein structures and all Tau ΔG° structures (described in Figure S12 legend). A non-parametric Wilcoxon rank sum test was used to test whether for significant difference in the thermodynamic stability of α-synuclein and tau amyloid fibrils formed under different conditions, (described in Figure S18 legend).

Supplemental Information

Document S1. Figures S1-S20 and Tables S1-S4.

Table S1: A list of all published α-synuclein amyloid structures deposited on Amyloid Atlas[2,13], related to STAR methods This table describes the reason for any structures that were not included in our analysis.

Table S2: A list of all published 3R-4R or 4R Tau amyloid structures deposited on Amyloid Atlas [2,13], related to STAR methods. This table describes the reason for any structures that were not included in our analysis.

Table S3: α-Synuclein contact frequency between stable regions for each RMSD cluster group, related to Figure 3.

References

1. Iadanza, M. G., Jackson, M. P., Hewitt, E. W., Ranson, N. A., and Radford, S. E. (2018). A new era for understanding amyloid structures and disease. Nat Rev Mol Cell Biol *19*, 755–773. doi:10.1038/s41580-018-0060-8.

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. doi:10.1016/j.cell.2021.08.013.

3. Scheres, S. H. W., Ryskeldi-Falcon, B., and Goedert, M. (2023). Molecular pathology of neurodegenerative diseases by cryo-EM of amyloids. Nature *621*, 701–710. doi:10.1038/s41586-023-06437-2.

4. Buxbaum, J. N., Eisenberg, D. S., Fändrich, M., McPhail, E. D., Merlini, G., Saraiva, M. J. M., Sekijima, Y., and Westermark, P. (2024). Amyloid nomenclature 2024: update, novel proteins, and recommendations by the International Society of Amyloidosis (ISA) Nomenclature Committee. Amyloid *31*, 249–256. doi:10.1080/13506129.2024.2405948.

5. Glenner, G. G., and Wong, C. W. (1984). Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. Biochem Biophys Res Commun *120*, 885–890. doi:10.1016/s0006-291x(84)80190-4

6. Masters, C. L., Simms, G., Weinman, N. A., Multhaup, G., McDonald, B. L., and Beyreuther, K. (1985). Amyloid plaque core protein in Alzheimer disease and Down syndrome. Proc Natl Acad Sci USA *82*, 4245-4249. doi:10.1073/pnas.82.12.4245.

7. Goedert, M., Wischik, C. M., Crowther, R. A., Walker, J. E., and Klug, A. (1988). Cloning and sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: identification as the microtubule-associated protein tau. Proc Natl Acad Sci USA *85*, 4051-4055. doi:10.1073/pnas.85.11.4051

8. Spillantini, M. G., Schmidt, M. L., Lee, V. M.-Y., Trojanowski, J. Q., Jakes, R., and Goedert, M. (1997). α-Synuclein in Lewy bodies. Nature *388*, 839-840. doi:10.1038/42166.

9. Tuttle, M. D., Comellas, G., Nieuwkoop, A. J., Covell, D. J., Berthold, D. A., Kloepper, K. D., Courtney, J. M., Kim, J. K., Barclay, A. M., Kendall, A., Wan, W., Stubbs, G., Schwieters, C. D., Lee, V. M. Y., George, J. M., and Rienstra, C. M. (2016). Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. Nature Struc Mol Biol *23*, 409-415. doi:10.1038/nsmb.3194.

10. Guerrero-Ferreira, R., Taylor, N. M., Mona, D., Ringler, P., Lauer, M. E., Riek, R., Britschgi, M., and Stahlberg, H. (2018). Cryo-EM structure of alpha-synuclein fibrils. eLife *7*, e36402. doi:10.7554/eLife.36402.

11. Li, Y., Zhao, C., Luo, F., Liu, Z., Gui, X., Luo, Z., Zhang, X., Li, D., Liu, C., and Li, X. (2018). Amyloid fibril structure of α-synuclein determined by cryo-electron microscopy. Cell Res *28*, 897-903. doi:10.1038/s41422-018-0075-x.

12. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research *28*, 235-242. doi:10.1093/nar/28.1.235.

13. Sawaya, M. R., Hughes, M. P., Rodriguez, J. A., Riek, R., and Eisenberg, D. S. Amyloid Atlas (2024). https://people.mbi.ucla.edu/sawaya/amyloidatlas/.

14. Yang, Y., Shi, Y., Schweighauser, M., Zhang, X., Kotecha, A., Murzin, A. G., Garringer, H. J., Cullinane, P. W., Saito, Y., Foroud, T., Warner, T. T., Hasegawa, K., Vidal, R., Murayama, S., Revesz, T., Ghetti, B., Hasegawa, M., Lashley, T., Scheres, S. H. W., and Goedert, M. (2022). Structures of α-synuclein filaments from human brains with Lewy pathology. Nature *610*, 791-795. doi:10.1038/s41586-022-05319-3.

15. Schweighauser, M., Shi, Y., Tarutani, A., Kametani, F., Murzin, A. G., Ghetti, B., Matsubara, T., Tomita, T., Ando, T., Hasegawa, K., Murayama, S., Yoshida, M., Hasegawa, M., Scheres, S. H. W., and Goedert, M. (2020). Structures of α-synuclein filaments from multiple system atrophy. Nature *585*, 464–469. doi:10.1038/s41586-020-2317-6.

16. Yang, Y., Garringer, H. J., Shi, Y., Lövestam, S., Peak-Chew, S., Zhang, X., Kotecha, A., Bacioglu, M., Koto, A., Takao, M., Spillantini, M. G., Ghetti, B., Vidal, R., Murzin, A. G., Scheres, S. H. W., and Goedert, M. (2023). New SNCA mutation and structures of α-synuclein filaments from juvenile-onset synucleinopathy. Acta Neuropathologica *145*, 561–572. doi:10.1007/s00401-023-02550-8.

17. 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. doi:10.1038/nature23002.

18. 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. doi:10.1038/s41586-019-1026-5.

19. Fowler, S. L., Behr, T. S., Turkes, E., O'Brien, D. P., Cauhy, P. M., Rawlinson, I., Edmonds, M., Foiani, M. S., Schaler, A., Crowley, G., Bez, S., Ficulle, E., Tsefou, E., Fischer, R., Geary, B., Gaur, P., Miller, C., D'Acunzo, P., Levy, E., Duff, K. E., and Ryskeldi-Falcon, B. (2025). Tau filaments are tethered within brain extracellular vesicles in Alzheimer's disease. Nature Neurosci *28*, 40-48. doi:10.1038/s41593-024-01801-5.

20. Kunach, P., Vaquer-Alicea, J., Smith, M. S., Monistrol, J., Hopewell, R., Moquin, L., Therriault, J., Tissot, C., Rahmouni, N., Massarweh, G., Soucy, J.-P., Guiot, M.-C., Shoichet, B. K., Rosa-Neto, P.,

Diamond, M. I., and Shahmoradian, S. H. (2024). Nature Commun *15*, 8497. doi:10.1038/s41467-024-52265-x.

21. Qi, C., Verheijen, B. M., Kokubo, Y., Shi, Y., Tetter, S., Murzin, A. G., Nakahara, A., Morimoto, S., Vermulst, M., Sasaki, R., Aronica, E., Hirokawa, Y., Oyanagi, K., Kakita, A., Ryskeldi-Falcon, B., Yoshida, M., Hasegawa, M., Scheres, S. H. W., and Goedert, M. (2023). Tau filaments from amyotrophic lateral sclerosis/parkinsonism-dementia complex adopt the CTE fold. Pro. Natl Acad Sci USA *120*, e2306767120. doi:10.1073/pnas.2306767120.

22. Zhang, W., Tarutani, A., Newell, K. L., Murzin, A. G., Matsubara, T., Falcon, B., Vidal, R., Garringer, H. J., Shi, Y., Ikeuchi, T., Murayama, S., Ghetti, B., Hasegawa, M., Goedert, M., and Scheres, S. H. W. (2020). Novel tau filament fold in corticobasal degeneration. Nature *580*, 283-287. doi:10.1038/s41586-020-2043-0.

23. Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M., and Orengo, C. A. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Research *35*, 291-297. doi:10.1093/nar/gkl959.

24. Knudsen, M., and Wiuf, C. (2010). The CATH database. Human Genomics *4*, 207. doi:10.1186/1479-7364-4-3-207.

25. Andreeva, A., Kulesha, E., Gough, J., and Murzin, A. G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Research *48*, 376-382. doi:10.1093/nar/gkz1064.

26. van der Kant, R., Louros, N., Schymkowitz, J., and Rousseau, F. (2022). Thermodynamic analysis of amyloid fibril structures reveals a common framework for stability in amyloid polymorphs. Structure *30*, 1178-1189.e3. doi:10.1016/j.str.2022.05.002.

27. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Research *33*, 382–388. doi:10.1093/nar/gki387.

28. Dhavale, D. D., Barclay, A. M., Borcik, C. G., Basore, K., Gordon, I. R., Liu, J., Milchberg, M. H., O'shea, J., Rau, M. J., Smith, Z., Sen, S., Summers, B., Smith, J., Warmuth, O. A., Chen, Q., Fitzpatrick, J. A. J., Schwieters, C. D., Tajkhorshid, E., Rienstra, C. M., and Kotzbauer, P. T. (2024). Structure of alpha-synuclein fibrils derived from human Lewy body dementia tissue. Nature Commun *15*, https://doi.org/10.1038/s41467-024-46832-5

29. Pintilie, G., Zhang, K., Su, Z., Li, S., Schmid, M. F., and Chiu, W. (2020). Measurement of atom resolvability in cryo-EM maps with Q-scores. Nature Methods *17*, 328–334. doi:10.1038/s41592-020-0731-1.

30. Lövestam, S., Schweighauser, M., Matsubara, T., Murayama, S., Tomita, T., Ando, T., Hasegawa, K., Yoshida, M., Tarutani, A., Hasegawa, M., Goedert, M., and Scheres, S. H. W. (2021). Seeded

assembly in vitro does not replicate the structures of α-synuclein filaments from multiple system atrophy. FEBS Open Bio *11*, 999–1013. doi:10.1002/2211-5463.13110.

31. Balana, A. T., Mahul-Mellier, A.-L., Nguyen, B. A., Horvath, M., Javed, A., Hard, E. R., Jasiqi, Y., Singh, P., Afrin, S., Pedretti, R., Singh, V., Lee, V. M.-Y., Luk, K. C., Saelices, L., Lashuel, H. A., and Pratt, M. R. (2024). O-GlcNAc forces an α-synuclein amyloid strain with notably diminished seeding and pathology. Nature Chem Biol *20*, 646–655. doi:10.1038/s41589-024-01551-2.

32. Tao, Y., Zhao, Q., Liu, C., and Li, D. Na (Unpublished). doi: doi.org/10.2210/pdb7WMM/pdb, doi.org/10.2210/pdb7YNM/pdb, doi.org/10.2210/pdb7YNN/pdb, doi.org/10.2210/pdb7YNO/pdb, doi.org/10.2210/pdb7YNQ/pdb, doi.org/10.2210/pdb7YNP/pdb, doi.org/10.2210/pdb7YNL/pdb, doi.org/10.2210/pdb7YNG/pdb, doi.org/10.2210/pdb7YNR/pdb, doi.org/10.2210/pdb7YNS/pdb, doi.org/10.2210/pdb7YNT/pdb.

33. Sokratian, A., Zhou, Y., Xu, E., Viverette, E., Dillard, L., Yuan, Y., Li, J. Y., Matarangas, A., Bouvette, J., Borgnia, M., Bartesaghi, A., and West, A (2022). Structural and functional landscape of α-synuclein fibril conformations amplified from cerebrospinal fluid. bioRxiv. doi:10.1101/2022.07.13.499896.

34. Hallinan, G. I., Hoq, M. R., Ghosh, M., Vago, F. S., Fernandez, A., Garringer, H. J., Vidal, R., Jiang, W., and Ghetti, B. (2021). Structure of Tau filaments in Prion protein amyloidoses. Acta Neuropathologica *142*, 227-241. doi:10.1007/s00401-021-02336-w.

35. Chang, A., Xiang, X., Wang, J., Lee, C., Arakhamia, T., Simjanoska, M., Wang, C., Carlomagno, Y., Zhang, G., Dhingra, S., Thierry, M., Perneel, J., Heeman, B., Forgrave, L. M., DeTure, M., DeMarco, M. L., Cook, C. N., Rademakers, R., Dickson, D. W., Petrucelli, L., Stowell, M. H. B., Mackenzie, I. R. A., and Fitzpatrick, A. W. P. (2022). Homotypic fibrillization of TMEM106B across diverse neurodegenerative diseases. Cell *185*, 1346-1355.e15. doi:10.1016/j.cell.2022.02.026.

36. Seidler, P. M., Murray, K. A., Boyer, D. R., Ge, P., Sawaya, M. R., Hu, C. J., Cheng, X., Abskharon, R., Pan, H., DeTure, M. A., Williams, C. K., Dickson, D. W., Vinters, H. V., and Eisenberg, D. S. (2022). Structure-based discovery of small molecules that disaggregate Alzheimer's disease tissue derived tau fibrils in vitro. Nature Commun *13*, 5451. doi:10.1038/s41467-022-32951-4.

37. Eisenberg, D., and McLachlan, A. D. (1986). Solvation energy in protein folding and binding. Nature *319*, 199-203. doi:10.1038/319199a0.

38. Eisenberg, D., Wesson, M., and Yamashita, M. (1986). Interpretation of protein folding and binding with atomic solvation parameters. Chemica Scripta *29A*, 217–221.

39. Boyer, D. R., Li, B., Sun, C., Fan, W., Sawaya, M. R., Jiang, L., and Eisenberg, D. S. (2019). Structures of fibrils formed by α-synuclein hereditary disease mutant H50Q reveal new polymorphs. Nature Struct Mol Biol *26*, 1044-1052. doi:10.1038/s41594-019-0322-y.

40. Schrödinger, L., and DeLano, W. PyMOL (2020). http://www.pymol.org/pymol.

41. Errico, S., Fani, G., Ventura, S., Schymkowitz, J., Rousseau, F., Trovato, A., Vendruscolo, M., Bemporad, F., and Chiti, F. (2025). Structural commonalities determined by physicochemical principles in the complex polymorphism of the amyloid state of proteins. Biochem J *482*, 87–101. doi:10.1042/BCJ20240602.

42. Fändrich, M., and Dobson, C. M. (2002). The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. EMBO J *21*, 5682-5960. doi:10.1093/emboj/cdf573

43. Makin, O. S., Atkins, E., Sikorski, P., Johansson, J., and Serpell, L. C. (2005). Molecular basis for amyloid fibril formation and stability. Proc Natl Aca. Sci USA. *102*, 315-320. doi:10.1073/pnas.0406847102

44. Frey, L., Ghosh, D., Qureshi, B. M., Rhyner, D., Guerrero-Ferreira, R., Pokharna, A., Kwiatkowski, W., Serdiuk, T., Picotti, P., Riek, R., and Greenwald, J. (2024). On the pH-dependence of α-synuclein amyloid polymorphism and the role of secondary nucleation in seed-based amyloid propagation. eLife *12*, RP93562. doi:10.7554/eLife.93562.3.

45. Boyer, D. R., Li, B., Sun, C., Fan, W., Zhou, K., Hughes, M. P., Sawaya, M. R., Jiang, L., and Eisenberg, D. S. (2020). The α-synuclein hereditary mutation E46K unlocks a more stable, pathogenic fibril structure. Proc Natl Acad Sc. USA *117*, 3592-3602. doi:10.1073/pnas.1917914117.

46. J.Y.-C, H., and Wu, K.-P. CryoEM structure of human alpha-synuclein A53T fibril (Unpublished). doi:10.2210/pdb7wnz/pdb, 10.2210/pdb7wo0/pdb.

47. Sun, Y., Hou, S., Zhao, K., Long, H., Liu, Z., Gao, J., Zhang, Y., Su, X.-D., Li, D., and Liu, C. (2020). Cryo-EM structure of full-length α-synuclein amyloid fibril with Parkinson's disease familial A53T mutation. Cell Res *30*, 360–362. doi:10.1038/s41422-020-0299-4.

48. Sun, C., Zhou, K., DePaola, P., Shin, W. S., Hillyer, T., Sawaya, M. R., Zhu, R., Peng, C., Zhou, Z. H., and Jiang, L. (2023). Cryo-EM structure of amyloid fibril formed by α-synuclein hereditary A53E mutation reveals a distinct protofilament interface. J Biol Chem *299*, 104566. doi:10.1016/j.jbc.2023.104566.

49. Zhao, K., Li, Y., Liu, Z., Long, H., Zhao, C., Luo, F., Sun, Y., Tao, Y., Su, X.-d., Li, D., Li, X., and Liu, C. (2020). Parkinson's disease associated mutation E46K of α-synuclein triggers the formation of a distinct fibril structure. Nature Commun *11*, 2643. doi:10.1038/s41467-020-16386-3.

50. Long, H., Zheng, W., Liu, Y., Sun, Y., Zhao, K., Liu, Z., Xia, W., Lv, S., Liu, Z., Li, D., He, K.-W., and Liu, C. (2021). Wild-type α-synuclein inherits the structure and exacerbated neuropathology of E46K mutant fibril strain by cross-seeding. Proc Natl Acad Sc. USA *118*, e2012435118. doi:10.1073/pnas.2012435118.

51. Sun, Y., Long, H., Xia, W., Wang, K., Zhang, X., Sun, B., Cao, Q., Zhang, Y., Dai, B., Li, D., and Liu, C. (2021). The hereditary mutation G51D unlocks a distinct fibril strain transmissible to wild-type α-synuclein. Nature Commun *12*, 6252. doi:10.1038/s41467-021-26433-2.

52. Mammeri, N. E., Dregni, A. J., Duan, P., and Hong, M. (2024). Structures of AT8 and PHF1 phosphomimetic tau: Insights into the posttranslational modification code of tau aggregation. Proc Natl Acad Sci USA *121*I, e2316175121. doi:10.1073/pnas.2316175121.

53. Qi, C., Kobayashi, R., Kawakatsu, S., Kametani, F., Scheres, S. H. W., Goedert, M., and Hasegawa, M. (2024). Tau filaments with the chronic traumatic encephalopathy fold in a case of vacuolar tauopathy with VCP mutation D395G. Acta Neuropathologica *147*, 86. doi:10.1007/s00401-024-02741-x.

54. Qi, C., Lövestam, S., Murzin, A. G., Peak-Chew, S., Franco, C., Bogdani, M., Latimer, C., Murrell, J. R., Cullinane, P. W., Jaunmuktane, Z., Bird, T. D., Ghetti, B., Scheres, S. H. W., and Goedert, M. Tau filaments with the Alzheimer fold in cases with MAPT mutations V337M and R406W (2024). bioRxiv. doi:10.1101/2024.04.29.591661.

55. Fernandez, A., Hoq, M. R., Hallinan, G. I., Li, D., Bharath, S. R., Vago, F. S., Zhang, X., Ozcan, K. A., Newell, K. L., Garringer, H. J., Jiang, W., Ghetti, B., and Vidal, R. (2024). Cryo-EM structures of amyloid-β and tau filaments in Down syndrome. Nature Struct Mol Biol *31*, 903–909. doi:10.1038/s41594-024-01252-3.

56. Ghosh, U., Tse, E., Yang, H., Shi, M., Caro, C. D., Wang, F., Merz, G. E., Prusiner, S. B., Southworth, D. R., and Condello, C. (2024). Cryo-EM structures reveal tau filaments from Down syndrome adopt Alzheimer's disease fold. Acta Neuropathologica Commun *12*, 94. doi:10.1186/s40478-024-01806-y.

57. Shi, Y., Murzin, A. G., Falcon, B., Epstein, A., Machin, J., Tempest, P., Newell, K. L., Vidal, R., Garringer, H. J., Sahara, N., Higuchi, M., Ghetti, B., Jang, M.-K., Scheres, S. H. W., and Goedert, M. (2021). Cryo-EM structures of tau filaments from Alzheimer's disease with PET ligand APN-1607. Acta Neuropathologica *141*, 697–708. doi:10.1007/s00401-021-02294-3.

58. Qi, C., Hasegawa, M., Takao, M., Sakai, M., Sasaki, M., Mizutani, M., Akagi, A., Iwasaki, Y., Miyahara, H., Yoshida, M., Scheres, S. H. W., and Goedert, M. (2023). Identical tau filaments in subacute sclerosing panencephalitis and chronic traumatic encephalopathy. Acta Neuropathologica Commun *11*, 74. doi:10.1186/s40478-023-01565-2.

59. Shi, Y., Ghetti, B., Goedert, M., and Scheres, S. H. W. Cryo-EM structures of chronic traumatic encephalopathy tau filaments with PET ligand flortaucipir (2023). J Mol Biol *435*, 168025 doi: 10.1016/j.jmb.2023.168025

60. Tarutani, A., Lövestam, S., Zhang, X., Kotecha, A., Robinson, A. C., Mann, D. M. A., Saito, Y., Murayama, S., Tomita, T., Goedert, M., Scheres, S. H. W., and Hasegawa, M. (2023). Cryo-EM structures of tau filaments from SH-SY5Y cells seeded with brain extracts from cases of Alzheimer's disease and corticobasal degeneration. FEBS Open Bio *13*, 1394-1404. doi:10.1002/2211-5463.13657.

61. Shi, Y., Zhang, W., Yang, Y., Murzin, A. G., Falcon, B., Kotecha, A., van Beers, M., Tarutani, A., Kametani, F., Garringer, H. J., Vidal, R., Hallinan, G. I., Lashley, T., Saito, Y., Murayama, S., Yoshida, M., Tanaka, H., Kakita, A., Ikeuchi, T., Robinson, A. C., Mann, D. M. A., Kovacs, G. G., Revesz, T.,

Ghetti, B., Hasegawa, M., Goedert, M., and Scheres, S. H. W. (2021). Structure-based classification of tauopathies. Nature *598*, 359–363. doi:10.1038/s41586-021-03911-7.

62. Guerrero-Ferreira, R., Taylor, N. M., Arteni, A.-A., Kumari, P., Mona, D., Ringler, P., Britschgi, M., Lauer, M. E., Makky, A., Verasdonck, J., Riek, R., Melki, R., Meier, B. H., Böckmann, A., Bousset, L., and Stahlberg, H. (2019). Two new polymorphic structures of human full-length alpha-synuclein fibrils solved by cryo-electron microscopy. eLife *8*, e48907. doi:10.7554/eLife.48907.

63. Li, B., Ge, P., Murray, K. A., Sheth, P., Zhang, M., Nair, G., Sawaya, M. R., Shin, W. S., Boyer, D. R., Ye, S., Eisenberg, D. S., Zhou, Z. H., and Jiang, L. (2018). Cryo-EM of full-length α-synuclein reveals fibril polymorphs with a common structural kernel. Nature Commun *9*, 3609. doi:10.1038/s41467-018-05971-2.

64. Tao, Y., Zhao, Q., Liu, C., and Li, D. Formed alpha-synuclein fibrils after incubation with heparin for 1 hour (Hep-remod-2) (Unpublished). doi:10.2210/pdb8hzb/pdb, 10.2210/pdb8hzc/pdb, 10.2210/pdb8hzs/pdb.

65. Burger, D., Kashyrina, M., Lewis, A. J., Nuccio, F. D., Mohammed, I., Seiglière, H. d. L., Heuvel, L. v. d., Verchère, J., Feuillie, C., Berbon, M., Arotçarena, M.-L., Retailleau, A., Bezard, E., Laferrière, F., Loquet, A., Bousset, L., Baron, T., Lofrumento, D. D., Giorgi, F. D., Stahlberg, H., and Ichas, F. 1.94 Å structure of synthetic α-synuclein fibrils seeding MSA neuropathology (2024). bioRxiv. doi:10.1101/2024.07.01.601498.

66. Hojjatian, A., Dasari, A. K. R., Sengupta, U., Taylor, D., Daneshparvar, N., Yeganeh, F. A., Dillard, L., Michael, B., Griffin, R. G., Borgnia, M. J., Kayed, R., Taylor, K. A., and Lim, K. H. (2021). Tau induces formation of α-synuclein filaments with distinct molecular conformations. Biochem Biophys Res Commun *554*,145–150. doi:10.1016/j.bbrc.2021.03.091.

67. Frieg, B., Antonschmidt, L., Dienemann, C., Geraets, J. A., Najbauer, E. E., Matthes, D., de Groot, B. L., Andreas, L. B., Becker, S., Griesinger, C., and Schröder, G. F. (2022). The 3D structure of lipidic fibrils of α-synuclein. Nature Commun *13*, 6810. doi:10.1038/s41467-022-34552-7.

68. McGlinchey, R. P., Ni, X., Shadish, J. A., Jiang, J., and Lee, J. C. (2021). The N terminus of α-synuclein dictates fibril formation. Proc Natl Acad Sci USA *118*, e2023487118. doi:10.1073/pnas.2023487118.

69. Stern, A. M., Yang, Y., Jin, S., Yamashita, K., Meunier, A. L., Liu, W., Cai, Y., Ericsson, M., Liu, L., Goedert, M., Scheres, S. H. W., and Selkoe, D. J. (2023). Abundant Aβ fibrils in ultracentrifugal supernatants of aqueous extracts from Alzheimer's disease brains. Neuron, *111*, 2012–2020.e4. doi:10.1016/j.neuron.2023.04.007.

70. Merz, G. E., Chalkley, M. J., Tan, S. K., Tse, E., Lee, J., Prusiner, S. B., Paras, N. A., De-Grado, W. F., and Southworth, D. R. (2023). Stacked binding of a PET ligand to Alzheimer's tau paired helical filaments. Nature Commun *14*, 3048. doi:10.1038/s41467-023-38537-y.

71. Coughlin, D. G., Dickson, D. W., Josephs, K. A., & Litvan, I. (2021). Progressive Supranuclear Palsy and Corticobasal Degeneration. Advances in Experimental Medicine and Biology, *1281*, 151–176. doi:10.1007/978-3-030-51140-1_11.

72. Zhao, W., Liu, K., Fan, Y., Zhao, Q., Tao, Y., Zhang, M., Gan, L., Yu, W., Sun, B., Li, D., Liu, C., and Wang, J. (2024). Cryo-EM structures reveal variant Tau amyloid fibrils between the rTg4510 mouse model and sporadic human tauopathies. Cell Discovery *10*, 27. doi:10.1038/s41421-023-00637-w.

73. Schweighauser, M., Shi, Y., Tarutani, A., Kametani, F., Murzin, A. G., Ghetti, B., Matsubara, T., Tomita, T., Ando, T., Hasegawa, K., Murayama, S., Yoshida, M., Hasegawa, M., Scheres, S. H. W., and Goedert, M. (2020). Structures of α-synuclein filaments from multiple system atrophy. Nature *585*, 464–469. doi:10.1038/s41586-020-2317-6.

74. Zhang, W., Falcon, B., Murzin, A. G., Fan, J., Crowther, R. A., Goedert, M., and Scheres, S. H. (2019). Heparin-induced tau filaments are polymorphic and differ from those in Alzheimer's and Pick's diseases. eLife *8*, e43584. doi:10.7554/eLife.43584.

75. Wilkinson, M., Xu, Y., Thacker, D., Taylor, A. I. P., Fisher, D. G., Gallardo, R. U., Radford, S. E., and Ranson, N. A. (2023). Structural evolution of fibril polymorphs during amyloid assembly. Cell *86*, 5798–5811.e26. doi:10.1016/j.cell.2023.11.025.

76. Lövestam, S., Li, D., Wagstaff, J. L., Kotecha, A., Kimanius, D., McLaughlin, S. H., Murzin, A. G., Freund, S. M. V., Goedert, M., and Scheres, S. H. W. (2024). Disease-specific tau filaments assemble via polymorphic intermediates. Nature *625*, 119–125. doi:10.1038/s41586-023-06788-w.

77. Van Rossum, G., and Drake Jr, F. L. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam (1995).

78. The wwPDB Consortium (2024). EMDB—the Electron Microscopy Data Bank. Nucleic Acids Research. doi:10.1093/nar/gkad1019.

79. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria (2022). https://www.R-project.org/.

80. Galili, T. (2015). Dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. doi:10.1093/bioinformatics/btv428.

81. Tao, Y., Sun, Y., Lv, S. *et al.* Heparin induces α-synuclein to form new fibril polymorphs with attenuated neuropathology. *Nat Commun* **13**, 4226 (2022). https://doi.org/10.1038/s41467-022-31790-7

82. Radusky, L. G., and Serrano, L. (2022). PyFoldX: enabling biomolecular analysis and engineering along structural ensembles. Bioinformatics *31*, 3718–3720. doi:10.1093/bioinformatics/btac072.

83. Grosjean, P., Ibanez, F., and Etienne, M. pastecs: Package for Analysis of Space-TimeEcological Series (2024). https://cran.r-project.org/web/packages/pastecs/index.html.

84. Wilmot, C., and Thornton, J. (1990). β-Turns and their distortions: a proposed new nomenclature. Prot Eng, Design and Selection *3*, 479-493. doi:10.1093/protein/3.6.479.

85. Richardson, L. (2007). Beautiful soup documentation. https://pypi.org/project/beautifulsoup4/.

86. RStudio Team. RStudio: Integrated Development Environment for R. RStudio, PBC Boston, MA (2022). http://www.rstudio.com/.

87. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York (2016). ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

88. Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Systems, 1695. https://igraph.org.

89. Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D., and M ¨uller, K. igraph: Network Analysis and Visualization in R (2025). https://CRAN.R-project.org/package=igraph. doi:10.5281/zenodo.7682609 r package version 2.1.4.