scientific reports

Check for updates

OPEN

Geometry based prediction of tau protein sites and motifs associated with misfolding and aggregation

Masumi Sugiyama[1,2], Kenneth S. Kosik[3] & Eleni Panagiotou[4]✉

Recent studies of tau proteins point to specific sites or motifs along the protein related to its misfolding and aggregation propensity, which is associated with neurodegenerative diseases of structure-dependent pathology. In this manuscript we employ topology and geometry to analyze the local structure of tau proteins obtained from the Protein Data Bank. Our results show that mathematical topology/geometry of cryo-EM structures alone identify the PGGG motifs, and the PHF6(*) motifs as sites of interest and reveal a geometrical hierarchy of the PGGG motifs that differs for 3R+4R and 4R tauopathies. By employing the Local Topological Free Energy (LTE), we find that progressive supranuclear palsy (PSP) and globular glial tauopathy (GGT) have the highest LTE values around residues 302–305, which are inside the jR2R3 peptide and in the vicinity of the 301 site, experimentally associated with aggregation. By extending the LTE definition to estimate a global topological free energy, we find that the jR2R3 peptide of PSP and GGT, has in fact the lowest global topological free energy among other tauopathies. These results point to a possible correlation between the global topological free energy of parts of the protein and the LTE of specific sites.

Keywords Neurodegenerative disease, Tau protein, Structure, Topology, Geometry, Aggregation

Neurodegenerative diseases affect millions of people worldwide[1] and are often characterized by the presence of neurofibrillary lesions in the cerebral cortex. Neurodegenerative diseases that are characterized by the deposition of tau protein fibrillar aggregates are referred to as tauopathies. These include Alzheimer's disease (AD), primary age-related tauopathy (PART), chronic traumatic encephalopathy (CTE), Pick's disease (PiD), corticobasal degeneration (CBD), argyrophilic grain disease (AGD), progressive supranuclear palsy (PSP), and globular glial tauopathy (GGT). Cryo-electron microscopy (cryo-EM) has enabled the identification of different tau fibril shapes associated with each of those diseases[2]. The findings by cryo-EM are consistent with the hypothesis that fibril shapes are the determinant of the selectivity of conformation-dependent antibodies, such as GT-7, GT-38, and 18F12 that can distinguish AD-tau from pathological tau of other tauopathies[3,4]. Hence, verifying and understanding the mechanisms of tau fibril shape formation and propagation is crucial for therapeutic development[5,6]. In this manuscript we employ topology and geometry in order to identify features of tau proteins that may be key to tau fibril shape formation and aggregation.

Tau protein is a microtubule-associated protein that binds to microtubules (MTs) and regulates their dynamics by stabilizing and promoting their assembly[7]. Although it most notoriously binds to MTs, it has many binding partners[8], such as filamentous-actin (F-actin)[9,10]. Tau protein can undergo various post-translational modifications (PTMs), including phosphorylation, ubiquitination and truncation[11–13]. Abnormal hyperphosphorylation and other PTMs of tau protein result in decreased binding to MTs and subsequently self-assemble into amyloid fibrils. Tau amino acid (residue) sequence can be divided into several regions[14], see Figure 1a. One region that has attracted attention for the discovery of antibodies[15], is the microtubule-binding repeat region (MTBR) formed by repeats R1 to R4 (31 or 32 residues for each repeat and inter-repeat region) that have similar sequences[16]. Tau protein is present as six isoforms in human brain (ranging from 352 to 441 residues), and they differ by differential inclusion of N1, N2, and R2. Depending on the predominant tau isoforms present, tauopathies can be classified into 3R+4R (approximately having an equal ratio of 3R and 4R isoforms: AD, PART and CTE), 3R (mainly having 3R isoform: PiD) and 4R (mainly having 4R isoforms: CBD, AGD, PSP, GGT and GPT (PSP/GGT hybrid)) tauopathies[14,17]. All the different tau fibril shapes (tau folds) of tauopathies contain the

[1]Department of Mathematics, University of Tennessee at Chattanooga, Chattanooga, TN 37403, USA. [2]International Institute for Sustainability with Knotted Chiral Meta Matter (WPI-SKCM2), Hiroshima University, Higashi-Hiroshima, Hiroshima, Japan. [3]Neuroscience Research Institute and Department of Molecular, Cellular, and Developmental Biology, University of California Santa Barbara, Santa Barbara, CA 93106, USA. [4]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85281, USA. ✉email: Eleni.Panagiotou@asu.edu

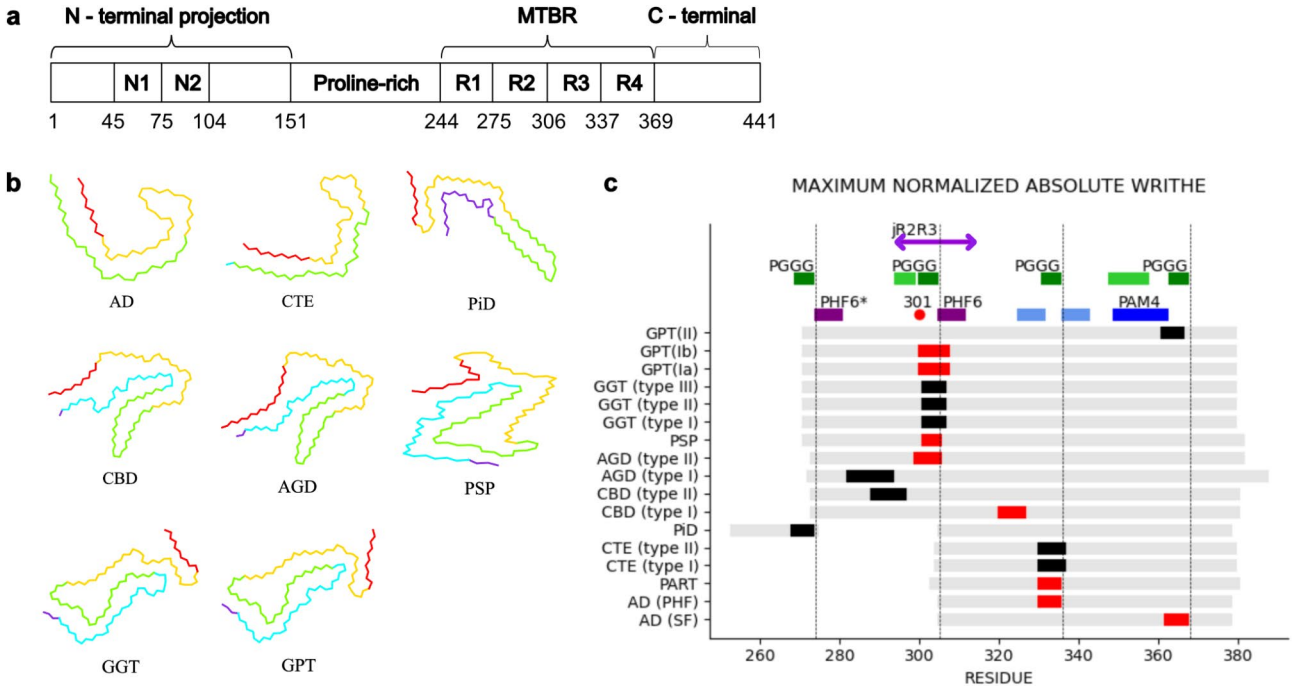

Fig. 1. Tau protein regions, different folds and the locations of maximum normalized absolute Writhe in each fold. (**a**) Tau amino acid sequence and regions of the longest 4R tau isoform (2N4R) consisting of 441 amino acids. Six isoforms differ by differential inclusion of N1, N2, and R2. The microtubule-binding repeat region (MTBR) of 4R tau isoforms comprise all four repeats (R1–R4) while that of 3R tau isoforms are missing R2. (**b**) Cartoon representations of different tau folds seen in tauopathies[2,14,25–29]. Residues in R1–R4 repeats and the C-terminal region are colored purple, blue, green, yellow and red, respectively. Each tau fold consists of the following residues: 306–378 for AD (PDB: *5o3t*), 305–379 for CTE (PDB: *6nwp*), 254–274 and 306–378 for PiD (PDB: *6gx5*), 274–380 for CBD (PDB: *6tjo*), 273–387 for AGD (PDB: *7p6d*), 272–381 for PSP (PDB: *7p65*), 272–379 for GGT (PDB: *7p66*) and GPT (PDB: *7p6a*). The cartoon representations are obtained by projecting the corresponding three-dimensional PDB coordinates to the *xy*-plane. (**c**) The maximum normalized absolute Writhe fragments in tauopathy filaments (black if positive, red if negative). The gray bars indicate the range of residues in filaments. The columns created by dotted lines from the left indicate the R1, R2, R3, R4 repeats and C-terminal region, respectively. The top rows indicate the locations of the jR2R3 motif (295–313)[21] shown by a purple arrow, the PGGG motifs[18] colored in green, the stabilizing fragments (295–300, 349–356)[19] colored in light green, the PHF6* (275–280) and PHF6 (306–311) motifs[30,31] colored in purple, the peptides (326–331)[32] and (337–342)[19] colored in light blue, the PAM4 (350–362) motif[32] colored in blue and the 301 mutation site[22] shown by a red dot. We see a distinction between 3R+4R and 4R tauopathies. Differences between 3R+4R, 3R and 4R tauopathies persist when examining maximum normalized absolute Writhe values restricted in the 3R, 4R repeats and C-terminal region and restricted in the 4R repeat and C-terminal region (see Supplementary Fig. S2).

common ordered core which includes the R3 and R4 repeats in the MTBR plus an additional 10–13 residues in the C-terminal region. Cartoon representations of tau folds are shown in Figure 1b.

Studies have revealed the importance of specific regions/motifs of the protein for the formation of fibrils. For example, amyloid-forming motifs such as two hexapeptides, PHF6* (^{275}VQIINK^{280}) and PHF6 (^{306}VQIVYK^{311}), control tau assembly into fibrils by interacting with other parts of the sequence[18,19]. These amyloid-forming motifs are preceded by PGGG sequence motifs shown to be the center of hairpin-like structures across each interface of the MTBR of tau protein[20]. Recently, a 19 amino acid segment (jR2R3 peptide) in the junction between the R2 and R3 repeats of the MTBR was shown to form fibrils and adopt a fold characteristic of 4R tauopathies[21]. Moreover, it has been shown that specific mutations can promote aggregation. The most common such mutation is the P301L mutation[22–24].

Methods from mathematical topology and geometry can provide more detailed and quantitative characterization of protein structural complexity both globally and locally[33–43] than descriptive methods. Recent work has shown a connection between novel topological measures and protein kinetics, namely, that experimental protein folding rates correlate with the topological structural complexity of the native state of simple, 2-state proteins without knots or slipknots[39]. Prior work on tau proteins has shown that, even though tau filaments of tauopathies are unknotted and pairwise unlinked when stacked in fibrils, they have non zero topological signatures for all these measures that enable the comparison and classification of tau filaments based on their subtle quantitative differences, which agree with tau pathology and clinical symptoms[44]. Moreover, it

was shown that these tau filaments each have a weak but observable knotoid structure, with PSP being the most prevalent.

In this manuscript, we employ the Writhe to quantify and compare the local structural complexity of tau filaments of tauopathies and identify specific sites of topological/geometrical interest that may be key to the formation of aggregates. The Writhe is a measure of interwinding of a filament around itself[45]. When applied locally, it captures geometrical characteristics of a part of a protein. We also employ data analysis methods to understand the structures of tau filaments in the context of the ensemble of folded proteins in the Protein Data Bank (PDB)[46]. Namely, the Local Topological Free Energy (LTE), introduced in[47,48], is used to assign a measure of rarity of a local conformation of 4 residues of a protein with respect to the folded state ensemble. LTE has recently been connected to stability of proteins via studies of protein evolution and conformational dynamics[49]. We also extend this method to account for the global topological free energy of longer parts of the protein.

Our results demonstrate that the Writhe and the LTE can predict sites that have been experimentally shown to be important for protein structure and function. In fact, maximum normalized absolute Writhe fragments are identified near the PGGG motifs of tauopathies and reveal a hierarchy of those motifs, in terms of geometrical signal, that is different for 3R+4R and 4R tauopathies. Overall, 3R+4R tauopathies have lower LTE values compared to 4R tauopathies, indicating that local conformations of 3R+4R tauopathies are more stable. PSP is the only tauopathy with a high LTE conformation (indicative of a rare/unstable structure in the PDB) at residues 302–305, next to the critical 301 mutation site and the PHF6 motif in the jR2R3 peptide. The jR2R3 peptide in 3R+4R tauopathies has, however, higher global topological energy than 4R tauopathies, suggesting it is more stable in 4R tauopathies. Finally, we find that the R2 repeat of tau proteins bound to MTs and to F-actin has higher LTE values than in tauopathies, but it has a lower global topological free energy as a repeat. The latter results suggest that binding sites might be detectable via high LTE geometries (something which we corroborate with data on known antibodies) and that high LTE conformations may result in lower global topological energies, possibly stabilizing larger parts of a protein.

The manuscript is organized as follows: The Results section presents the results of this study. The Methods section lists the PDB structures and datasets used, and gives the rigorous mathematical defitions of the structural parameters used in this study. The Discussion section summarizes our conclusions. Additional data can be found in the Supplementary information.

Sites and motifs of interest in tau protein

In this paragraph we present more background information about local sites and motifs of interest in tau filaments that are experimentally known to be relevant for protein aggregation and will be helpful as a context for our analysis. A PGGG motif creates a turn or a short loop between two strands, stabilizing this resulting hairpin structure[20,50,51]. These hairpin structures are important for the formation of oligomers which ultimately become neurofibrillary tangles (NFTs)[52,53]. The PGGG motifs are located at the end of each repeat: (1) residue 270–273 in the R1 repeat, (2) 301–304 in the R2 repeat, (3) 332–335 in the R3 repeat, (4) 364–367 in the R4 repeat[18,20]. In the vicinity of these PGGG motifs, amyloid-forming motifs are present. We consider amyloid-forming motifs that have been experimentally shown to form amyloid fibrils: (1) PHF6* (residues 275–280), (2) PHF6 (306–311), (3) jR2R3 (295–313), (4) ^{326}GNIHHK^{331}, (5) ^{337}VEVKSE^{342} and (6) PAM4 (350–362). The PHF6* and PHF6 motifs located at the beginning of R2 and R3 repeats, respectively, are essential for tau aggregation[30,31]. Although the PHF6(*) (PHF6* and PHF6) motifs spontaneously aggregate in solution in contrast to the full length tau protein, which is a highly soluble protein, the PHF6* is thought to be the stronger driver of aggregation[16,54–56]. It was shown that the PHF6 motif becomes less stable when a disease-associated mutation is present such as the P301L mutation[22–24], increasing susceptibility to conformational changes that expose the motif and drive tau aggregation[20]. While standalone PHF6(*) motifs can assemble into fibrils, these fibrils are unable to propagate monomeric tau by templating[57–59]. To achieve propagation by templated seeding, these motifs need to be stably held in a specific orientation with a counterstrand when forming fibrils[60]. It was reported that the 19 residue peptide (residues 295–313), jR2R3, holds the PHF6 in place by a counterstrand, can readily form fibrils and template monomeric tau for fibril propagation[21]. In contrast to the PHF6(*) motifs which adapt very similar conformations across tau folds, PAM4 (polymorphic amyloid segment in repeat 4: ^{350}VQSKIGSLDNITH^{362}) exhibits four unique conformations, influencing the organization of different tau folds[32]. In addition to PAM4 which was initially predicted by amyloid prediction algorithms[32], the two peptides, ^{326}GNIHHK^{331} in the R3 repeat and ^{337}VEVKSE^{342} in the R4 repeat, have also been predicted and experimentally shown to form fibrils[19,32].

Results

In this work, we analyze local structures of tau filaments of tauopathies using the Writhe and the LTE. We also introduce and use the Repeat Topological Free Energy (RTE) and the Peptide Topological Free Energy (PTE). Based on the local structures of tau filaments, we identify regions of geometrical interest along the filaments in relation with experimentally studied motifs and the 301 mutation site associated with misfolding and aggregation. We point out that the results in this section are all obtained from PDB structures, some of which, come from different laboratories (see references and data in Supplementary Table S5).

Maximum normalized absolute writhe fragments in tauopathies

We first examine the local Writhe of each tauopathy filament. We scan along each filament at intervals of varying length and compute the normalized absolute Writhe of each resulting fragment (normalized by the fragment length) to determine where, along the filament, the maximum normalized absolute Writhe is stored; we denote Wr_{max} (see Materials and Methods for the definition, and Supplementary Fig. S1 for examples of Wr_{max} fragments).

The results are shown in Fig. 1c. The columns in the figure show the R1, R2, R3, R4 repeats and the C-terminal region from the left. The locations of Wr_{max} fragments of tauopathy filaments are highlighted in black, resp. red, if the sign of Wr_{max} is positive, resp. negative, which is indicative of a right-handed, resp. left-handed, local conformation. Wr_{max} fragments, which range from 4 to 11 residues long, are found at the following residues: 363–367 for AD (SF), 331–335 for AD (PHF) and PART, 331–336 for CTE, 269–273 for PiD, 321–326 for CBD (type I), 289–296 for CBD (type II), 283–293 for AGD (type I), 300–305 for AGD (type II), 302–305 for PSP, 302–306 for GGT, 301–307 for GPT (type Ia and Ib) and 362–366 for GPT (type II). Supplementary Table S1 shows the corresponding maximum normalized Writhe values.

Interestingly, Wr_{max} fragments are located in a few common regions among tauopathies. These locations reveal that Wr_{max} fragments are highly likely to be part of any one of the PGGG motifs. This result is reasonably expected due to turns or short loops formed by the PGGG motifs which may cause increased Writhe[51]. Interestingly however, the PGGG motif that is part of the site identified by Wr_{max} depends on the predominant isoform composition of tauopathy, namely 4R tauopathies have Wr_{max} near the 2nd PGGG motif, while 3R+4R tauopathies have their Wr_{max} near the 3rd PGGG motif. Wr_{max} fragments of 4R tauopathies are also located inside the jR2R3 motif which is known to adopt a fold characteristic of 4R tauopathies[21].It is noteworthy that these results persist for the energy minimized structures as well (results shown in Supplementary Table S6). When the search of Wr_{max} fragment is restricted to the R3, R4 repeats and the C-terminal region or the R4 repeat and the C-terminal region for each tauopathy filament, Wr_{max} fragments continue being highly likely to be part of the PGGG motif(s) present (see Supplementary Table S2, S3 and Supplementary Fig. S2). As the search region of Wr_{max} fragment changes, there seems to be a hierarchy of the PGGG motifs that are part of the Wr_{max} fragment, which is different between 3R, 3R+4R and 4R tauopathies. Namely, when restricted to the R3, R4 repeats and the C-terminal region for 3R+4R tauopathies, majority of Wr_{max} is identified overlapping with the 3rd PGGG motif, while for 4R tauopathies it is either overlapping the 4th PGGG motif or in the region between the 2nd and 3rd motif. When restricted to the R4 repeat and C-terminal region, almost all tauopathy filaments have Wr_{max} in a region overlapping with the 4th PGGG motif.

There are notable exceptions in these trends. When the search region of each Wr_{max} fragment is the entire filament, Wr_{max} fragments of CBD (type I and II) and AGD (type I) are located outside of the PGGG motifs, but in the vicinity of (at most one residue away from) the minimal fragment ^{295}KDNIKHV^{300} of the jR2R3 peptide. This minimal fragment is reported to stabilize and hold in place the PHF6 motif[60]. Moreover, this minimal fragment is one of two fragments that interact with the PHF6 motif in CBD and AGD[19]. It is possible that the Wr_{max} fragment of CBD (type I), which is not in the vicinity of any one of the stabilizing motifs, may also point to an important stabilizing region for this filament.

Finally, we notice that Wr_{max} fragment of PiD, which is a 3R tauopathy, encompasses a region that includes the 1st PGGG motif whose conformation has been shown to be influenced by the E264G mutation, and the conformational change was shown to accelerate aggregation[61]. When the search region is restricted to the 4R repeat and the C-terminal region, PSP is the only filament among 3R+4R and 4R tauopathies that does not overlap with the 4th PGGG motif. Its Wr_{max} fragment is part of a stabilizing fragment ^{349}RVQSKIGS^{356} with which the PHF6 motif interacts in PSP[19].

Local topological free energy of tauopathies

To further analyze the local conformations of tauopathy filaments, we employ the LTE in Writhe, which was introduced in[47] (see Materials and Methods for the definition). The LTE distribution (see Supplementary Fig. S3) is defined based on the Writhe values of local conformations of 13,193 proteins in the PDB. According to this LTE distribution of the ensemble of folded proteins, the LTE of a local protein conformation indicates the degree of rarity of its conformation, consisting of 4 residues, relative to local conformations of proteins in the folded ensemble. A higher (lower) LTE of a protein fragment indicates a more rare (common) local conformation in the folded ensemble. LTE can be thought of as a free energy estimator and differences in LTE values as free energy differences[47,62,63]. An LTE value greater than 3.07 indicates values outside the 75th percentile of the LTE distribution, and it is called a high LTE. A local conformation of a protein attaining a high LTE is considered to be rare in the folded ensemble.

The LTEs of each tauopathy filament are shown in Fig. 2a (and Supplementary Fig. S4). Each data point denotes the LTE of a local conformation of 4 consecutive residues starting at that residue in the filament. PSP is the only filament that attains a high LTE at residues 302–305, which is indicative of a rare local conformation. This high LTE fragment, which is inside the jR2R3 motif, overlaps with the ^{301}PGGG^{304} motif and is adjacent to the most common 301 mutation site and the PHF6 motif. In addition to PSP, the filaments of GGT attain their highest LTE values around these sites of interest. This cluster of highest LTEs of PSP and GGT are in the vicinity of residue 305 where they are observed to be oriented inwardly within the hairpin structure[21]. In contrast, GPT (type Ib and II) attains its minimum LTEs near residue 305, suggesting a more stable conformation there. This location of GPT is observed to be oriented outwardly from the hairpin structure. It is interesting to compare these results with DEER experiments showing that jR2R3 P301L fibrils most closely resemble GPT distances between a pair of residues selected for investigation of the adapted fold. Another cluster of relatively high LTEs for (PART, PiD, PSP, CBD (type II), GGT, GPT (type Ia)) is formed around the residue 345. These two clusters of high or relatively high LTE values along tauopathy filaments are found near the strong tau fold stabilizing motifs, PHF6 and PAM4[32].

We note that the PGGG motifs (which appeared in the previous section), with the exception of ^{301}PGGG^{304} motif in PSP, are not in high LTE. To further examine the local conformation of PGGG motifs we analyzed our culled protein sample for the Writhe values of PGGG motifs only, and computed the LTE of PGGG motifs in tauopathies based solely on those values. This led to a small sample of 58 values with respect to which the LTE was defined. Based on that sample, we find that approximately 40% of the PGGG motifs in tauopathies are in

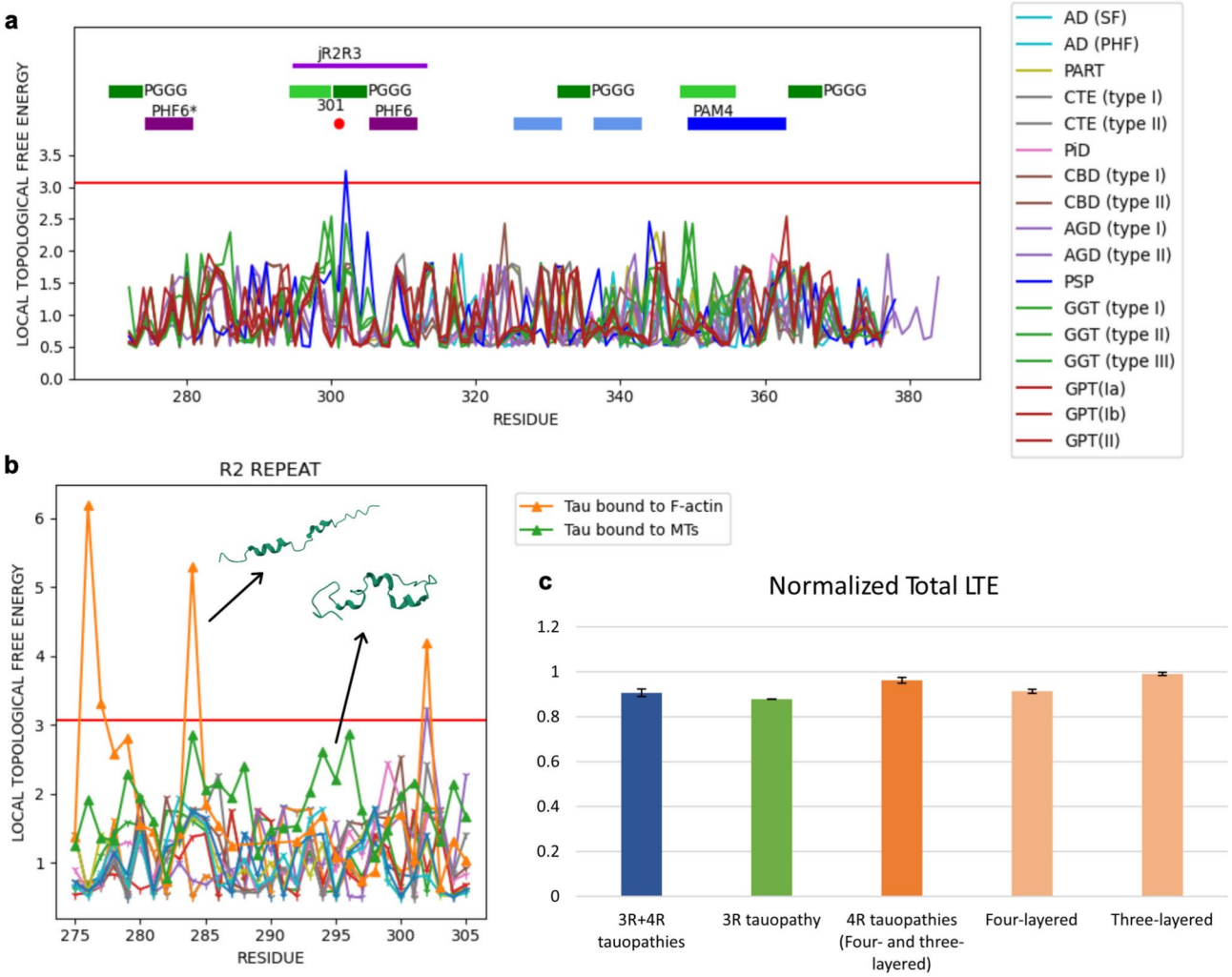

Fig. 2. The Local Topological Free Energy (LTE) of different tau protein folds. (**a**,**b**) Each data point is the LTE of a local conformation of 4 consecutive residues starting at that point. The horizontal line at $\Pi(p) = 3.07$ corresponds to the 75th percentile of LTEs of the ensemble of folded proteins in the culled sample of the PDB. (**a**) The LTE along tauopathy filaments. The top rows are as in Fig. 1c. PSP has its high LTE inside the jR2R3 motif, being near the PHF6 motif and the 301 mutation site. Similarly, GGT has some of its highest LTE values in that region. (**b**) The LTE along the R2 repeat of tauopathies and tau filaments bound to MTs (PDB: *2mz7*) and to F-actin (PDB: *5n5a* and *5n5b*). We see that bound tau proteins have higher LTEs. Some of the high or maximum LTE conformations are at or near binding sites. (**c**) The average normalized total LTE of 3R+4R (navy), 3R (green) and 4R (dark orange) tauopathies. Within 4R tauopathies, the orange light bars indicate the average normalized total LTE of the four-layered and the three-layered 4R tauopathies from the left, respectively. We see that 3R tauopathy is followed by 3R+4R and then by 4R tauopathies in terms of their normalized total LTE, which may imply the opposite order in terms of stability. Three-layered 4R tauopathies have higher normalized total LTE than four-layered. Each error bar indicates a standard error of different tau filaments.

high LTE relative to other PGGG motifs (outside the 75th percentile of the LTE distribution). This indicates that PGGG motifs in tauopathies adopt disproportionally uncommon conformations which may be related to the global structure of the protein.

We also compare tauopathies in terms of their (normalized) total LTEs, each of which is the average sum of all LTE values along filaments normalized by the length of the filament. Overall, we see that 3R+4R and 3R tauopathies have lower total LTE than 4R tauopathies, see Fig. 2c, suggesting that conformations of 3R+4R and 3R tauopathies are locally more stable than 4R tauopathies. Additionally, within 4R tauopathies, four-layered tauopathies have lower total LTE than three-layered, thus a higher local stability of the four-layered tauopathies than the three-layered is likely. Pick's disease, a 3R tauopathy, which has the lowest total LTE, is not a common tauopathy, but it is interesting to point out that AD, a 3R+4R tauopathy with low total LTE, is the most prevalent in the population (affects 10.9% of people age 65 and older[64]) than all 4R tauopathies[65–72]. Additionally, AGD, a

four-layered 4R tauopathy is more common than PSP, a three-layered 4R tauopathy[73], in agreement with their total LTE ranking from lower to higher values.

Binding sites and high LTE

The R2 repeat is one of the four repeats in the MTBR, and it is present in 4R tauopathy filaments. Unlike the R3 repeat, which is very rigid, the R2 repeat is more flexible and can produce different conformations, depending on environmental conditions such as temperature[74]. It is noteworthy that the total LTE of the R2 repeat is the highest among the R1, R3 and R4 repeats for taopathies (see Supplementary Fig. S5), in agreement with this being the more flexible/less stable repeat than the R3 repeat.

Figure 2b shows the LTEs of 4R tauopathy filaments along the R2 repeat, as well as the LTEs of tau filaments bound to MTs (PDB: *2mz7*) and to F-actin (PDB: *5n5a* and *5n5b*). Each data point in the figure corresponds to the initial residue of 4 consecutive residues of a filament. Since each PDB file of tau filament bound to MTs and to F-actin contain 20 models, the average LTEs over the 20 model are used for analyses.

Overall, we see that the LTEs of bound tau filaments attain higher values than those of tauopathies. This suggests that the local conformations of misfolded tau R2 repeats are topologically preferred than those of tau filaments bound to MTs and to F-actin.

The high LTE fragments of tau filaments bound to F-actin are found at the following residues: 276–280 (helix), 284–287 and 302–305. The helical high LTE fragment is part of the PHF6* motif and also part of one of the two binding sites (residues 277–283) through which F-actin interacts[10]. For tau protein bound to MTs, the R2 repeat is one of small groups of residues that that bind tightly to MTs[8]. Although there are no high LTEs of tau filament bound to MTs, its maximum LTE conformation is found at residues 296–299 which is immediately before one of two binding regions (residues 300–310) shown to fold into a hairpin conformation upon binding to MTs. These results point to binding sites possibly be reflected by higher LTEs.

To test the hypothesis that binding sites being at relatively high LTE conformations, we examine the LTEs of known tau antibodies relative to their binding sites. We use renumbered structure (Chothia) files in PDB format for antibodies, and the locations of CDRs for each antibody are obtained from the Structure Antibody Database (SAbDab)[75].

The average probability of an antibody residue in high LTE fragments being a binding site is higher than that of any residue being a binding site (23% versus 16%), see Fig. 3 (and Supplementary Table S4 for the probability of each antibody and Supplementary Fig. S6 for examples of high LTE fragments of antibodies). This result suggests that the high LTE fragment of an antibody could be used to point to a location of its antigen binding site.

Among the antibodies we examined, VHH Z70 is the only antibody of single variable domain on a heavy chain antibody (nanobody). VHH Z70 antibody precisely binds to the PHF6 motif and inhibits tau aggregation in heparin-induced assays in vitro[56]. The interaction between the CDRH3 (residues 95–102) of the antibody and the PHF6 motif is stabilized by the 5 intermolecular hydrogen bonds[56]. The antibody has two high LTE conformations, and the maximum LTE of those is located inside the CDRH3 and captures one of the 5 intermolecular hydrogen bonds.

Global topological free energy

In order to analyze the more global/topological information of a filament relative to conformations of proteins in the folded ensemble, we introduce the global topological free energy (see Materials and Methods for more information). In particular, we focus on the global topology of the repeats and of the jR2R3 peptide. Instead of protein fragments of 4 residues, we use fragments of 31 residues, which is the average length of repeats in the MTBR, to compute the Repeat Topological Free Energy (RTE) and 19 residues, which is the length of the jR2R3 peptide (residues 264–313 for 3R tauopahty and 295–313 for 4R tauopathies), to compute the Peptide Topological Free Energy (PTE), respectively (see Materials and Methods for more information, and Supplementary Fig. S7 and S8 for the RTE and PTE distributions, respectively). Since the Writhe accounts for cancellations due to sign, which may be influenced by the presence of secondary structure elements, we instead use the average Crossing

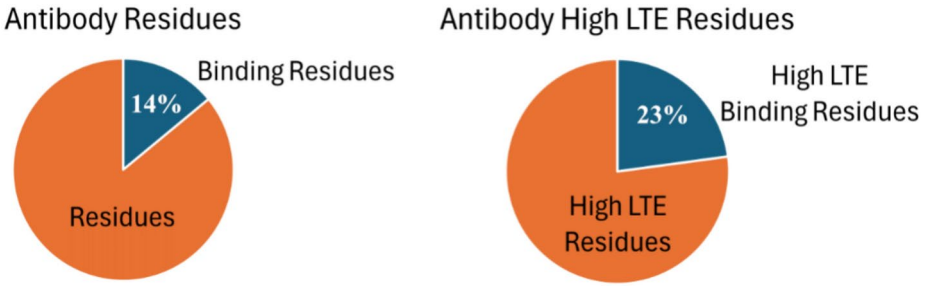

Fig. 3. Binding residues in known antibodies and their LTE. (Left) The average percentage distribution of residues of an antibody that are binding sites. (Right) The average percentage distribution of high LTE residues of an antibody that are binding sites. We hypothesize that relatively high LTE values can help predict binding sites. The standard error of binding residues and of binding high LTE residues in antibodies are 0.008 and 0.029, respectively.

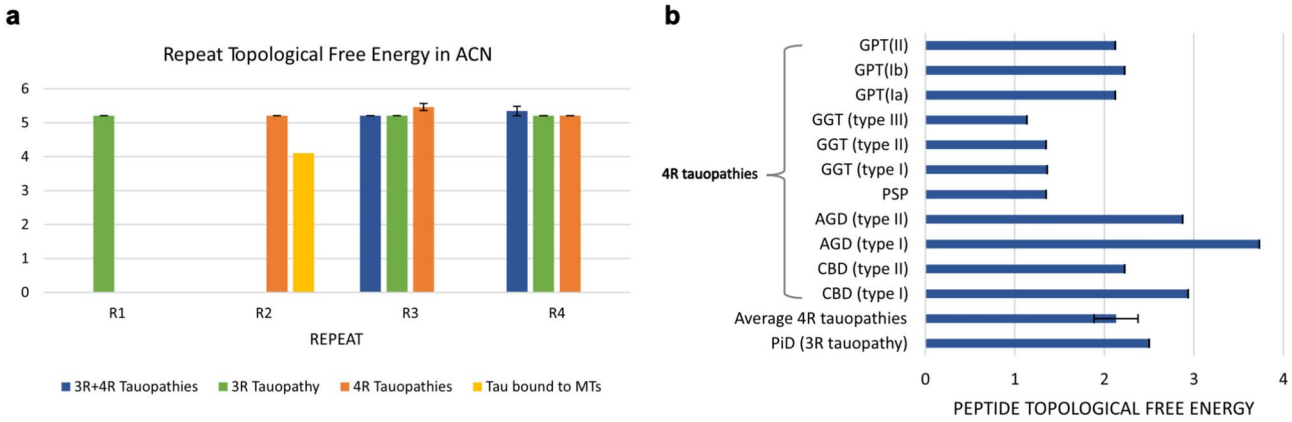

Fig. 4. The global topological energy of repeats and of the jR2R3 peptide in different tau protein folds. (**a**) The average Repeat Topological Free Energy (RTE) in ACN for 3R (green), 3R+4R (navy) and 4R (orange) tauopathies, and the RTE for a tau filament bound to MTs (yellow). Each error bar indicates a standard error of different tau filaments. We see that the global topological free energy of bound proteins is less than that of tauopathies. (**b**) The Peptide Topological Free Energy (PTE) in ACN for 3R and 4R tauopathies. The average PTEs of ACN for 4R tauopahties are approximately 2.13. In comparison, the PTEs of ACN for 3R tauopathy are approximately 2.50. We see that the jR2R3 repeat of 4R tauopathies has lower global topological free energy than the 3R tauopathy, with PSP and GGT having the lowest.

Number (ACN) as a meaningful geometrical/topological parameter for the global topological energy, which will not be affected by the sign (see Materials and Methods for more information).

Topological free energy of repeats

Our results in Fig. 4a show similar average RTEs of the R3 and R4 repeats between 3R+4R, 3R and 4R tauopathies, suggesting similar stability for their repeat conformations. The lower RTE of the R2 repeat of tau filaments bound to MTs than that of 4R tauopathies implies that the R2 repeat conformation of bound tau filament is topologically more stable. Since our results earlier showed that local R2 repeat conformations of 4R tauopathies are more favorable, we hypothesize that the global R2 repeat conformation of tau filament bound to MTs compensates for its unfavorable local conformations.

Within 4R tauopathies, our result shows overall similar average RTEs between the three- and four-layered tauopathies, suggesting similar stability for their repeat conformations (see Supplementary Fig. S9).

Topological free energy of the jR2R3 peptide

Our results in Fig. 4b show that the 19 residue peptides of all tauopahy filaments are in high PTE conformation, attaining values greather than 75th percentile of the PTE distribution. In fact, the peptides of PiD, CBD (type I), AGD and GPT (type Ia) each have a PTE value greater than 90th percentile of the PTE distribution. This implies that the peptide of these tauopahies attain rare conformations than most of proteins in the folded ensemble. PSP and GGT have lower PTEs in ACN among 4R tauopathy filaments, suggesting a more stable conformation of the peptide than other tauopathies. Overall, PiD, a 3R tauopathy, has higher PTE than the average PTE of 4R tauopathies.

Discussion

Neurodegenerative diseases like Alzheimer's, are associated with the misfolding and aggregation of tau proteins. As pathology appears to be structure-dependent, structure-based therapeutics have been proposed. Recent studies of misfolding and aggregation of tau proteins point to specific sites or fragments of interest along the protein that are important for misfolding and aggregation.

In this manuscript we employ topology and geometry in order to rigorously quantify and compare the local structural complexity of tau proteins obtained from the PDB. We demonstrate that mathematical topology and geometry alone and only by using structures from the PDB predicts sites at or near the most important motifs from experiments. More precisely, sites near the PGGG motifs are found to be part of or intersecting the fragments with maximum Writhe in tauopathies. Moreover, the local structures of tau filaments can be classified according to their Writhe, revealing a hierarchy of sites containing the PGGG motifs that differentiates 3R and 4R from 3R+4R tauopathies. For PSP, AGD and CBD, our mathematical analysis predicts important sites outside the PGGG motifs, near known stabilizing fragments. It is interesting to point out that these diseases are associated to different pathologies; tufted astrocytes for example are unique to PSP and astrocytic plaques unique to CBD.

To further identify local conformations of interest, we employ the local and global topological free energy. More precisely, by analyzing a culled dataset of 13,193 proteins from the PDB (2,929,754 local writhe values), we define the Local Topological Free Energy (LTE)[47]. A local protein conformation of 4 residues is assigned an LTE value which can be interpreted as a measure of rarity of its geometry in the folded state ensemble. By employing

the LTE we see that 3R+4R tauopathies have lower total LTE than 4R tauopathies, meaning that conformations of 3R+4R tauopathies are geometrically more stable. Among 4R tauopathies, four-layered tauopathies have lower total LTE than three-layered tauopathies. PSP is the only tauopathy with a high LTE conformation (a value of LTE that is outside the 75th percentile of the PDB) at 302–305, which is between the 301 site and the PHF6 motif. The 301 site has been shown to be an important site for mutations that lead to aggregation[23,24,76], while the PHF6 motif is also shown experimentally to be an important driver for aggregation[30,31].

The R2 repeat of tau proteins is more flexible than the R3 repeat and known to bind to MTs and to F-actin[7,10]. By comparing the LTEs of the R2 repeat of bound tau proteins with those of tauopathies, we find that tauopathies have lower LTEs, implying that the conformations of misfolded R2 repeat are geometrically locally more stable. In fact, tau protein bound to F-actin has multiple high LTE conformations, including at the 302–305 fragment, suggesting that high LTE values in this case may reflect binding sites. To further explore this idea, we use the LTE to identify binding sites in known tau antibodies. Our results show that binding occurs with higher probability at high LTE sites of an antibody. These results point to a possibility of screening for binding sites by using the LTE.

To make a connection of local conformations to global conformations of the protein, we introduce a method to obtain a global topological free energy. By analyzing the same culled data set of the PDB with a scanning window size of 31 residues, which is the length of average repeats in tauopathies, we compute the Repeat Topological Free Energy (RTE). The RTE analysis shows that, even though bound tau proteins have higher LTEs, the R2 repeat of tau protein bound to MTs has lower RTE. These results may suggest that the global R2 repeat conformation compensates for its unfavorable local conformations.

By analyzing a culled data set with a scanning window of 19 residues, which is the length of the jR2R3 peptide, we obtain the Peptide Topological Free Energy (PTE). We find that the jR2R3 peptide of PSP and GGT has lower PTE, implying that these 19 residue conformations of these tauopathies are more stable. Moreover, we find that the PTE of 4R tauopathies is on average smaller than that of the 3R tauopathy. This is in agreement with experimental results that show the jR2R3 peptide preferably forms 4R tauopathy fibrils, as their conformation seems to be more accessible in the folded state ensemble of proteins.

Overall, our results reveal that topological/geometrical characteristics of static structures of tau filaments alone can predict regions at or near sites that are experimentally shown to be important for the protein folding and aggregation. These mathematical methods not only identify these structures, but can also quantify their complexity in a way that could help understand protein misfolding and aggregation. The topological energy of local structures reveals a complex interplay of local conformations of 4 residues with global conformations of parts of the protein. For example, local conformations with relatively high topological energy are associated with binding which in turn reflects lower topological energy of repeats. In fact, local conformations with high topological energy could provide a mechanism to detect binding sites. Mutations that cause an increase of LTE but a decrease of global topological free energy of tauopathies may increase the propensity of aggregation, while increasing both the LTE and the global topological free energy may destabilize a protein. These results suggest that topological/geometrical methods could be used as a method for designing structure- and site-specific therapeutics.

Methods

Representations of tau filaments

Proteins are represented by their consecutive alpha carbon atoms (CA atoms) as linear open-ended polygonal curves in a three-dimensional space (3-space), which are used as approximations of the protein backbones.

Some tauopathies consist in different types of spatial organization of filaments present. Namely, two different types of AD and CTE are formed from two identical filaments that differ in their interaction, the straight filament (SF) and the paired helical filaments (PHF) for AD, and the type I and type II filaments for CTE. The type I and type II filaments in CBD and AGD consist of a single filament and a pair of identical filaments, respectively. For GGT and GPT, three types of filaments are observed with the type I being composed of a single filament and the type II and type III consisting of two identical filaments.

In this manuscript, the following three dimensional crystal structures from the PDB are used for tau filaments of tauopathies: *5o3t* for AD (SF), *5o3l* for AD (PHF), *7nrq* for PART, *6nwp* for CTE (type I), *6nwq* for CTE (type II), *6gx5* for PiD, *6tjo* for CBD (type I), *6tjx* for CBD (type II), *7p6d* for AGD (type I), *7p6e* for AGD (type II), *7p65* for PSP, *7p66* for GGT (type I), *7p67* for GGT (type II), *7p68* for GGT (type III), *7p6a* for GPT (type Ia), *7p6b* for GPT (type Ib) and *7p6c* for GPT (type II).

Measures of topological characterization of tau filaments

The Writhe of a curve in 3-space provides the degree to which a curve winds around itself. It is defined as the Gauss linking integral over one curve[77]:

Definition 1 (Writhe) For an oriented curve l with an arc-length parametrization γ, the Writhe, Wr, is the double integral over l:

$$Wr(l) = \frac{1}{2\pi} \int_{[0,1]^*} \int_{[0,1]^*} \frac{(\dot{\gamma}(t) \times \dot{\gamma}(s)) \cdot (\gamma(t) - \gamma(s))}{||\gamma(t) - \gamma(s)||^3} \, dtds$$

where $\dot{\gamma}$ denotes the derivative of γ, and the integral runs over $[0,1]^* \times [0,1]^*$ for all $s, t \in [0,1]$ such that $s \neq t$.

The Writhe can have both positive and negative values depending on the orientation of a curve. It can also be expressed as the average algebraic sum of signs of all crossings in a projection of a curve with itself over all possible projection directions.

By taking the absolute value of the integrand of the Writhe, we obtain the Average Crossing Number:

Definition 2 (Average Crossing Number) For an oriented curve l with an arc-length parametrization γ, the Average Crossing Number, ACN, is the double integral over l:

$$ACN(l) = \frac{1}{2\pi} \int_{[0,1]^*} \int_{[0,1]^*} \frac{|(\dot{\gamma}(t) \times \dot{\gamma}(s)) \cdot (\gamma(t) - \gamma(s))|}{||\gamma(t) - \gamma(s)||^3} \, dtds$$

where $\dot{\gamma}$ denotes the derivative of γ, and the integral runs over $[0,1]^* \times [0,1]^*$ for all $s, t \in [0,1]$ such that $s \neq t$.

The ACN is a measure of the number of crossings (without signs) of a curve and can only be a positive real number. The ACN can be expressed as the average sum of crossings in a projection of a curve with itself over all possible projection directions.

The Writhe and ACN are continuous functions of the curve coordinates for both open and closed curves. They can be applied to a protein as a whole or to fragments of it (see Fig. S10 in SI for an example of the Writhe and ACN).

For polygonal curves, the Writhe and ACN each can be expressed as a finite sum of signed geometric probabilities that any two edges cross in any projection direction[78].

Maximum normalized absolute Writhe

The maximum normalized absolute Writhe of a protein is computed by first scanning along the protein at intervals of varying length and computing the absolute Writhe of each resulting fragment (normalized by the fragment length).

Definition 3 (Maximum normalized absolute Writhe) Let $p_{k,l}$ denote the part of a protein from residue k to residue l and let the first residue of the protein be K and the last be N. The *maximum normalized absolute Writhe* of a protein is defined by $Wr_{max} = \max |Wr(p_{i,j})|/(N-K)$, where $K \leq i \leq (N-1)$ and $(i+1) \leq j \leq N$.

We call a fragment of a protein with the maximum normalized absolute Writhe, the maximum normalized absolute Writhe fragment of the protein (see Fig. S2 in SI for examples).

Topological free energy

The topological free energy of a protein was introduced in[47], as a measure of rarity of a (local) protein conformation (characterized by a geometrical/topological parameter) relative to (local) conformations of proteins in the ensemble of folded proteins. As a representative of ensemble of folded proteins, we use a culled dataset of 13,193 proteins in the PDB with less than 60% of homology[79]. By using a sliding window approach of N residues, we scan along each protein and compute its Writhe of N consecutive residues. If there is a gap in these residue numbers, the Writhe is not be computed. Out of proteins in the dataset, proteins of length less than N are excluded. By doing this for the entire set of proteins, we obtain a distribution of Writhe values for fragments of N residues, from which the topological free energy of a protein in the folded ensemble is defined:

Definition 4 (N-topological free energy of a protein in the ensemble of folded proteins) Let d_{Wr} denote the density (i.e. the number of occurrences) of a given Writhe value (window) of a segment of length N in the folded ensemble. Let m denote the (window) value of Writhe of maximum occurrence. To any value p of the Writhe, a normalized quantity is associated and called a Topological Free Energy in Writhe:

$$\Pi(p) = \ln[d_{Wr}(m)/d_{Wr}(p)].$$

The topological free energies obtained from all Writhe values of proteins in the folded ensemble constitute the Topological Free Energy distribution of the folded ensemble. According to this distribution, a topological free energy of a protein, $\tilde{\Pi}$, is assigned based on its writhe value. We say that a conformation has a high topological free energy when it attains a higher topological free energy than the 75th percentile of the distribution (a value outside the 75th percentile). A conformation of a protein with a high topological free energy is called a high topological free energy conformation, and in this case, it is considered to be rare in the folded ensemble.

When $N = 4$ (which is the minimum length at which we can define the Writhe), we call the topological free energy of a conformation of 4 residues, the Local Topological Free Energy (LTE). The culled dataset of 13,193 proteins in the PDB results in a sample of 2,929,754 local conformations of length 4, for which the Writhe values are computed. The LTE distribution of the folded ensemble as a function of Writhe is shown in Supplementary Fig. S3.

An analogue of LTE can be defined for other values of scanning window size. By increasing the window size, we capture more global/topological information of a protein conformation, obtaining a global topological free energy. In this manuscript, we use $N = 31$ and $N = 19$ for a scanning window, as these correspond to the average length of repeats in the MTBR and to the jR2R3 peptide, respectively. For the purpose of this study, we call these the Repeat Topological Free Energy (RTE) and the Peptide Topological Free Energy (PTE), respectively. For the ensemble of folded proteins, we use a culled dataset of 13,082 proteins in the PDB, which is a subset of the culled dataset used for the LTE and results in a sample of 82,309 local conformations of length 31 and in a sample

of 143,842 local conformations of length 19. In order to better capture the geometrical complexity without cancellations that occur in the Writhe calculation, we define the RTE and PTE in ACN. The same Definition 4 for *N*-topological free energy holds except that the Writhe is replaced by the ACN. The RTE and PTE distributions of the folded ensemble as a function of the ACN are shown in Supplementary Fig. S7 and S9, respectively.

Data availability

All data generated or analysed during this study are included in this published article and Supplementary information.

Received: 18 September 2024; Accepted: 5 March 2025
Published online: 25 March 2025

References

1. Hamza, T. H. et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. *PLoS Genet.* **7**, e1002237 (2011).
2. Shi, Y. et al. Structure-based classification of tauopathies. *Nature* **598**, 359–363 (2021).
3. Gibbons, G. S. et al. Detection of Alzheimer Disease (AD)-specific tau pathology in AD and nonad tauopathies by immunohistochemistry with novel conformation-selective tau antibodies. *J. Neuropathol. Exp. Neurol.* **77**, 216–228 (2018).
4. Verelst, J. et al. A novel tau antibody detecting the first amino terminal insert reveals conformational differences among tau isoforms. *Frontiers Mol. Bio.* **7**, 48 (2020).
5. Zeng, Z., Fichou, Y., Vigers, M. & Han, S. Illuminating the structural basis of tau aggregation by intramolecular distance tracking: A perspective on methods. *J. Phys. Chem. B* **126**, 6384–6395 (2022).
6. Mirbaha, H. et al. Inert and seed-competent tau monomers suggest structural origins of aggregation. *Elife* **7**, e36584 (2018).
7. Mandelkow, E. & Mandelkow, E.-M. Microtubules and microtubule-associated proteins. *Curr. Opin. Cell Biol.* **7**, 72–81 (1995).
8. Kadavath, H. et al. Tau stabilizes microtubules by binding at the interface between tubulin heterodimers. *Proc. Natl. Acad. Sci.* **112**, 7501–7506 (2015).
9. Knops, J. et al. Overexpression of tau in a nonneuronal cell induces long cellular processes. *J. Cell Biol.* **114**, 725–733 (1991).
10. Cabrales Fontela, Y. et al. Multivalent cross-linking of actin filaments and microtubules through the microtubule-associated protein tau. *Nat. Commun.* **8**, 1981 (2017).
11. El Mammeri, N., Gampp, O., Duan, P. & Hong, M. Membrane-induced tau amyloid fibrils. *Commun. Biol.* **6**, 467 (2023).
12. Mietelska-Porowska, A., Wasik, U., Goras, M., Filipek, A. & Niewiadomska, G. Tau protein modifications and interactions: Their role in function and dysfunction. *Int. J. Mol. Sci.* **15**, 4671–4713 (2014).
13. Alquezar, C., Arya, S. & Kao, A. W. Tau post-translational modifications: Dynamic transformers of tau function, degradation, and aggregation. *Front. Neurol.* **11**, 595532 (2021).
14. Oakley, S. S. et al. Tau filament self-assembly and structure: Tau as a therapeutic target. *Front. Neurol.* **11**, 590754 (2020).
15. Horie, K., Barthélemy, N. R., Sato, C. & Bateman, R. J. CSF tau microtubule binding region identifies tau tangle and clinical stages of Alzheimer's disease. *Brain* **144**, 515–527 (2021).
16. Fichou, Y. et al. The elusive tau molecular structures: Can we translate the recent breakthroughs into new targets for intervention?. *Acta Neuropathol. Commun.* **7**, 1–17 (2019).
17. Zhang, Y., Wu, K.-M., Yang, L., Dong, Q. & Yu, J.-T. Tauopathies: New perspectives and challenges. *Mol. Neurodegener.* **17**, 28 (2022).
18. Li, L. et al. Disease-associated patterns of acetylation stabilize tau fibril formation. *Structure* **31**, 1025–1037 (2023).
19. Mullapudi, V. et al. Network of hotspot interactions cluster tau amyloid folds. *Nat. Commun.* **14**, 895 (2023).
20. Chen, D. et al. Tau local structure shields an amyloid-forming motif and controls aggregation propensity. *Nat. Commun.* **10**, 2493 (2019).
21. Longhini, A. P. et al. Precision proteoform design for 4R tau isoform selective templated aggregation. *Proc. Natl. Acad. Sci.* **121**, e2320456121 (2024).
22. Gamache, J. E. et al. Developmental pathogenicity of 4-repeat human tau is lost with the P301L mutation in genetically matched tau-transgenic mice. *J. Neurosci.* **40**, 220–236 (2020).
23. Kawasaki, R. & Tate, S.-I. Impact of the hereditary P301L mutation on the correlated conformational dynamics of human tau protein revealed by the paramagnetic relaxation enhancement nmr experiments. *Int. J. Mol. Sci.* **21**, 3920 (2020).
24. Pounot, K. et al. Mutations in tau protein promote aggregation by favoring extended conformations. bioRxiv 2023-05 (2023).
25. Fitzpatrick, A. W. et al. Cryo-em structures of tau filaments from Alzheimer's disease. *Nature* **547**, 185–190 (2017).
26. Shi, Y. et al. Cryo-em structures of tau filaments from Alzheimer's disease with pet ligand apn-1607. *Acta Neuropathol.* **141**, 697–708 (2021).
27. Falcon, B. et al. Novel tau filament fold in chronic traumatic encephalopathy encloses hydrophobic molecules. *Nature* **568**, 420–423 (2019).
28. Falcon, B. et al. Structures of filaments from Pick's disease reveal a novel tau protein fold. *Nature* **561**, 137–140 (2018).
29. Zhang, W. et al. Novel tau filament fold in corticobasal degeneration. *Nature* **580**, 283–287 (2020).
30. Von Bergen, M. et al. Mutations of tau protein in frontotemporal dementia promote aggregation of paired helical filaments by enhancing local β-structure. *J. Biol. Chem.* **276**, 48165–48174 (2001).
31. Lyons, A. J., Gandhi, N. S. & Mancera, R. L. Molecular dynamics simulation of the phosphorylation-induced conformational changes of a tau peptide fragment. *Proteins: Struct. Funct. Bioinform.* **82**, 1907–1923 (2014).
32. Louros, N. et al. Local structural preferences in shaping tau amyloid polymorphism. *Nat. Commun.* **15**, 1028 (2024).
33. Rana, M. M. & Nguyen, D. Geometric graph learning to predict changes in binding free energy and protein thermodynamic stability upon mutation. *J. Phys. Chem. Lett.* **14**, 10870–10879 (2023).
34. Chen, D., Liu, J. & Wei, G.-W. Multiscale topology-enabled structure-to-sequence transformer for protein–ligand interaction predictions. *Nat. Mach. Intell.* **6**, 799–810 (2024).
35. Tang, W. S. et al. A topological data analytic approach for discovering biophysical signatures in protein dynamics. *PLOS Comp. Biol.* **18**, e1010045 (2022).
36. Cang, Z., Mu, L., Opron, K., Xia, K. & Wei, G.-W. A topological analytic approach for protein classification. *Mol. Based Math. Biol.* **3**, 140–162 (2015).
37. Dabrowski-Tumanski, P. et al. Knotprot 2.0: A database of proteins with knots and other entangled structures. *Nucleic Acids Res.* **47**, D367–D375 (2019).
38. Baiesi, M., Orlandini, E., Seno, F. & Trovato, A. Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding. *Sci. Rep.* **9**, 1–12 (2019).
39. Wang, J. & Panagiotou, E. The protein folding rate and the topology and geometry of the native state. *Sci. Rep.* **12**, 6384 (2022).
40. Bale, A., Rambo, R. & Prior, C. The SKMT algorithm: A method for assessing and comparing underlying protein entanglement. *PLOS Comp. Biol.* **19**, e1011248 (2023).

41. Darcy, I. K. & Vazquez, M. Determining the topology of stable protein–DNA complexes. *Biochem. Soc. Trans.* **41**, 601–605 (2013).
42. Orlandini, E. & Micheletti, C. Topological and physical links in soft matter systems. *J. Phys: Cond. Matter* **34**, ac28bf (2022).
43. Beccara, S., Skrbic, T., Covino, R., Micheletti, C. & Faccioli, O. Folding pathways of a knotted protein with a realistic atomistic force field. *PLOS Comp. Biol.* **34**, ac28bf (2013).
44. Sugiyama, M., Kosik, K. S. & Panagiotou, E. Mathematical topology and geometry-based classification of tauopathies. *Sci. Rep.* **14**, 7560 (2024).
45. Panagiotou, E. & Plaxco, K. W. A topological study of protein folding kinetics. *Topol. Biopolym. AMS Contemp. Math. Ser.* **223** (2020).
46. Berman, H. et al. The Protein Data Bank. *Nucleic Acid Res.* **28**, 235–242 (2020).
47. Baldwin, Q. & Panagiotou, E. The local topological free energy of proteins. *J. Theor. Biol.* **529**, 110854 (2021).
48. Baldwin, Q., Sumpter, B. & Panagiotou, E. The local topological free energy of the sars-cov-2 spike protein. *Polymers* **14**, 3014 (2022).
49. Nikhil, R., Panagiotou, E. & Banu, O. The geometry and topology of the native state and conformational dynamics of proteins (in preparation) (2024).
50. Vaquer-Alicea, J., Diamond, M. I. & Joachimiak, L. A. Tau strains shape disease. *Acta Neuropathol.* **142**, 57–71 (2021).
51. DuPai, C. D., Davies, B. W. & Wilke, C. O. A systematic analysis of the beta hairpin motif in the protein data bank. *Protein Sci.* **30**, 613–623 (2021).
52. Khaled, M. et al. A hairpin motif in the amyloid-β peptide is important for formation of disease-related oligomers. *J. Am. Chem. Soc.* **145**, 18340–18354 (2023).
53. Hill, E., Wall, M. J., Moffat, K. G. & Karikari, T. K. Understanding the pathophysiological actions of tau oligomers: A critical review of current electrophysiological approaches. *Front. Mol. Neurosci.* **13**, 155 (2020).
54. Malhis, M. et al. Potent tau aggregation inhibitor d-peptides selected against tau-repeat 2 using mirror image phage display. *ChemBioChem* **22**, 3049–3059 (2021).
55. Seidler, P. et al. Structure-based inhibitors of tau aggregation. *Nat. Chem.* **10**, 170–176 (2018).
56. Danis, C. et al. Inhibition of tau seeding by targeting tau nucleation core within neurons with a single domain antibody fragment. *Mol. Ther.* **30**, 1484–1499 (2022).
57. Ganguly, P. et al. Tau assembly: The dominant role of PHF6 (VQIVYK) in microtubule binding region repeat R3. *J. Phys. Chem. B* **119**, 4582–4593 (2015).
58. Goux, W. J. et al. The formation of straight and twisted filaments from short tau peptides. *J. Biol. Chem.* **279**, 26868–26875 (2004).
59. Fagnen, C., Giovannini, J., Catto, M., Voisin-Chiret, A. S. & Sopkova-de Oliveira Santos, J. On the tracks of the aggregation mechanism of the PHF6 peptide from tau protein: Molecular dynamics, energy, and interaction network investigations. *ACS Chem. Neurosci.* **13**, 2874–2887 (2022).
60. Vigers, M. et al. Tau P301L mutation promotes core 4R tauopathy fibril fold through near-surface water structuring and conformational rearrangement. bioRxiv 2023-11 (2023).
61. Sahayaraj, A. E., Viswanathan, R., Pinhero, F., Abdul Vahid, A. & Vijayan, V. Sequence-dependent conformational properties of PGGG motif in tau repeats: Insights from molecular dynamics simulations of narrow Pick filament. *ACS Chem. Neurosci.* **14**, 136–147 (2022).
62. Penner, R. C. Backbone free energy estimator applied to viral glycoproteins. *J. Comput. Biol.* **27**, 1495–1508 (2020).
63. Finkelstein, A. V., Badretdinov, A. Y. & Gutin, A. M. Why do protein architectures have Bolzmann-like statistics? *Proteins: Struct. Funct. Genet.* **23**, 142–150 (1995).
64. Hebert, L. E., Weuve, J., Scherr, P. A. & Evans, D. A. Alzheimer disease in the united states (2010–2050) estimated using the 2010 census. *Neurology* **80**, 1778–1783 (2013).
65. Togasaki, D. M. & Tanner, C. M. Epidemiologic aspects. *Adv. Neurol.* **82**, 53–59 (2000).
66. Ferrer, I., Santpere, G. & van Leeuwen, F. W. Argyrophilic grain disease. *Brain* **131**, 1416–1432 (2008).
67. Rizzi, L., Rosset, I. & Roriz-Cruz, M. Global epidemiology of dementia: Alzheimer's and vascular types. *Biomed. Res. Int.* **2014**, 908915 (2014).
68. Hartnell, I. J. et al. Glial reactivity and T cell infiltration in frontotemporal lobar degeneration with tau pathology. *Brain* **147**, 590–606 (2024).
69. Buciuc, M. et al. The many faces of globular glial tauopathy: A clinical and imaging study. *Eur. J. Neurol.* **30**, 321–333 (2023).
70. Logroscino, G. et al. Incidence of syndromes associated with frontotemporal lobar degeneration in 9 European countries. *JAMA Neurol.* **80**, 279–286 (2023).
71. Swallow, D. M., Zheng, C. S. & Counsell, C. E. Systematic review of prevalence studies of progressive supranuclear palsy and corticobasal syndrome. *Movem. Disorders Clin. Pract.* **9**, 604–613 (2022).
72. Shoeibi, A., Olfati, N. & Litvan, I. Frontrunner in translation: Progressive supranuclear palsy. *Front. Neurol.* **10**, 493674 (2019).
73. Rodriguez, R. D. & Grinberg, L. T. Argyrophilic grain disease: An underestimated tauopathy. *Dementia Neuropsychologia* **9**, 2–8 (2015).
74. El Mammeri, N., Duan, P., Dregni, A. J. & Hong, M. Amyloid fibril structures of tau: Conformational plasticity of the second microtubule-binding repeat. *Sci. Adv.* **9**, eadh4731 (2023).
75. Dunbar, J. et al. Sabdab: the structural antibody database. *Nucleic Acids Res.* **42**, D1140–D1146 (2014).
76. Longhini, A. P. et al. A small tau fragment specifically templates four repeat tau aggregates through multiple generations. biorxiv 1 (2023).
77. Gauss, K. F. Gauss 1833 werke. *Königlichen Gesellschaft der Wissenschaften zu Göttingen* (1877).
78. Banchoff, T. Self-linking numbers of space polygons. *Indiana Univ. Math. J.* **25**, 1171–1188 (1976).
79. Wang, G. & Dunbrack, R. L. Jr. PISCES: A protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).

Acknowledgements

Masumi Sugiyama (MS) and Eleni Panagiotou (EP) were supported by the National Science Foundation, NSF CAREER 2047587 and the National Institutes of Health R01GM152735. MS acknowledges support from the Research Institute at the University of Tennessee at Chattanooga. We acknowledge support from the Arizona Alzheimer's Consortium (Pilot Grant) and the Arizona State University Women in Philanthropy.

Author contributions

E.P. (Eleni Panagiotou) and K.S. (Kenneth Kosik) conceived of the work; M.S. (Masumi Sugiyama) acquired the data; M.S. and E.P. analyzed the data and created new software used in the work; M.S., K.S. and E.P. interpreted the results and participated in writing the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-93304-x.

Correspondence and requests for materials should be addressed to E.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.