Non-standard proteins in the lens of AlphaFold 3 - a case study of amyloids

Alicja W. Wojciechowska[1,2], Jakub W. Wojciechowski[2], Gert Vriend[3], Malgorzata Kotulska[1]✉
✉ malgorzata.kotulska@pwr.edu.pl
1 Wroclaw University of Science and Technology, Faculty of Fundamental Problems of Technology, Department of Biomedical Engineering, Wroclaw, Poland
2 Sano Centre for Computational Medicine, Krakow, Poland
3 Baco Institute of Protein Science, Mindoro, Philippines

CONTENT

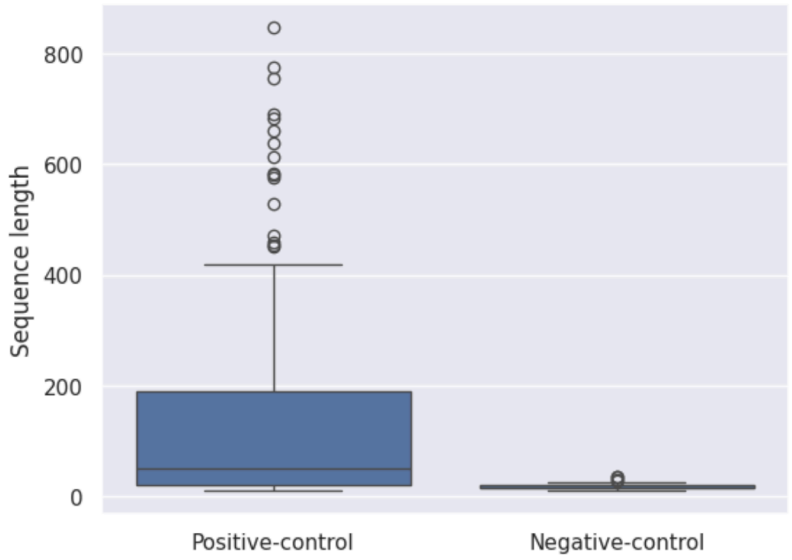


Figure S1. Distributions of sequence length for *Positive-* and *Negative*-control.

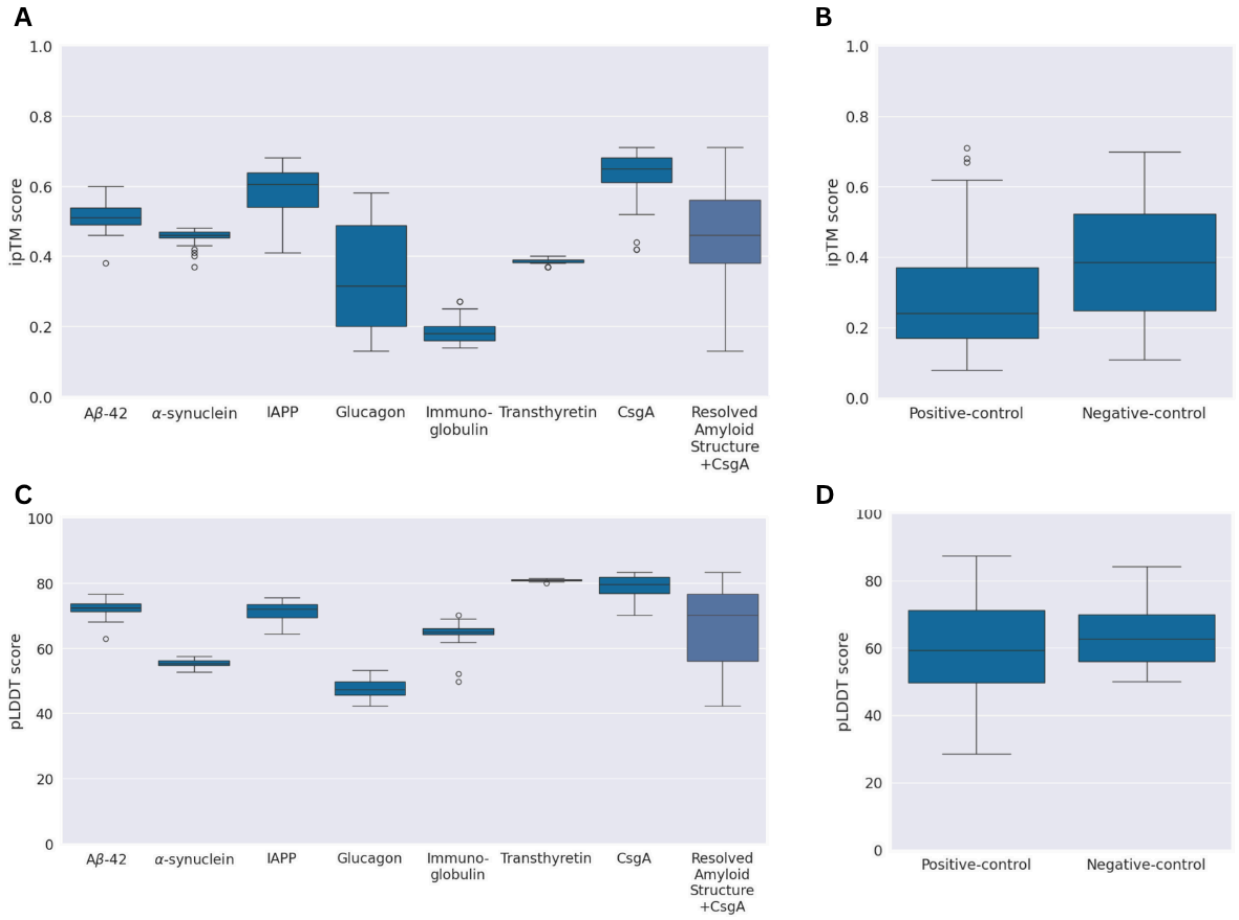


Figure S2. Distributions of AF3 quality metrics for all datasets. A. ipTM score distribution for *ResolvedAmyloidStructure* and CsgA. B. ipTM score distribution for *Positive-* and *Negative-control*. C. pLDDT score distribution for *ResolvedAmyloidStructure* and CsgA. D. pLDDT score distribution for *Positive-* and *Negative-control*.

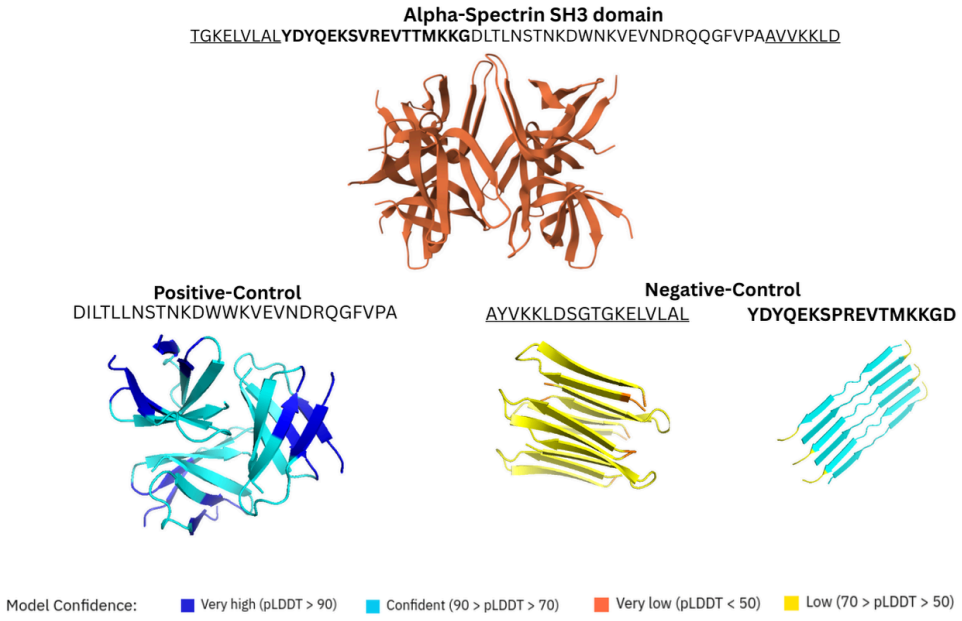


Figure S3. A case study of Alpha-Spectrin SH3 domain deposited in AmyLoad. The entire structure of the domain is not predicted as an amyloid. The aggregation-prone region (present in *Positive-control* dataset) is not predicted as an amyloid. The non-aggregating peptides of this domain (present in *Negative-control* dataset) are both predicted as amyloid structures.

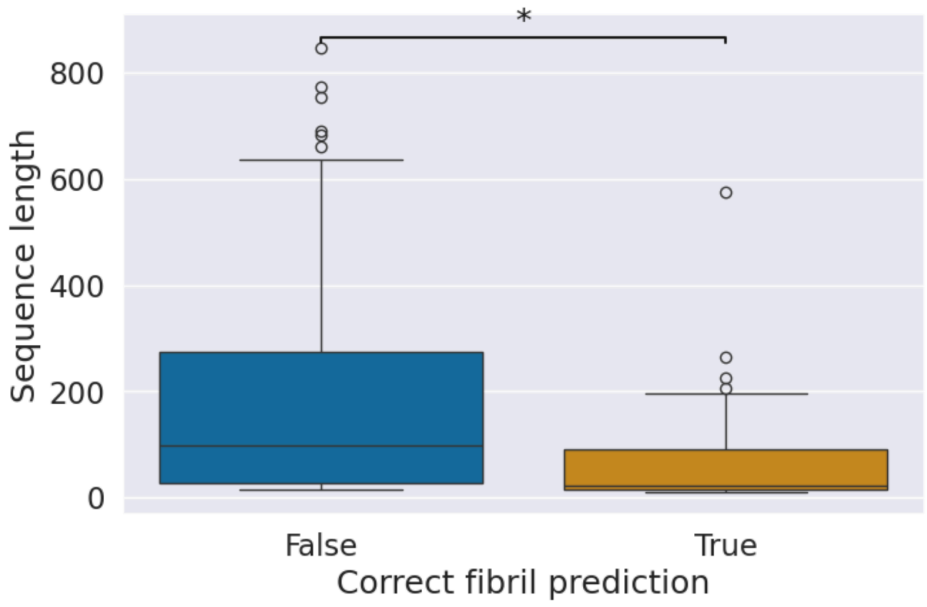


Figure S4. Distribution of the sequence length for *correct* and *wrong* AF3 predictions for the *Positive-control*.

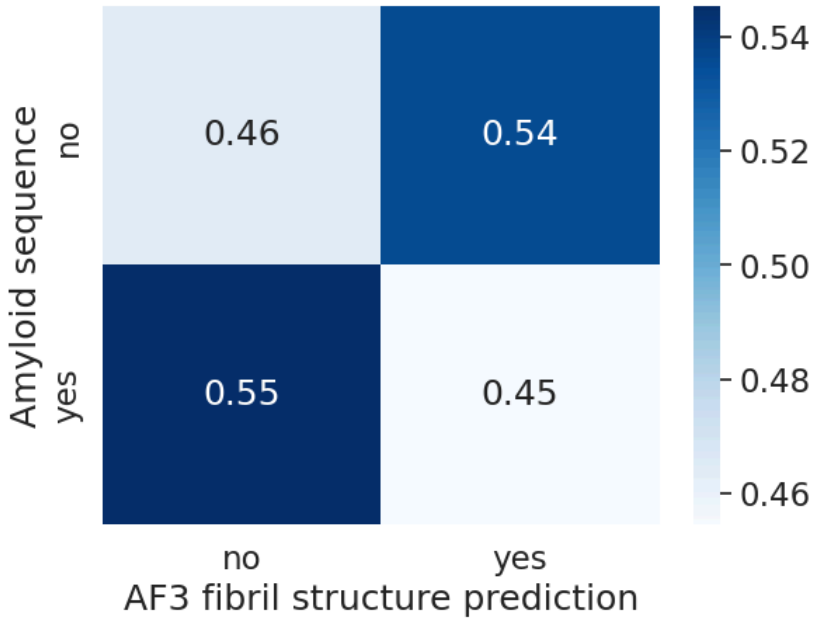


Figure S5. Confusion matrix for the predictions of the *Positive-control* and *Negative-control* datasets. To create this confusion matrix, we used only sequences of lengths 36 or shorter in the *Positive-control*. All sequences in the *Negative-control* are also of length 36 or shorter.

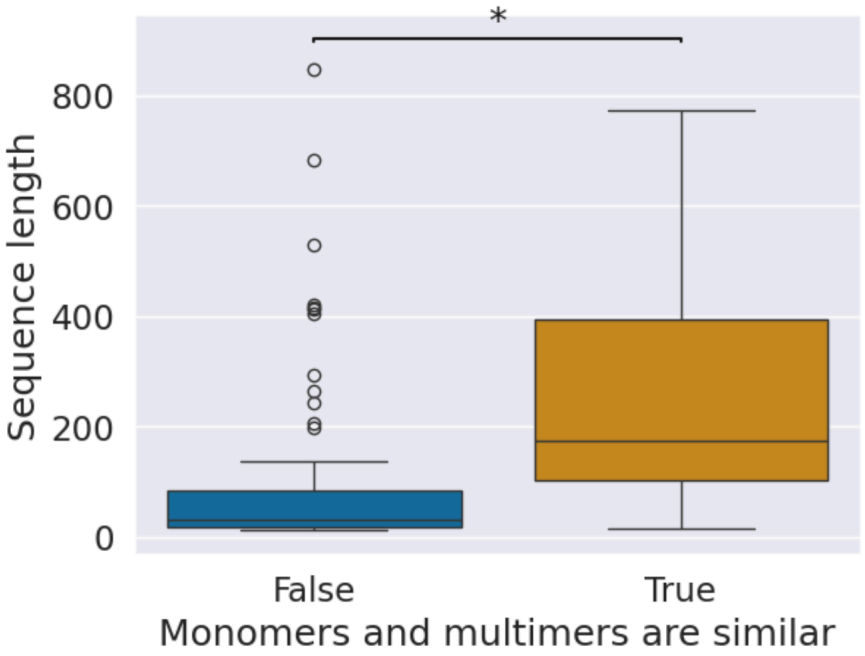


Figure S6. Sequence length distribution and similarity between monomeric and multimeric models for the *Positive-control* dataset. The difference between the left and right distributions was statistically significant; Mann-Whitney p-value=3e-10. When models for sequences with a length above 36 amino acids are only considered, the similarity between multimeric and monomeric models implies an increase in the pLDDT score by 25 points (statistically significant, Mann-Whitney p-value=2e-12).

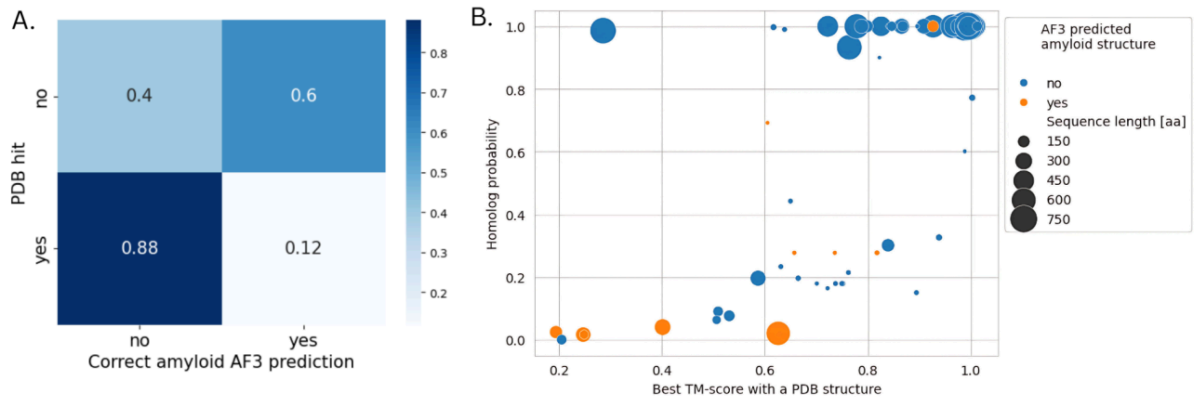


Figure S7. Comparison of AF3 models for amyloids to PDB structures. Results of the search of AF3 models in the PDB for amyloid proteins (*Positive-control*). **A.** Confusion matrix for amyloid proteins from the *Positive-control* dataset concerning AF3 predictions and similarity of the models to PDB structures. **B.** Foldseek homolog probability as a function of the TM-score between the AF3 model and the most similar PDB file. Higher homolog probability was associated with higher pLDDT (mean pLDDT with a homolog probability above 0.95 was 67, with no homolog 57, p-value=1e-4).

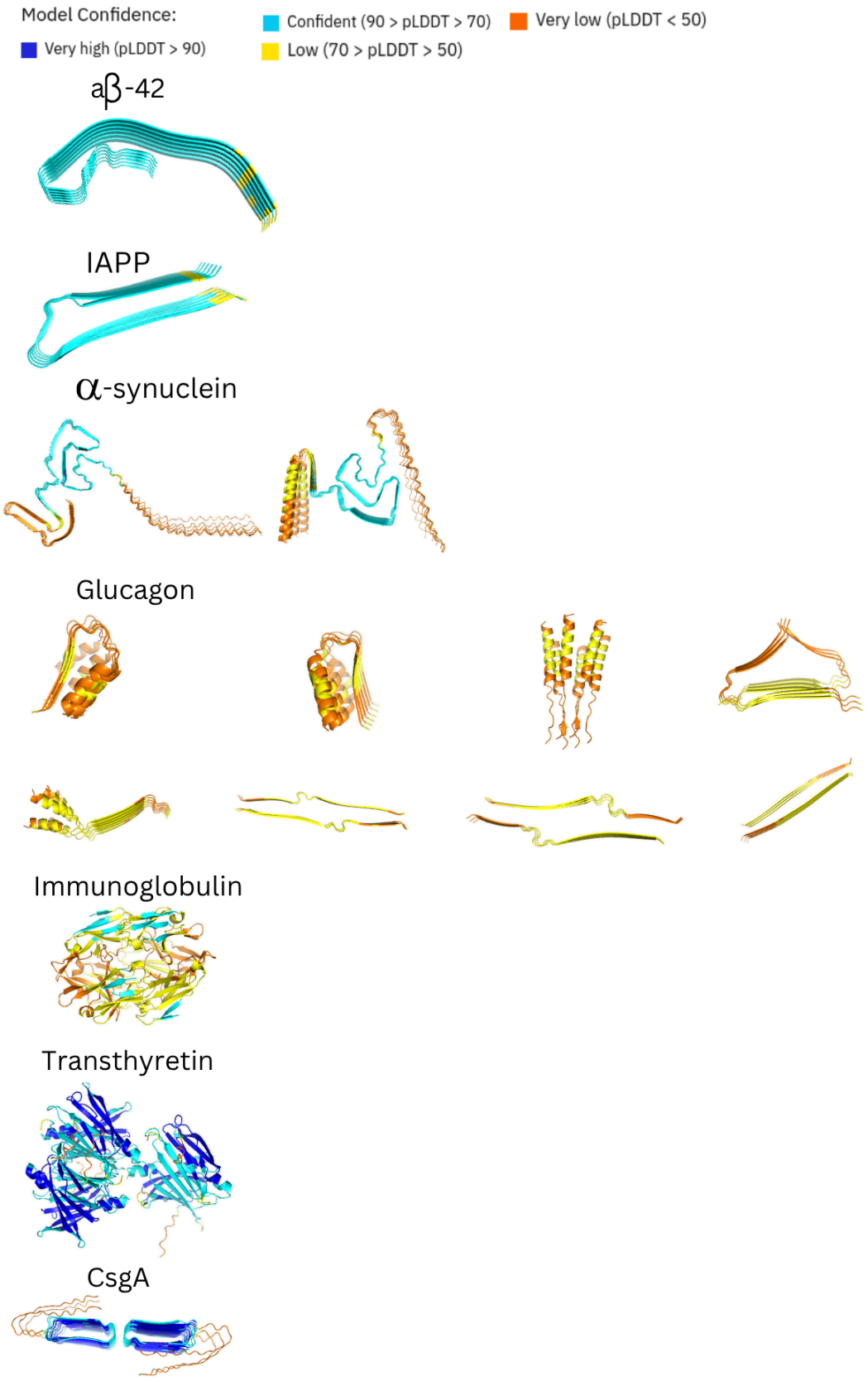


Figure S8. Cluster representatives for *ResolvedAmyloidStructure* and CsgA proteins. Structures are coloured according to the pLDDT score per residue.

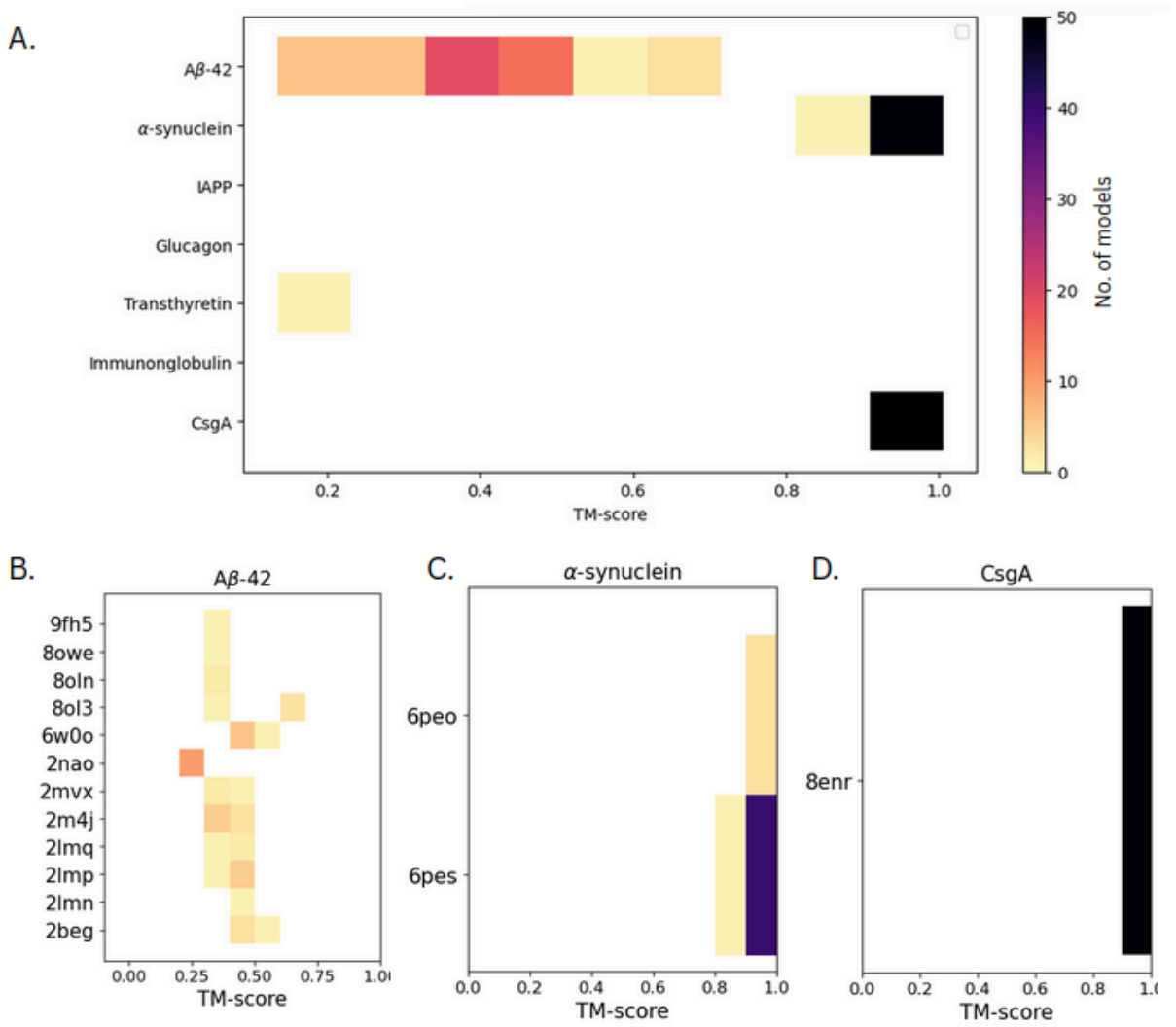


Figure S9. Best Foldseek hits for each AF3 model for *ResolvedAmyloidStructure* dataset and CsgA in the STAMP database of amyloid structures. The colors represent the number of models that had the best match to a PDB identifier. Foldseek TM-scores for one chain from the n=6 model and the hit are provided; single-chain. **A.** TM-scores of all hits **B.** Similarity between 50 Aβ-42 AF3 models and amyloid structures of this protein in STAMP **C.** Similarity between 50 α-synuclein AF3 models and amyloid structures of this protein in STAMP. **D.** Similarity between 50 CsgA AF3 models and amyloid structures of this protein in STAMP (extended by 8enr and 8enq PDB files).

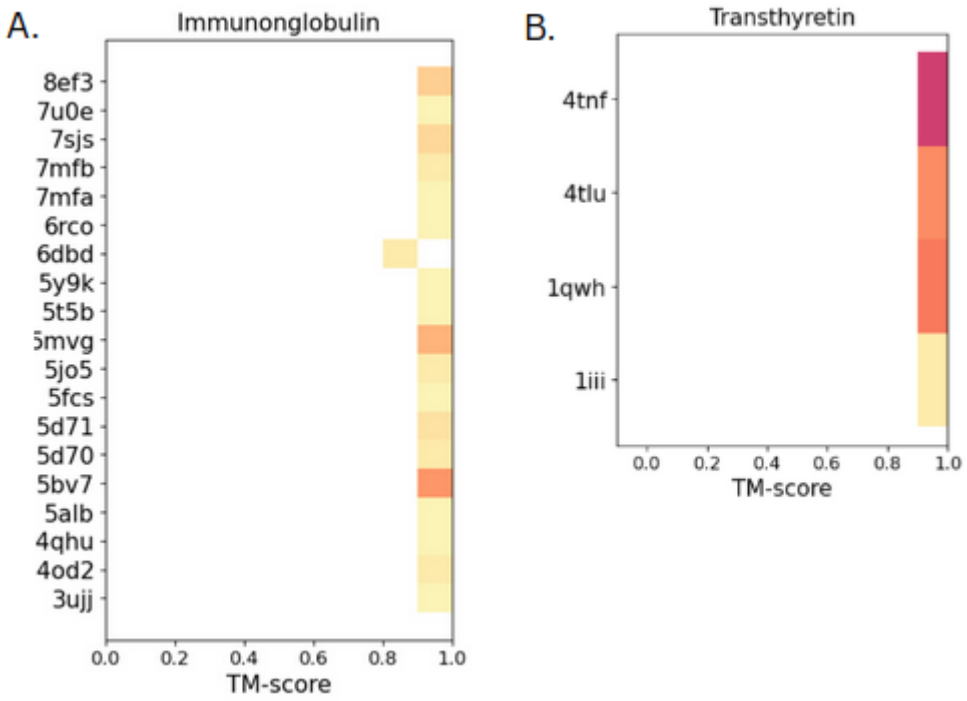


Figure S10. A list of the best Foldseek hits for AF3 models in the PDB for A. immunoglobulin and B. transthyretin. All hits are globular structures. Foldseek TM-scores for one chain from the n=6 model and the hit are provided; single-chain.

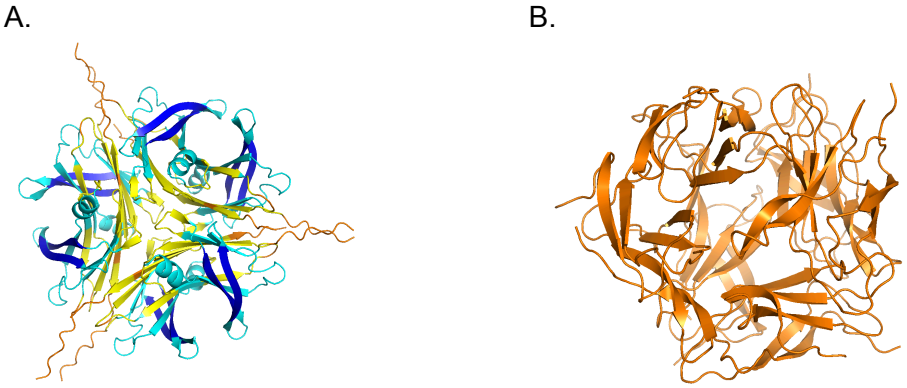


Figure S11. The cluster representative of all AF3 models for A. transthyretin and B. immunoglobulin.

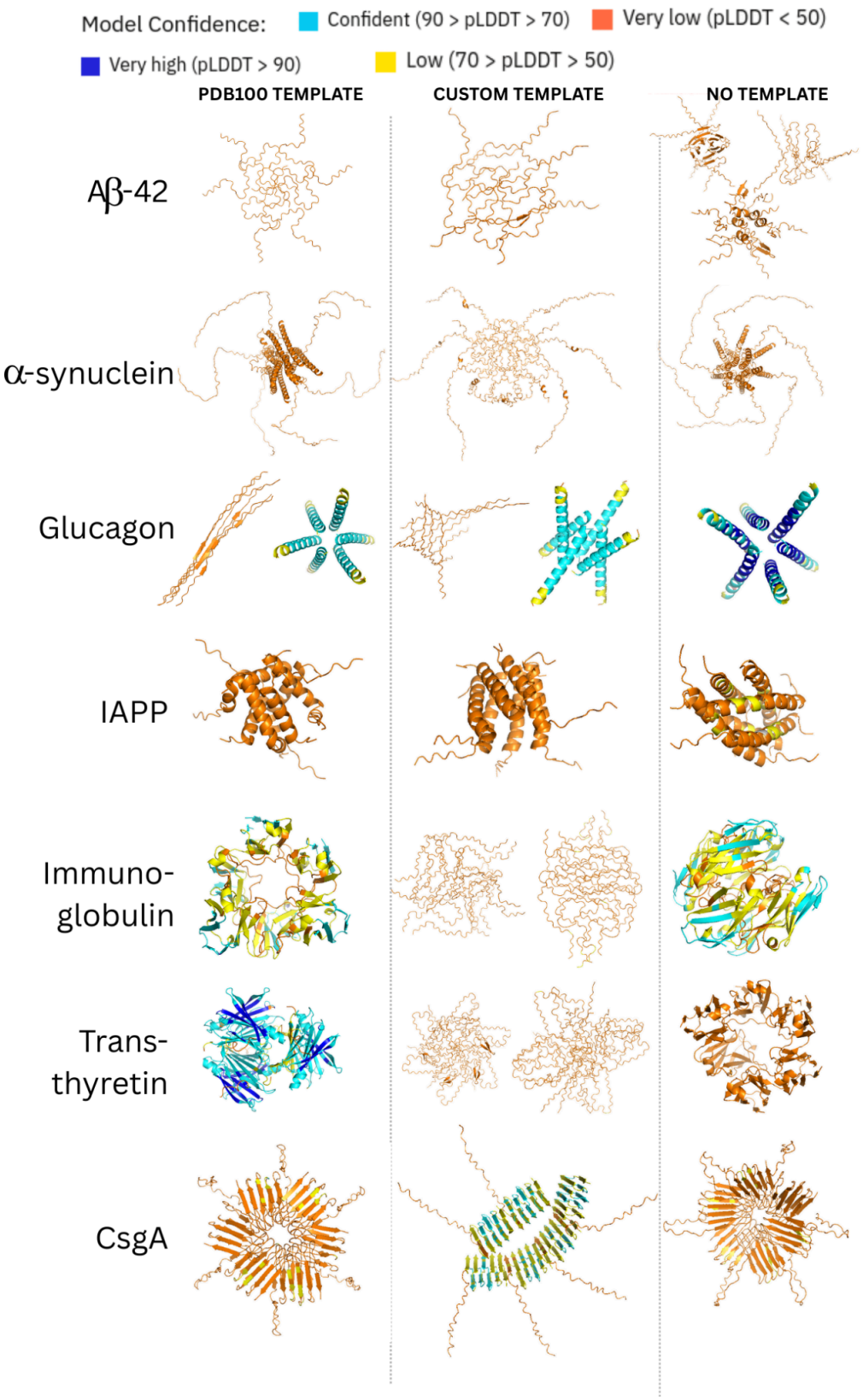


Figure S12. AF-Multimer models for the *ResolvedAmyloidStructure* dataset (and CsgA): using a PDB100 template, a custom template from Table S2 and without any template**.** Five models generated with AF-Multimer in each case were clustered with Foldseek and their cluster representatives are shown.

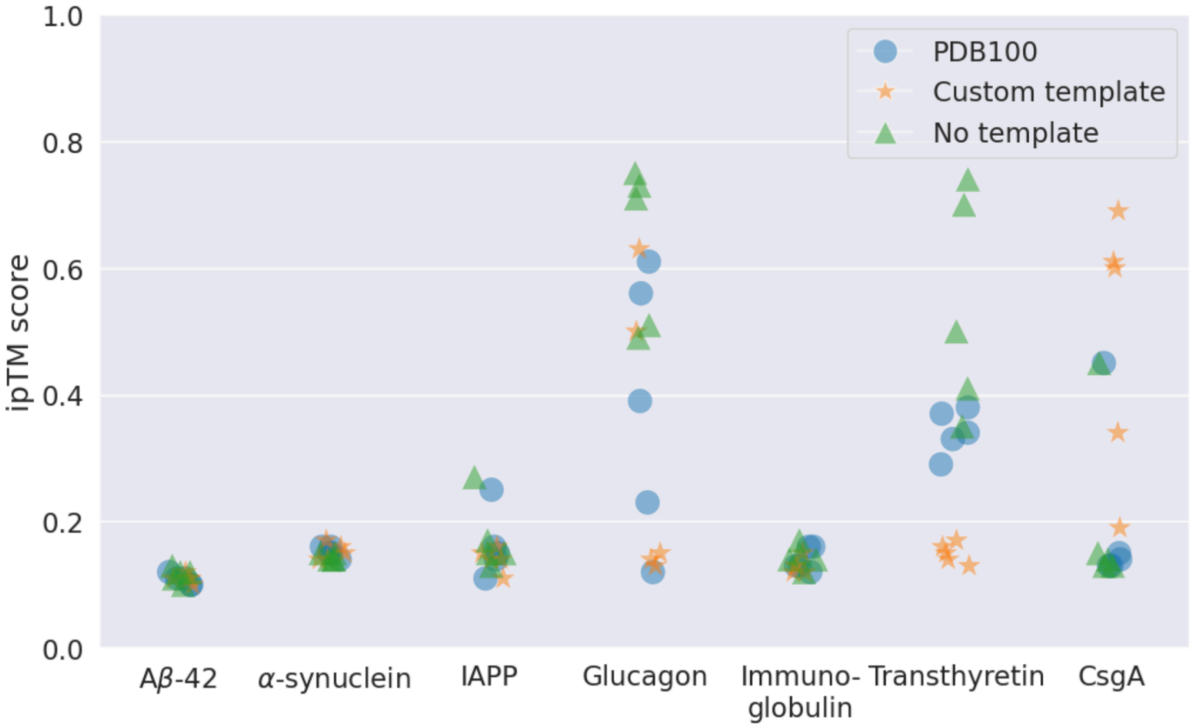


Figure S13. ipTM score distribution for *ResolvedAmyloidStructure* and CsgA models produced with AF-Multimer.

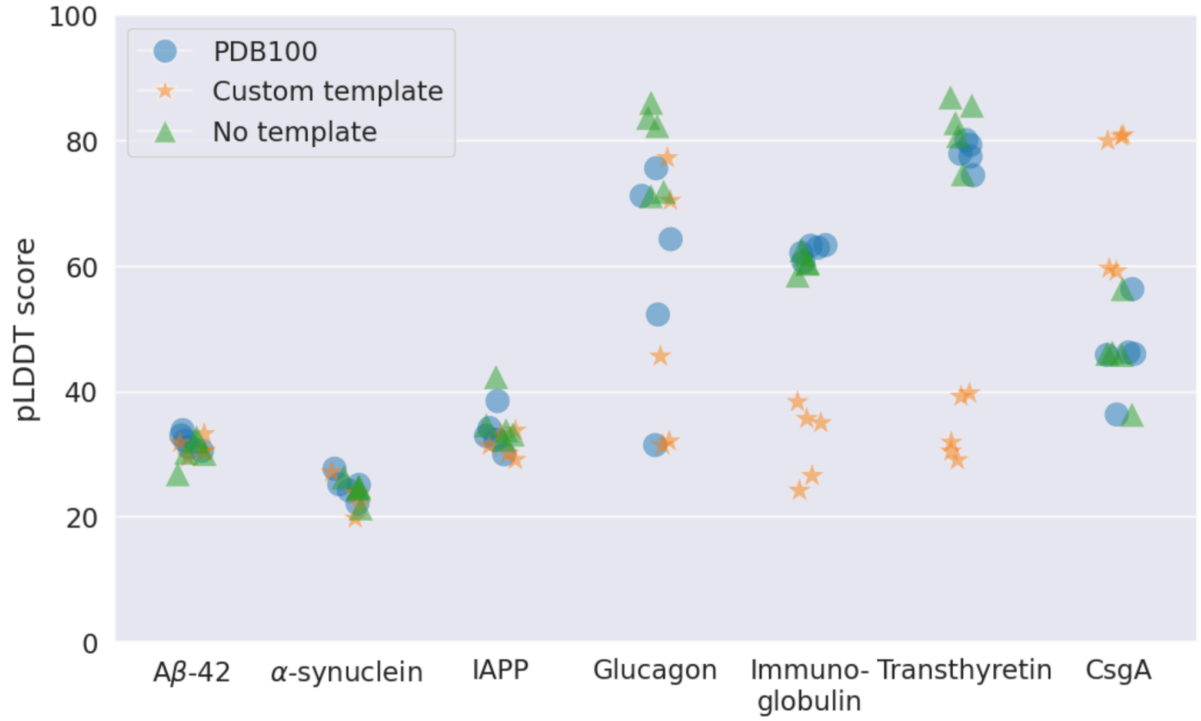


Figure S14. pLDDT score distribution for *ResolvedAmyloidStructure* and CsgA models produced with AF-Multimer.

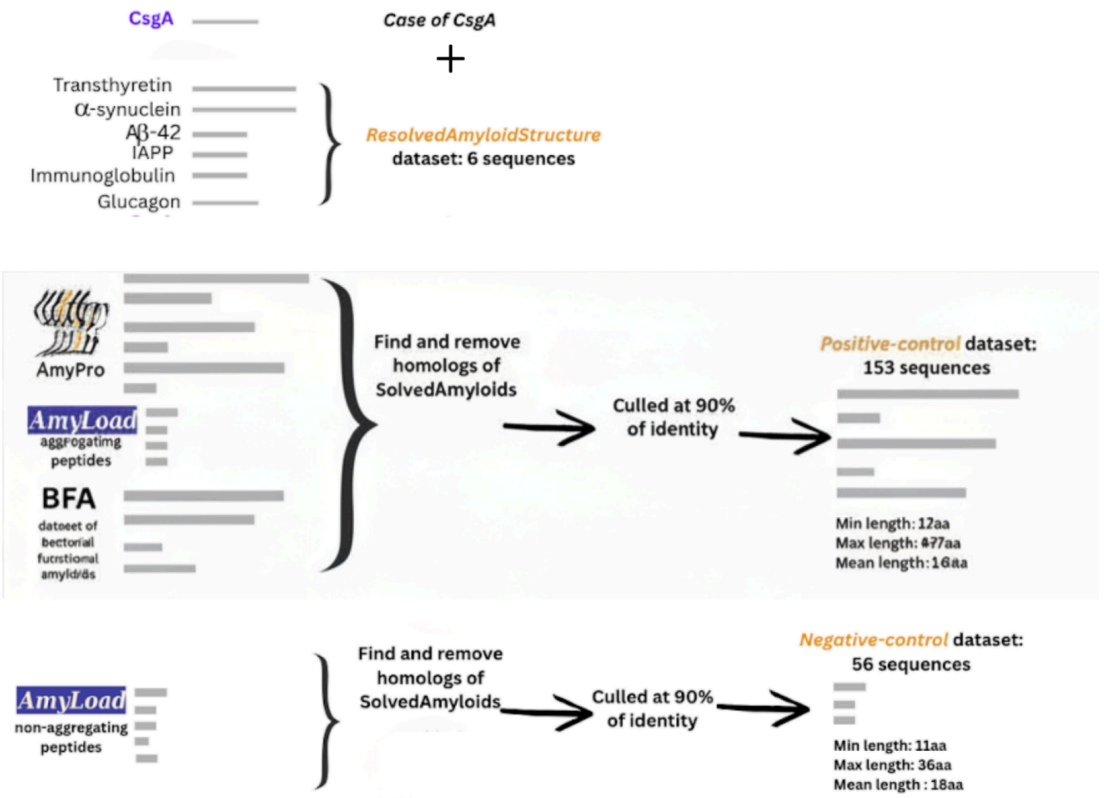
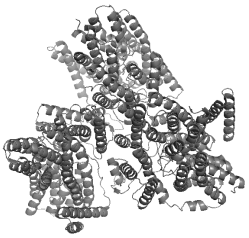
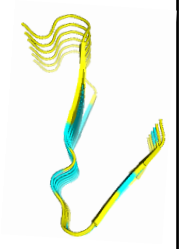
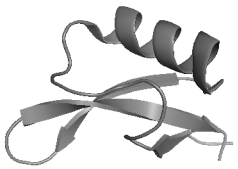
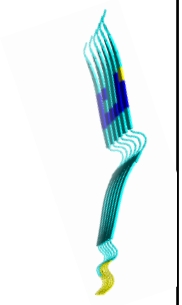
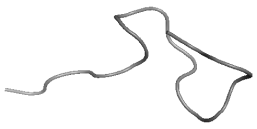
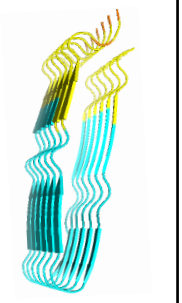


Figure S16. Summary of the dataset preparation process.

Protein	AF3 multimer model (models colored by pLDDT)	mmseqs hit in the PDB	mmseqs alignment E-value, (identity)	Comment about the hit PDB file
Microcin E492		7dyr 	1.7e-42 (1.0)	Protein complex with globular microcin

PB1-F2, C-terminal domain		2hn8 	4e-19 (0.95)	Globular monomer of the protein with the C-terminal fragment
Defensin-like protein 1, C-terminal peptide		2n2r 	2.3e-17 (0.97)	Globular monomer of a homolog
Obestatin		2jsj 	6.8e-8 (0.87)	Globular monomer of a homolog
Merozoite surface protein 2, MSP2		2mu8 	6.3e-4 (1.0)	Globular monomer of a small fragment of this protein
Calcitonin			4.9e-14 (0.85)	Non-amyloidogenic analogue of the protein

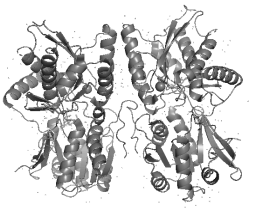
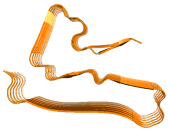
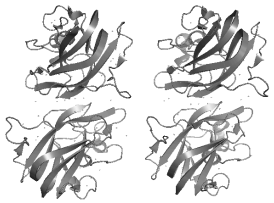
Alpha-S2-casein			5.5e-15 (1.0)	Distant homologous globular fragment of a protein
Natriuretic peptides B, ProBNP			1.4e-6 (1.0)	Homologous globular fragment of a protein
Medin			2.4e-17 (0.66)	Globular monomer of a homolog

Table S1. Mmseqs hits found for the sequences of correct *Positive-control* models in the PDB.

Protein	**Sequence**	**Example of a PDB identifier of an amyloid structure that was part of the AF3 training set**
Aβ-42	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA	2mxu
α-synuclein	MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA	2n0a
IAPP	KCNTATCATQRLANFLVHSSNNFGAILSSTNVGSNTY	6vw2

Glucagon	HSQGTFTSDYSKYLDSRRAQDFVQW LMNT	6nzn
Immunoglobulin lambda variable 3-19 light chain	AVSVALGQTVRITCQGDSLRSYSASW YQQKPGQAPVLVIFRRFSGSSSGNTA SLTITGAQAEDEADYYCNSRDSSANH QVFGGGTKLTV	6z1i
Transthyretin	GPTGTGESKCPLMVKVLDAVRGSPAI NVAMHVFRKAADDTWEPFASGKTSES GELHGLTTEEEFVEGIYKVEIDTKSYW KALGISPFHEHAEVVFTANDSGPRRYT IAALLSPYSYSTTAVVTNPKE	6sdz
CsgA	MGVVPQYGGGGNHGGGGNNSGPNS ELNIYQYGGGNSALALQTDCRNSDLTI TQHGGGNGADVGQGSDDSSIDLTQR GFGNSATLDQWNGKNSEMTVKQFGG GNGAAVDQTASNSSVNVTQCGFGNN ATAHQY	

Table S2. Sequences of *ResolvedAmyloidStructure* datasets and examples of PDB identifiers of amyloid structures that were part of the AF3 training set. In the grey colour, the used sequence of CsgA is given.

Protein	**Depth of the full Colab MSA**	**Neff: Colab MSA full**	**Neff: MSA with first 1000 sequences**	**Neff: MSA with first 500 sequences**	**Neff: MSA with first 100 sequences**
Immunoglobulin lambda variable 3-19 light chain	13297	885	27	6	2
Transthyretin	6176	385	47	13	2

Table S3. Number of effective sequences (Neff) in the Colab-generated MSA and in the cases when only the first 1000, 500 and 100 sequences of it were extracted and applied in AF3 modelling.

Supplementary Note S1. On the automatic classification of amyloid versus non-amyloid structures.

Problem statement

Currently there is no software that can determine whether a structure is amyloid or not, which leads to a strong reliance on manual curation in the field. In our work, we supplement the manual curation with an automatic classifier. The automatic classifier takes a structure in PDB format as input, and as output it provides the classification (0: it is not an amyloid structure, 1: it is an amyloid structure). Below we provide information on the classifier construction and performance, and a comparison with manual curation.

Dataset

The dataset consists of two parts:

1. Structures considered to be amyloid were extracted from the STAMP database that has 823 entries. NMR structures were discarded for simplicity, giving the final set of 491 amyloid structures. These structures have label 1. **These are amyloid structures.**

2. All homo-oligomeric structures were extracted from the PDB. These were culled at 30% sequence identity (by the PDB). Again, only structures solved using X-Ray diffraction or electron microscopy were taken. From the remaining 14845 structures only those in which all chains were of identical length were considered (some structures in the PDB have missing residues). Structures present in the STAMP dataset were excluded from the PDB dataset. This resulted in a final set of non-amyloid structures with 1884 entities. These structures have label 0. **These are non-amyloid homo-oligomeric structures.**

   Details on the PDB query used on the PDB website https://www.rcsb.org/ (access date: 12.06.2026):

Entry Polymer Types	is Protein (only)
Total Number of Polymer Instances (Chains)	> 2
Oligomeric state	is not empty
Refinement Resolution	< 3
Experimental Method	is X-Ray Diffraction OR is Electron Microscopy
[Tab -> Macromolecules] Group results by	Sequence Identity < 30%

[Tab -> Macromolecules] Represented by	Resolution: Best
Tab -> Representatives	

Summary of the used data:

	Number of entities
Amyloid structures from STAMP (label=1)	491
Non-amyloid homooligomeric structures (label=0)	1884

Features

A DSSP vector was calculated for each structure in the dataset using the DSSP programme. This gave:

- Eight continuous columns representing the content of each of the possible DSSP categories (no category '-', B, E, G, H, I, S, T) normalized by the length of the structure.

Symmetry numbers were calculated for each structure in the dataset with AnAnaS. This gave:

- Twenty binary columns representing the presence or absence of each of the twenty possible symmetry numbers (c11, c12, c13, c15, c19, c2, c3, c4, c5, c6, c7, c8, c9, d2, d3, d4, d5, d6, d7, d8).

The vector of RMSD distances between consecutive chains in the structure was calculated with the biotite package. This gave:

- Two continuous columns. The first represents the mean of RMSD vectors and the second its standard deviation.

The vector of rotational angles between consecutive chains in the structure was calculated with the biotite package [1]. This gave:

- Two continuous columns. The first represents the mean of the angle vector and the second its standard deviation.

This gives the feature vector of length 32 for each of the structures in the dataset.

Model training

The dataset was divided into training and testing sets [80-20%]. The partitioning utilized a stratification strategy, so that both the training and testing subsets maintain identical class proportions of amyloid and non-amyloid structures.

The Lasso regression model was chosen as it reduces the multicollinearity problem. It was evaluated using 5-fold cross-validation on the training set to ensure robustness to the dataset choice. The cross-validation gave stable and satisfactory results (see Table below).

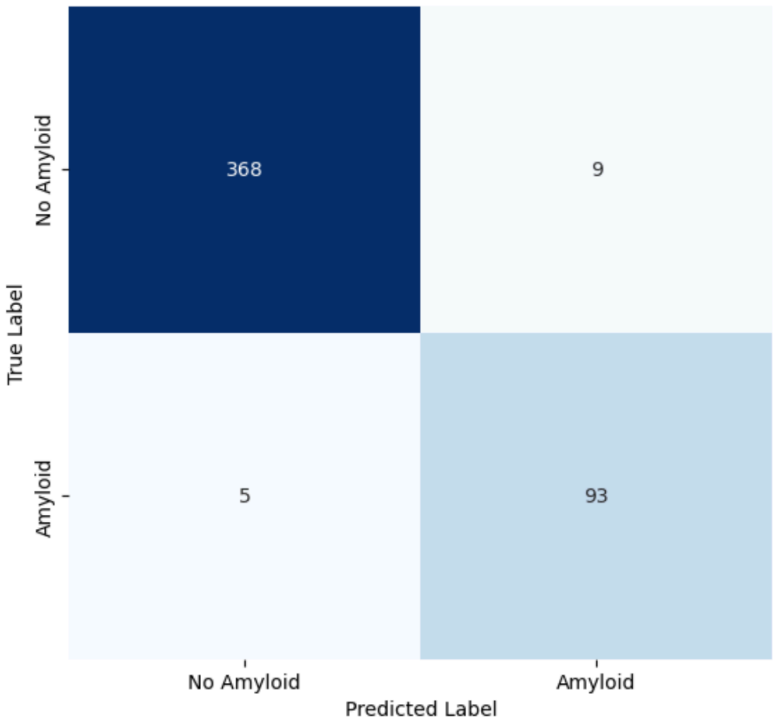
Metric	Mean	Std
F1-score	0.98	0.008
Accuracy	0.98	0.005
Precision	0.95	0.02
Recall	0.98	0.01

Lasso regression was then trained on the entire training set. Coefficient values for non-zeroed columns were the following:

Ananas	presence of c2 symmetry	0.36
	presence of c3 symmetry	-1.91
	presence of c5 symmetry	-0.55
Rotation and translation features	Mean rotation angle between consecutive chains	-0.04
	Standard deviation of rotation angle between consecutive chains	-1.19
	Mean RMSD between consecutive chains	-0.11
DSPP content	Fraction of '-'	6.07
	Fraction of 'E'	1.9
	Fraction of 'H'	-8.5
	Fraction of 'S'	0.41
	Fraction of 'T'	-5.14

Model testing

The model was tested on a testing set giving Accuracy=0.97, F1-score=0.96, Precision=0.94, Recall=0.94. The confusion matrix of the model is the following:

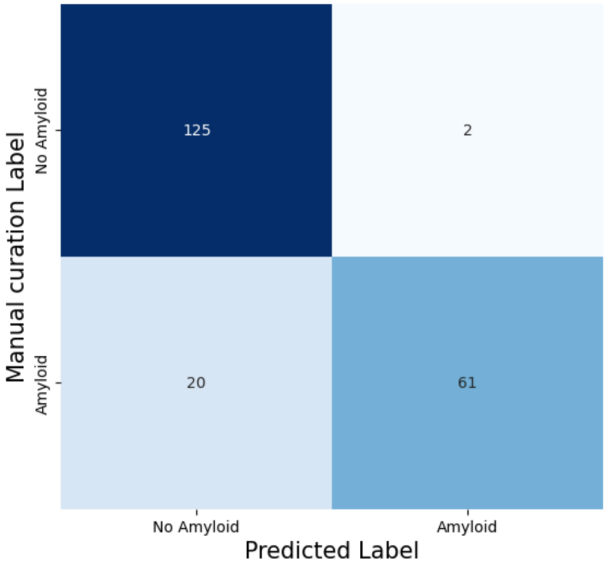


Implementation

The Lasso regression model was implemented in Python 3.15 using the scikit-learn package.

Comparison with manual curation

We applied the model to our dataset of predicted structure from AF3 for Positive-control and Negative-control datasets. The model was mostly consistent with our manual annotation. The confusion matrix is presented below:



Conclusion

We managed to construct an automatic classification method that differentiates between amyloid and non-amyloid structures based on simple structural features and Lasso logistic regression. In 90% of the cases the automatic classification agrees with the manual curation, showing the robustness of our approach in classifying amyloid versus non-amyloid structures.

References:

[1] Kunzmann, Patrick, and Kay Hamacher. "Biotite: a unifying open source computational biology framework in Python." *BMC bioinformatics* 19.1 (2018): 346.