bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Non-standard proteins in the lenses of AlphaFold3 - case study of amyloids

Alicja W. Wojciechowska[1*], Jakub W. Wojciechowski[2], Malgorzata Kotulska[1*]

[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, Cracow, Poland

*Corresponding authors: alicja.wojciechowska@pwr.edu.pl, malgorzata.kotuslka@pwr.edu.pl

Abstract

Motivation

The recent release of *AlphaFold3* raises a question about its powers and limitations. Here, we analyze the potential of *AlphaFold3* in correct reproduction of amyloid structures, which are an example of multimeric proteins characterized by polymorphism and low representation in protein structure databases.

Results

We show that *AlphaFold3* is capable of producing amyloid-like assemblies that have a high similarity to experimental structures, although its results are impacted by the number of monomers in the predicted fibril. It produces structurally diverse models of some amyloid proteins, which could reflect their polymorphism observed in nature. We hypothesize that the lower emphasis on multiple sequence analysis (MSA) in *AlphaFold3* improves the results quality, since for this class of proteins sequence homology is not necessary for their structural similarity. Notably, the structural landscape obtained from the modeling does not reflect the real one governed by thermodynamics, which does not hamper modeling amyloid proteins. Finally, *AlphaFold3* opens the door to fast structural modeling of fibril-like structures, including their polymorphic nature.

Introduction

AlphaFold2, the DeepMind algorithm for protein structure prediction Jumper et al. (2021), boosted the field of Molecular Biology resulting in more than 20,000 citations in less than 3 years. Its modification, *AlphaFold2 - Multimer* Evans et al. (2021), enabled not only the prediction of an individual protein structure but also the prognosis of contacts in multimeric complexes. Their younger brother, *AlphaFold3*, released this year, promises to surpass the master Abramson et al. (2024). Its ability to predict interactions between multiple molecular entities potentially breaks through the glass ceiling of faster and more accurate drug design.

Both models heavily depend on the complex neural network architectures; however, a couple of differences stand out. In the first place, *AlphaFold2*, in contrast to *AlphaFold3*, heavily depends on the evolutionary information encoded in Multiple Sequence Alignment (MSA). *AlphaFold3* also relies on this information, though in a lesser way. The *AlphaFold2* architecture contains an *evoformer* module for MSA processing in its core. The *AlphaFold3* is based on the *pairformer*, which works on the pair representations which embrace MSA

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

information. *AlphaFold3* is a generative model with a stochastic element, which does not require additional structure optimization, in contrast to *AlphaFold2.*

The launch of *AlphaFold2* posed the fundamental question if a neural network can learn thermodynamics laws. It has been shown by Brown et al. (2023) that *AlphaFold2* reproduces the Boltzmann distribution of contact distances in the protein. The pLDDT (Local Distance Difference Test) measure has been shown to correlate with protein flexibility Saldaño et al. (2022) and disordered regions Wilsone et al. (2022). On the other hand, *AlphaFold2* struggled with polymorphic structures Saldaño et al. (2022), Chakravarty and Porter (2022). The conceptual tricks of modifying the input MSA to its reduced versions have been shown to be a partial cure to this problem Monteiro da Silva (2024) Wayment-Steele et al. (2024). *AlphaFold2* also provided unreliable results for antibody-antigen interactions Yin et al. (2022) and hallucinated about knots' appearance in the structure Dabrowski-Tumanski and Stasiak (2023). The improvement of *AlphaFold3* in these tasks is still to be evaluated.

Amyloid proteins are a significant challenge in proteomics. Frequently, on a daily basis, they execute their cellular function correctly in a globular state. However, slight sequence modifications or certain environmental conditions can trigger their aggregation, which leads to formation of fibrils with beta-cross architecture Lutter et al. (2021). Amyloids are mostly associated with neurodegenerative conditions, such as Alzheimer's or Parkinson's diseases. What is even more confusing is that certain proteins are evolutionarily designed to become amyloids. For example, the high stability of amyloid fibrils and their low solubility can make them an ideal material for the construction and scaffolding of biofilm Otzen and Riek (2019). Furthermore, not all amyloid fibrils are the same. The amyloid polymorphism is widespread and a variability of the amyloid structures was observed Lutter et al. (2021).

Prediction of protein aggregation is a difficult task and most algorithms are based on sequence features Kotulska & Wojciechowski (2022). The prediction of the amyloid fibril structure, or even more precisely its polymorphism, is still a holy grail and no dedicated software has been developed. The availability of amyloid structures in PDB is low, because of experimental difficulties. Finally, *AlphaFold2* had poor performance for amyloids, frequently producing alpha helices instead of beta sheets Ragonis-Bachar et al. (2023). Amyloids are a significant challenge for machine learning approaches due to their unusual multi-funneled energy landscape Strodel (2021).

In light of these facts, amyloids seem to be a great case for evaluating the model performance, also due to their clinical significance. Therefore, in this work we focus on the prediction of the amyloid fibril structures using *AlphaFold3*. We observed that *AlphaFold3* reproduces amyloid fibril structures, accounting for their structural polymorphism, although its results depend on the modelled fibril length (number of units assumed in the model). The structural variability of the models was observed to be higher for pathological amyloids than for functional ones. Finally, structures of the lowest energy are not the ones most frequently produced.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Results

Using *AlphaFold3,* we analyzed six different amyloid proteins. Three pathological amyloids, which are involved in amyloid-related diseases and their aggregation is undesired: Aβ-42, α-synuclein, IAPP (amylin). In addition, we examined three functional amyloids which are evolutionarily programmed to aggregate in order to execute their cellular function: Orb2 (aggregation-prone fragment), CsgA, and Het-s (aggregation prone fragment). Aβ-42 was additionally benchmarked to identify the best number of protein copies used as an *AlphaFold3* parameter. To compare the results with those of other fibrillar structures that are not amyloids, we also studied a collagen model peptide as a negative reference.

Number of protein copies significantly impacts AlphaFold3 predictions

In the first step, we generated fifty Aβ-42 models for different numbers of protein copies, *n* in the range from 1 to 9. *AlphaFold3* provided models of high pLDDT and iPTM (both are measures of model confidence) for monomeric forms (Fig. 1A). For dimers, pLDDT and iPTM values dropped to increase again with *n*, until saturation of the model quality was reached for *n=5*. Clearly, the higher complexity of protein models, more than 5 or 6 protein copies, significantly burdened the quality of the model.

Additionally, we observed that *AlphaFold3* produced helical, not amyloid-like, structures, when only one or two protein copies were provided for structure prediction. It was also visible that the structures were not identical for each *AlphaFold3* prediction. Hence, to assess the influence of the number *n* on the variability of the modelled structures, we calculated the distribution of pairwise TM-score values (measure of structure similarity, TM-score $\epsilon\,[0, 1)$) between all models for each *n* (Supplemental Fig. S1). Aβ-42 monomers, with helical structures, had the highest values of TM-score indicating significant similarity between the generated structures. *AlphaFold3* produced the most structurally varied models for Aβ-42 dimers. In this case, dimers could be both helical, beta-sheet rich, or some mixture of both types.

Based on these observations, we concluded that too low *n* can result in no reproduction of a fibrillar structure, characteristic of amyloids. On the other hand, a too high *n* value would lead to a significant decrease in the model quality predictions.

AlphaFold3 produces models of different qualities for different amyloids

Finally, we modelled the rest of amyloids with six protein copies to balance between the fibril length and the model quality. Collagen was predicted as a trimer. We observed a high variability in the pLDDT and iPTM (and PTM) score values (Figs. 1B and C, Supplemental Fig. S2, Supplemental Table S1) for different amyloids.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

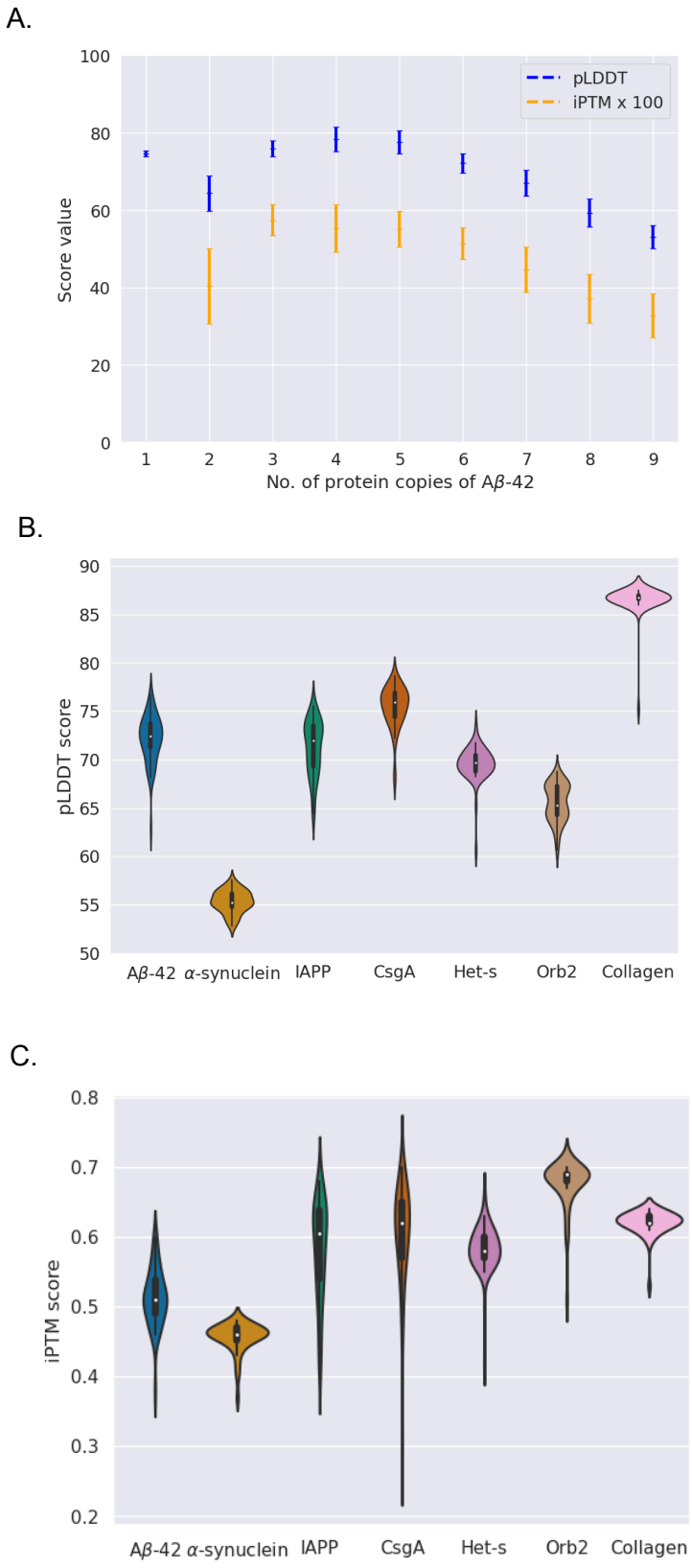


Fig. 1 Distribution of quality metrics for *AlphaFold3* predictions. A. Mean and standard deviation of pLDDT and iPTM for different numbers of Aβ-42 protein copies. B. and C. violin plots of pLDDT and iPTM, respectively, for the amyloid and collagen structures generated.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The worst models were obtained for α-synuclein with the mean pLDDT equal to 55. Similarly, poorer predictions were provided for Orb2 (majority of structures with 50<pLDDT<70). The quality of *AlphaFold3* models for the rest of amyloids was rather similar. No models of a very high quality were produced (pLDDT > 90, *very high*), though the mean quality was usually above 70 (*confident*). For most amyloids, although lower quality models (50<pLDDT<70) were less probable, they could still appear. The PTM and iPTM score distribution frequently contained long-tail characteristics, which was especially visible in the case of CsgA, although all proteins followed this pattern.

Finally, *AlphaFold3* produced very good results for collagen, which has a fibrillar structure though not amyloid-like. The difference in the distribution of pLDDT score for collagen and other amyloids structures was statistically significant (Mann Whitney test p-value<1e-26). These results suggest the troublesome nature of the amyloid fold for structure prediction.

AlphaFold3 structures are not sampled according to physicochemical rules

Next, in light of the seed requirement for *AlphaFold3* predictions, we focused on its sampling methodology. We decided to verify if structures with the lowest energy are the most probable to be generated, as it should be according to the Boltzmann distribution. For each structure, we calculated its energy score using FoldX Delgado et al. (2019) (Fig. 2, Supplemental Table S2). We observed that structures with the lowest energy were not the most likely chosen by *AlphaFold3*. Most frequently it tended to generate the samples around the mean of the energy distribution, which was true for both amyloids and collagen. We also noted that the energy distributions were again long-tailed, with a frequently appearing small peak at some higher energy values. All this implied that *AlphaFold3* structure sampling does not follow the statistical mechanics laws. Finally, collagen fibrils had a much smaller variation in the energy (also when the standard deviation was divided by the mean) than amyloid fibrils, highlighting the previous observations about their lower structural variability, which is reproduced by the algorithm. No correlation between pLDDT and structure energy was observed (|r|<0.2).

AF3 generates structures matching their experimental counterparts

Visual inspection revealed the reproduction of the characteristic beta-cross pattern in all generated structures. We further examined the structural similarity between the generated structures and their experimental counterparts in more detail. For each protein, we compared all pairwise TM-scores between generated amyloid structures and the experimental ones available in the PDB. Then, we extracted the pair with the highest TM-score between a predicted model and its experimental structure. In all cases TM-score was above 0.49. This result indicates a statistically significant similarity between experimental and predicted structures (p-value<1e-6 Xu & Zhang (2010)). We performed structural alignment for each such pair (Fig. 3, Supplemental Table S2) and observed a generally reasonable overlap in each case, indicating the capability of *AlphaFold3* of reproducing the amyloid fold to a certain extent. Evidently, pathological amyloids were a greater challenge than the functional ones and higher deviations from the experimental structures could be observed. The poorest similarity between the best matching model of

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

AlphaFold3 and the experimental counterpart was observed for the amyloid structure of IAPP protein (TM-score = 0.49), and the best for the evolutionary conserved functional CsgA protein (TM-score=0.99).

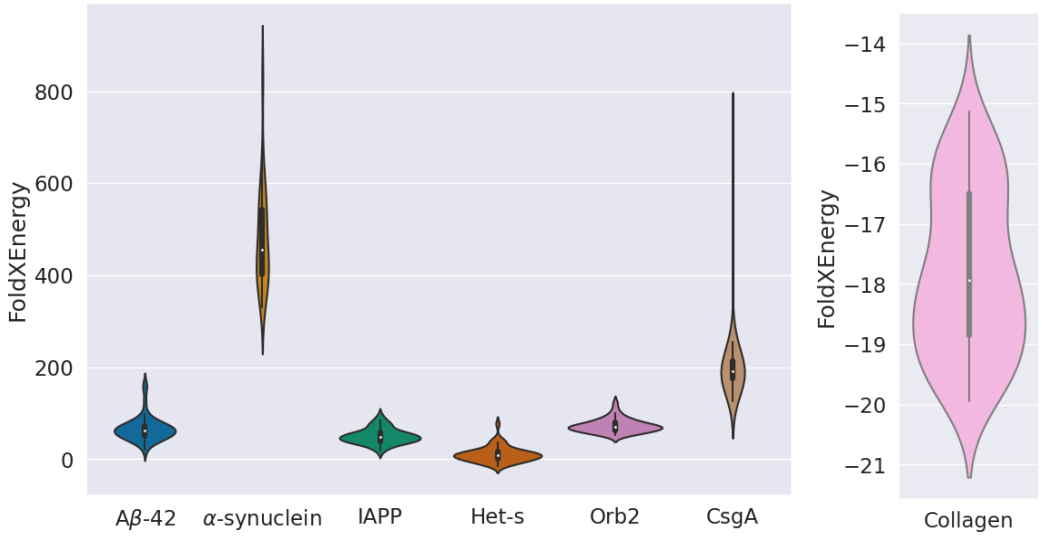


Fig. 2. Structure energy distribution for generated amyloid and collagen model structures.

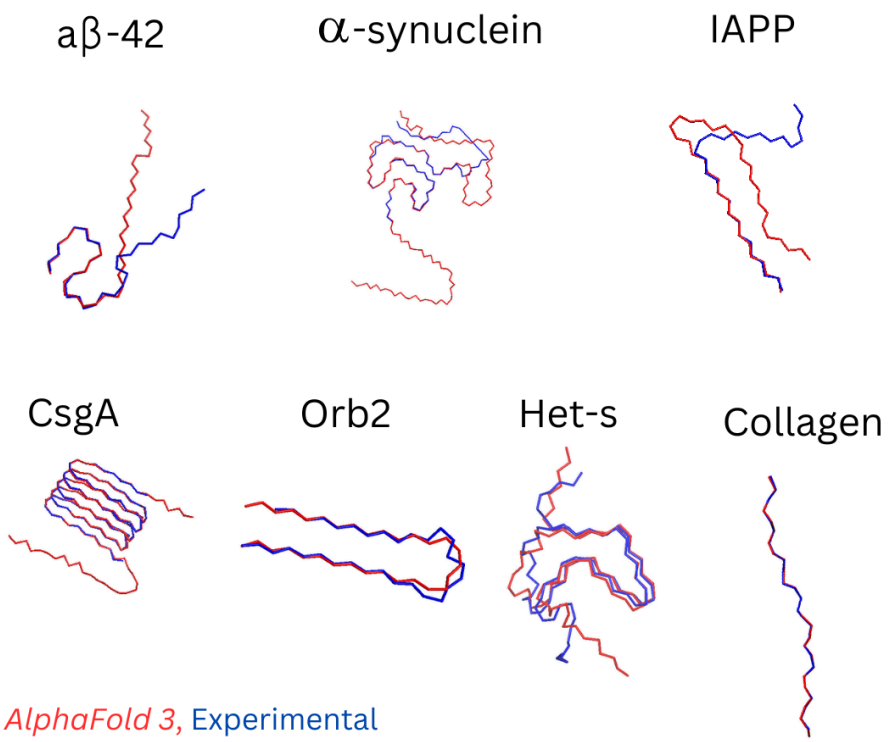


Fig. 3. Best matching pairs between *AlphaFold3* models (in red) and experimental PDB structures (in blue).

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

AlphaFold3 produces polymorphic amyloid structures

For finding the pair which is most similar, we only analyzed a fraction of the structural landscape generated by *AlphaFold3*. Therefore, we decided to broaden our horizons and look at all structures at once. Given the randomness factor implemented in *AlphaFold3*, we hypothesized that it could be capable of reproducing amyloid polymorphism.

To dive into the structural landscape, we compared all pairwise TM-score distances between *AlphaFold3* models and took a look at their distribution (Fig. 3A). For all amyloids, the distribution of TM-score is long-tailed. However, the difference in the distributions of functional and pathological amyloids is striking. In the case of pathological ones, *AlphaFold3* is capable of producing highly varied structures as if two completely unrelated proteins were compared (TM-score≃0). The TM-score distribution is much more concentrated for functional amyloids and collagen. The majority of structures were highly similar to one another (TM-score>0.8) Xu & Zhang (2010).

Next, we compared all structures: predicted and experimental at once, to explore the whole protein structure space. The new TM-score matrices, for each amyloid protein, contained pairwise TM-score values between all structures from the sum of experimental and predicted structure sets. Therefore, we obtained six different distance matrices. Collagen was not considered because its models and experimental structure were very similar. Then, for each amyloid protein, the resulting distance matrix was flattened to a two-dimensional space using PaCMAP dimension reduction Huang et al. (2022) (Fig. 4). Finally, we clustered the structures in the PaCMAP space, using k-means, and extracted a representative structure for each cluster (Fig. 5).

Firstly, we observed that the experimental structures were highly similar to each other, frequently forming an easily visible cluster. On the other hand, the predicted structures varied and only a couple of models would be located close to the experimental ones. *AlphaFold3* frequently sampled the structural space not only in the vicinity of the experimental structures but also a bit further. The highest number of clusters was obtained for Aβ-42, which resembled significant structural variability in the nature. Simple hairpin models and more complex amyloid cores were produced. A similar situation could be observed for α-synuclein. The beta-fold amyloid core structure differed depending on the cluster, although in each case *AlphaFold3* was fairly confident (blue color, 70<pLDDT<90) about its prediction. On the other hand, it struggled with the correct location of C and N-terminals, which varied for different clusters. More α-synuclein structures are presented in Supplemental Fig. S4. For IAPP two cluster representatives differed in the direction of the beta sheets, and the third had three beta folds, instead of two.

Regarding functional amyloids, the majority of the CsgA models were very similar to one another and to the experimental structure, and reproduced the regular beta-cross fold. The variability appeared in the location of the disordered region. Interestingly, in a few cases *AlphaFold3* produced two separated structures, a dimer and a four-unit length fibril (Supplemental Fig. S4). Orb2 structures were even more similar to each other, and the only difference regarded the direction of the beta folds in the hairpin. Het-s fibrils were also highly similar with little variation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The obtained results suggest that *AlphaFold3* captures the concept of amyloid polymorphism. It reproduces the higher structural variability of pathological amyloids and the more conserved structures of the functional ones do not show such variability.

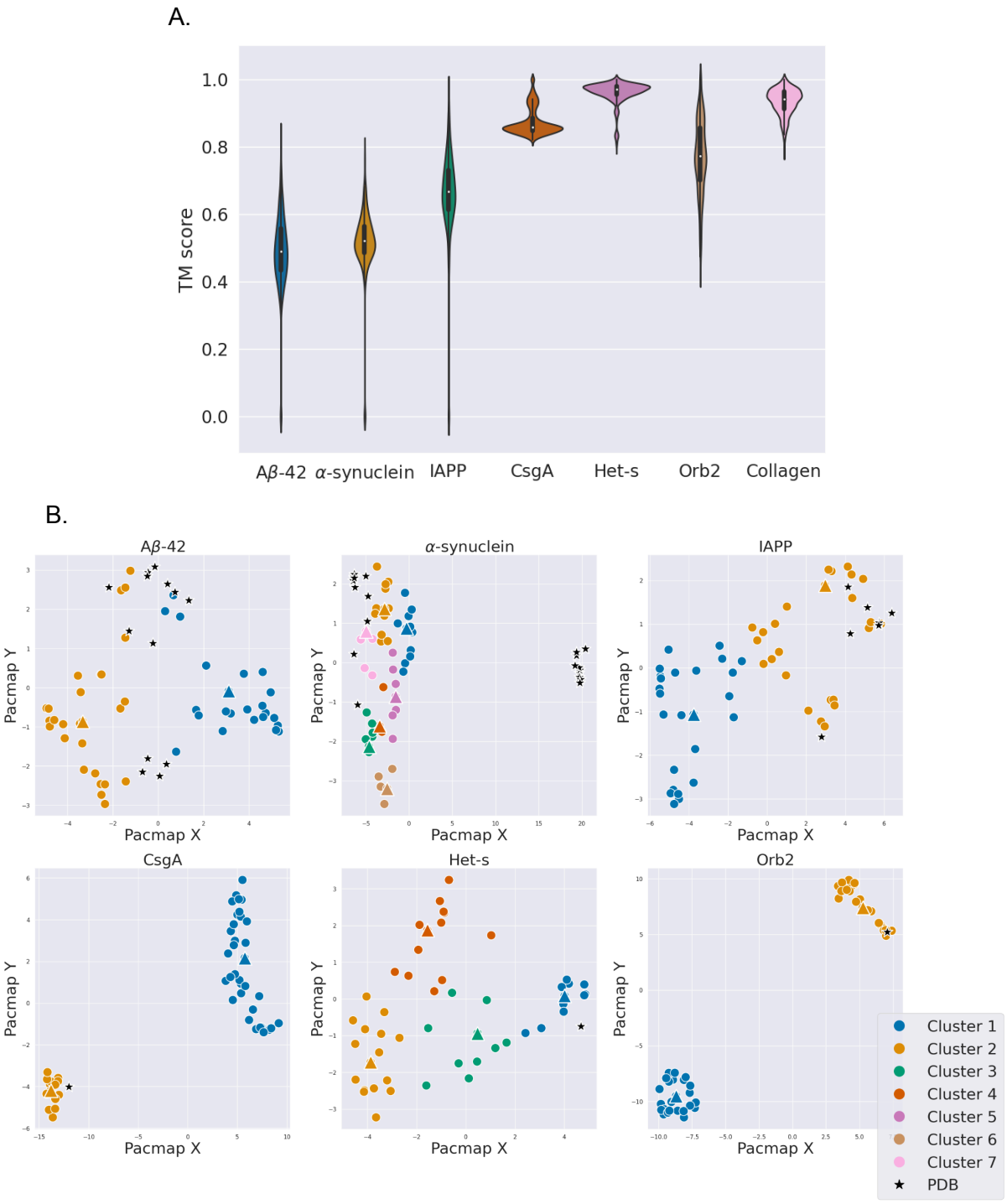


Fig. 4 A. TM-score distribution in the form of a violin plot between generated *AlphaFold3* models. B. Visualization of the amyloid structure space after PaCMAP and clustering with k-means; the triangles refer to cluster representatives.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

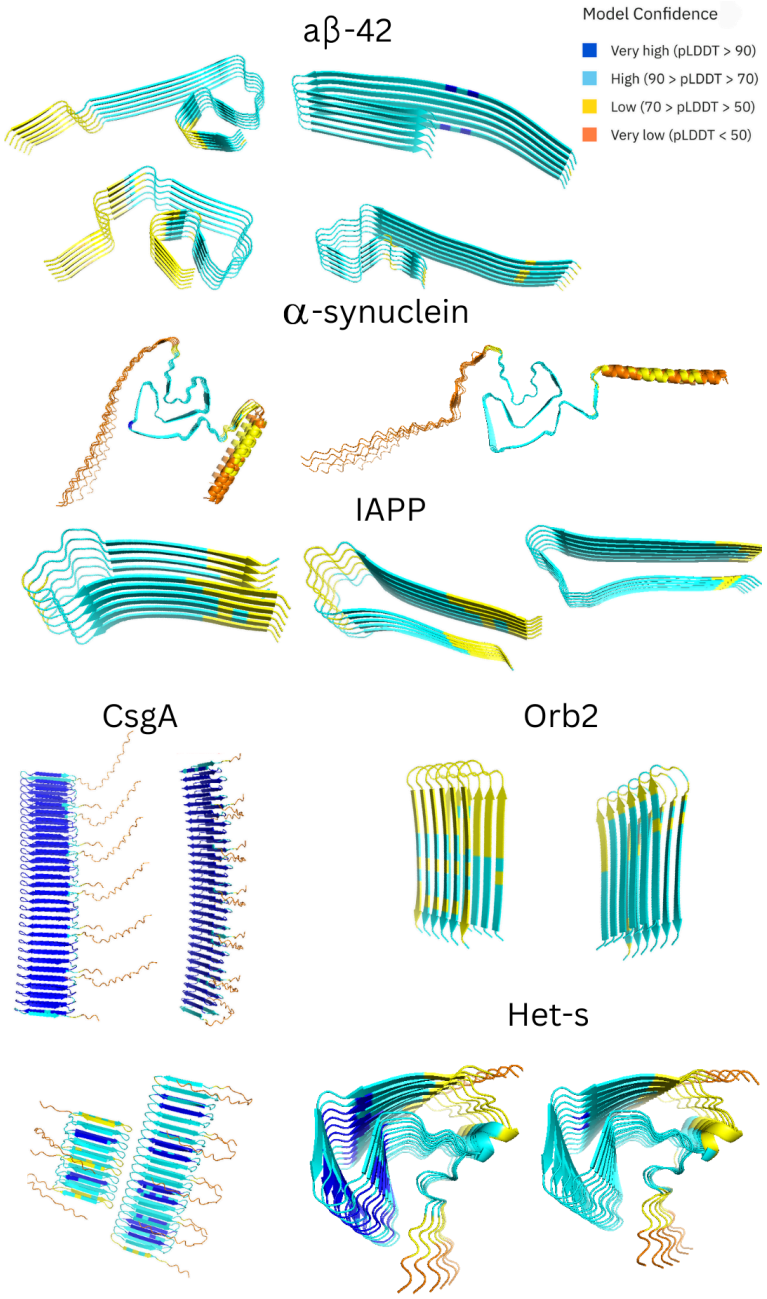


Fig. 5. Cluster representatives for the amyloid structure space produced by *AlphaFold3*.

AlphaFold2 provides poorer results for amyloids than AlphaFold3

Finally, for a comparison, we examined the capability of *AlphaFold2* to reproduce the amyloid fold for our proteins of choice. It failed to provide a beta-cross structure for Aβ-42, α-synuclein, IAPP and Orb2. This result was consistent with Ragonis-Bachar et al. (2023).

For Het-s, the beta-cross fold was reproduced by *AlphaFold2,* although the results somehow differed between both versions of the algorithm. The most matching Het-s model pair produced by *AlphaFold3* and *AlphaFold2* had a TM-score of 0.45. The mean TM-score between the *AlphaFold2* best structure (selected out of 5 provided) and 50 structures

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

produced by *AlphaFold3* was 0.4 (standard deviation = 0.01), the TM-score between *AlphaFold2* structure and the experimental one was 0.37. Finally, the structures of CsgA were highly similar to each other regardless of the *AlphaFold* version. The most similar pair had a TM-score of 0.88, the mean TM-score between the *AlphaFold2* model and *AlphaFold3* models was 0.85 (standard deviation equal to 0.01), the TM-score between the *AlphaFold2* structure and the experimental counterpart was 0.99.

Discussion

In this work, we tested the ability of *AlphaFold3* to model the special class of proteins, mainly amyloids, which have crucial clinical relevance. Amyloids are a great example of complex proteins with a multi-funneled energy landscape, which makes them a significant challenge for machine learning approaches Strodel (2021). Furthermore, their computational studies are limited by insufficient experimental data. Therefore, they provide a great testing set for evaluation of *AlphaFold3* predictions regarding non-typical proteins. We tested pathological amyloids, such as aβ-42, α-synuclein, IAPP and functional amyloids, which are evolutionarily designed to aggregate, i.e., Orb2, CsgA, and Het-s. As a negative control, we also studied a collagen model.

At first sight, a remarkable improvement in structure prediction of amyloids could be observed between *AlphaFold2* and *3*. *AlphaFold2* predicted amyloid structures, incorrectly assuming their helical character Ragonis-Bachar et al. (2023). When benchmarking *AlphaFold3* with different Aβ-42 fibril lengths, we observed helical structure only for monomers or dimers. Longer fibrils lacked helical characteristics, suggesting an important impact of the number of protein copies used for predictions. The difference in the performances of *AlphaFold2* and *3* probably lies in the lower emphasis on the MSA in the later version of the algorithm. Aggregation of pathological amyloids is undesired, which should not be evolutionary conserved in MSA (e.g. highly similar homologs of Aβ-42, identity >90%, do not share the aggregation propensity Tharp & Sarkar (2013)). This could contribute to incorrect prediction of amyloid structures by *AlphaFold2*. On the other hand, the aggregation propensity of functional amyloids is necessary for executing their function in the cell and should be preserved. Both versions of *AlphaFold* correctly predicted the structure of CsgA, which is a functional amyloid with a conserved aggregation propensity between homologs, although the aggregation rate can differ Bhoite (2022).

AlphaFold3, in contrast to its previous version, is a generative algorithm which produces a set of different hypotheses regarding protein structure predictions. We observed that its sampling procedure does not follow thermodynamics laws. The distribution of the structure energy was concentrated around the mean with a frequent bimodal shape, although the second peak was much smaller. Despite that, it is capable of reproducing a characteristic beta-cross pattern typical for amyloids. Although the structures are similar to the experimental ones, the overlap is not as good as for globular proteins and their interactions Abramson et al. (2024). This could be the effect of a low representation of amyloid structures in the PDB.

Interestingly, *AlphaFold3* seems to capture the concept of amyloid polymorphism. It produces a higher structural variability for pathological amyloids, which are not desired to be

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

evolutionary conserved. Prediction of the fibril core frequently has a good pLDDT (>70), meanwhile the exact location of surrounding chains is confusing to the algorithm. Such observations could reflect the rigidity of the core fibril, typical of amyloids, and the flexibility of other parts of the protein. This is supported by Saldaño et al. (2022) who showed that pLDDT anti-correlates with protein flexibility. The more conserved MSA seems to impact the structural variability produced by the model. For functional amyloids, whose aggregation propensity is evolutionary preserved, we observed significantly smaller variation between the models and almost no difference between the predictions for collagen.

In this report, we showed that *AlphaFold3* is capable of dealing with very challenging proteins whose energy landscape is far from typical and whose representation in the structural databases is very low. On the basis of these results, we conclude that *AlphaFold3* is more powerful than *AlphaFold2* in the case of proteins with a lower representation in the PDB database and lower MSA quality data due to its lower MSA reliance. The obtained predictions can capture even more, such as structural polymorphism. Nevertheless, the exact reproducibility of the experimental structure should be treated with caution in the case of non-standard proteins. We also recommend obtaining several sets of predictions as models can vary widely. Furthermore, the produced landscape is not representative of the real one, as structures of the lowest energy are not the most frequently predicted. In addition, in the case of complex predictions, the number of protein copies used in the modelling impacts the predicted structures and their quality, indicating the need for individual benchmarking. Finally, in light of these observations, we call for a desktop version of the *AlphaFold3,* as the limit of 20 structures per day significantly discourages such analyses.

Methods

Extraction of amyloid sequences and structures

For the analysis we chose six amyloids. Three of them belonged to the commonly studied group of human pathological amyloids: Aβ-42, α-synuclein and IAPP (amylin). To account for the evolutionary conserved functional amyloids, we added 3 other proteins: CsgA, Orb2 and Het-s, for which some experimental structures are available. In the case of Orb2 and Het-s, we generated models just for aggregation-prone regions, as only those were experimentally resolved. CsgA was modeled without its signal peptide. The exact sequences along with used PDB structures are in Supplemental Table S4.

For each amyloid protein, we generated fifty *AlphaFold3* models with six protein copies. The PTM (predicted TM-score, quality of individual subunits prediction, PTM $\epsilon\,[0,1)$), iPTM (interface predicted TM-score, quality of a whole complex prediction, iPTM $\epsilon\,[0,1)$) and pLDDT scores (per-residue *predicted local distance difference test*, pLDDT $\epsilon\,[0,100]$) were extracted from the generated files Evans et al. (2021). As collagen exists in a triple formation, here three protein copies were used when generating models.

Energy calculations

To perform energy calculation, we used pyFoldX and FoldX software Delgado et al. (2019). Each AlphaFold3 structure was first minimized according to the FoldX force field, using the

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

following command `foldx --command=RepairPDB --pdb=structure.pdb`. Then the structure energy was calculated using the `getTotalEnergy()` function implemented in pyFoldX Radusky & Serrano (2022).

Structure comparison and visualization

The structures were aligned using TM-align software Zhang & Skolnick (2005), which also provided TM-score values (structure similarity measure, TM-score $\epsilon\ [0, 1)$). For each protein, we calculated all pair TM-scores between all models and experimental structures, hence generating a pairwise TM-score distance matrix for the sum of sets of all models and experimental structures.

We performed dimension reduction of distance matrix M using the PaCMAP method Huang et al. (2022). The clusterization with k-means was performed with scikit-learn Python package Kramer (2016). For each protein, we identified the optimal number of clusters from 2 to 8 by maximizing Silhouette score. Structures were visualized using PyMol software Yuan et al. (2017).

General data analysis

All the data analyses were performed in Python 3.11 using the following packages: pandas Reback et al. (2020), biopandas Raschka (2017), numpy Harris et al. (2020), matplotlib Hunter & Dale (2007), seaborn Waskom (2021).

Funding

This work was partially supported by the National Science Centre, Poland [2019/35/B/NZ2/03997].

References

Abramson, Josh, et al. "Accurate structure prediction of biomolecular interactions with AlphaFold3." Nature (2024): 1-3.

Bhoite, Sujeet S., et al. "Mechanistic insights into accelerated α-synuclein aggregation mediated by human microbiome-associated functional amyloids." Journal of Biological Chemistry 298.7 (2022).

Brown, Benjamin P., et al. "Approximating conformational Boltzmann distributions with AlphaFold2 predictions." bioRxiv (2023).

Chakravarty, Devlina, and Lauren L. Porter. "AlphaFold2 fails to predict protein fold switching." Protein Science 31.6 (2022): e4353.

Dabrowski-Tumanski, Pawel, and Andrzej Stasiak. "Alphafold blindness to topological barriers affects its ability to correctly predict proteins' topology." Molecules 28.22 (2023): 7462.

Delgado, Javier, et al. "FoldX 5.0: working with RNA, small molecules and a new graphical interface." Bioinformatics 35.20 (2019): 4168-4169.

Dong, Runze, et al. "mTM-align: an algorithm for fast and accurate multiple protein structure alignment." Bioinformatics 34.10 (2018): 1719-1725.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.09.602655; this version posted July 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Evans, Richard, et al. "Protein complex prediction with AlphaFold-Multimer." biorxiv (2021): 2021-10.
Harris, Charles R., et al. "Array programming with NumPy." Nature 585.7825 (2020): 357-362.
Hunter, John, and Darren Dale. "The matplotlib user's guide." Matplotlib 0.90. 0 user's guide (2007).
Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." nature 596.7873 (2021): 583-589.
Kotulska, Małgorzata, and Jakub W. Wojciechowski. "Bioinformatics methods in predicting amyloid propensity of peptides and proteins." Computer Simulations of Aggregation of Proteins and Peptides. New York, NY: Springer US, 2022. 1-15.
Kramer, Oliver, and Oliver Kramer. "Scikit-learn." Machine learning for evolution strategies (2016): 45-53.
Lutter, Liisa, Liam D. Aubrey, and Wei-Feng Xue. "On the structural diversity and individuality of polymorphic amyloid protein assemblies." Journal of Molecular Biology 433.20 (2021): 167124.
Monteiro da Silva, Gabriel, et al. "High-throughput prediction of protein conformational distributions with subsampled AlphaFold2." nature communications 15.1 (2024): 2464.
Otzen, Daniel, and Roland Riek. "Functional amyloids." Cold Spring Harbor perspectives in biology 11.12 (2019): a033860.
Radusky, Leandro G., and Luis Serrano. "pyFoldX: enabling biomolecular analysis and engineering along structural ensembles." Bioinformatics 38.8 (2022): 2353-2355.
Ragonis-Bachar, Peleg, et al. "What can AlphaFold do for antimicrobial amyloids?." Proteins: Structure, Function, and Bioinformatics 92.2 (2024): 265-281.
Raschka, Sebastian. "BioPandas: Working with molecular structures in pandas DataFrames." J. Open Source Softw. 2.14 (2017): 279.
Reback, Jeff, et al. "pandas-dev/pandas: Pandas 1.0. 5." Zenodo (2020).
Saldaño, Tadeo, et al. "Impact of protein conformational diversity on AlphaFold predictions." Bioinformatics 38.10 (2022): 2742-2748.
Strodel, Birgit. "Energy landscapes of protein aggregation and conformation switching in intrinsically disordered proteins." Journal of molecular biology 433.20 (2021): 167182.
Tharp, William G., and Indra Neil Sarkar. "Origins of amyloid-β." BMC genomics 14 (2013): 1-15.
Waskom, Michael L. "Seaborn: statistical data visualization." Journal of Open Source Software 6.60 (2021): 3021.
Wayment-Steele, Hannah K., et al. "Predicting multiple conformations via sequence clustering and AlphaFold2." Nature 625.7996 (2024): 832-839.
Wilson, Carter J., Wing-Yiu Choy, and Mikko Karttunen. "AlphaFold2: a role for disordered protein/region prediction?." International Journal of Molecular Sciences 23.9 (2022): 4591.
Xu, Jinrui, and Yang Zhang. "How significant is a protein structure similarity with TM-score= 0.5?." Bioinformatics 26.7 (2010): 889-895.
Yin, Rui, et al. "Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants." Protein Science 31.8 (2022): e4379.
Yuan, Shuguang, HC Stephen Chan, and Zhenquan Hu. "Using PyMOL as a platform for computational drug design." Wiley Interdisciplinary Reviews: Computational Molecular Science 7.2 (2017): e1298.
Zhang, Yang, and Jeffrey Skolnick. "TM-align: a protein structure alignment algorithm based on the TM-score." Nucleic acids research 33.7 (2005): 2302-2309.