AlphaFold can be used to predict the oligomeric states of proteins

Yiechang Lin[†*], Ciara Wallis[†], Ben Corry

Research School of Biology, The Australian National University, Canberra, ACT 2600, Australia

* Correspondence: Yiechang.Lin@anu.edu.au

† Co-first authors

Homooligomerisation is a prevalent and important process that many proteins undergo to form the quaternary structure required to carry out their biological functions. However, determining the oligomeric states and structures of some proteins through experimental methods remains technically challenging and time consuming. Here, we show that the protein structure prediction tools AlphaFold2-Multimer and AlphaFold3 can be used to quickly and accurately predict the oligomeric states and structures of soluble and membrane proteins. We provide optimal parameters for minimizing computational cost while maintaining accuracy, apply this technique to several membrane proteins of interest, and highlight instances where multiple physiologically relevant oligomeric states of a given protein can be predicted using this method.

Introduction

Between 30-50% of all proteins are thought to self-associate into stable complexes consisting of multiple copies of themselves.[1,2] This oligomerisation process is important for the structural and functional properties of these proteins, as it enables the formation of repetitive structural elements, catalytic interfaces in enzymes and the central pore in many membrane proteins. Establishing the biologically relevant oligomeric state of proteins, and the structures they adopt is thus critical for understanding the way in which they perform their biological roles.

A variety of biochemical and biophysical techniques such as native mass spectrometry, size exclusion chromatography and native gel electrophoresis are available to characterize the oligomeric states of proteins.[3] However, these experimental techniques can sometimes be expensive, time-consuming and limited by technical constraints. Existing computational tools to characterize protein oligomerisation states include sequence similarity-based methods which predict the oligomerisation state using related proteins with resolved structures.[4] Once the number of subunits is known, a model of the oligomer can be built using a structural templating technique if a relevant template exists or ab initio docking where this is not possible. However, this method relies on the existence of oligomeric state information from a related protein. More recently developed deep learning tools trained on large datasets of curated subunit information from experimental data predicts oligomeric states with relatively high accuracy but do not provide structural information.[5,6] Currently, for yet to be characterized proteins, the AlphaFold Protein Structure Database (AFDB) provides predicted monomeric structures but lacks predictions for functionally relevant multimeric states which may provide key insights into their physiological behaviour.[7]

Since its release, AlphaFold2 Multimer has found wide application - including to identify plant-pathogen interactions across two different species,[8] to predict protein-protein complexes in yeast,[9] and capture interactions between intrinsically disordered protein regions.[10] Additionally, recent proteome wide predictions using AlphaFold2 highlighted the high proportion of proteins which form homomers and highlight the potential for structural prediction tools to model protein homooligomerization.[2] The recently released AlphaFold3 provides further improvements in accuracy for the modelling of protein-protein interactions, in addition to the ability to model structures of other biological molecules.[11]

Here, we investigate whether AlphaFold-Multimer v.2.3 (AF2-M) and AlphaFold3 (AF3) can correctly predict the oligomeric states of proteins and their structures. We show that the interface predicted TM (ipTM) score reported by both AF2-M and AF3 is an accurate predictor of the oligomeric state, that structures of oligomers resemble experimentally resolved structures, and present optimised parameters for the application of this method to predict the oligomerisation state of yet uncharacterised proteins of interest. Additionally, we apply this technique to a select set of membrane proteins and highlight several cases where the ipTM distribution correctly predicts the presence of multiple biologically relevant oligomeric states.

Results

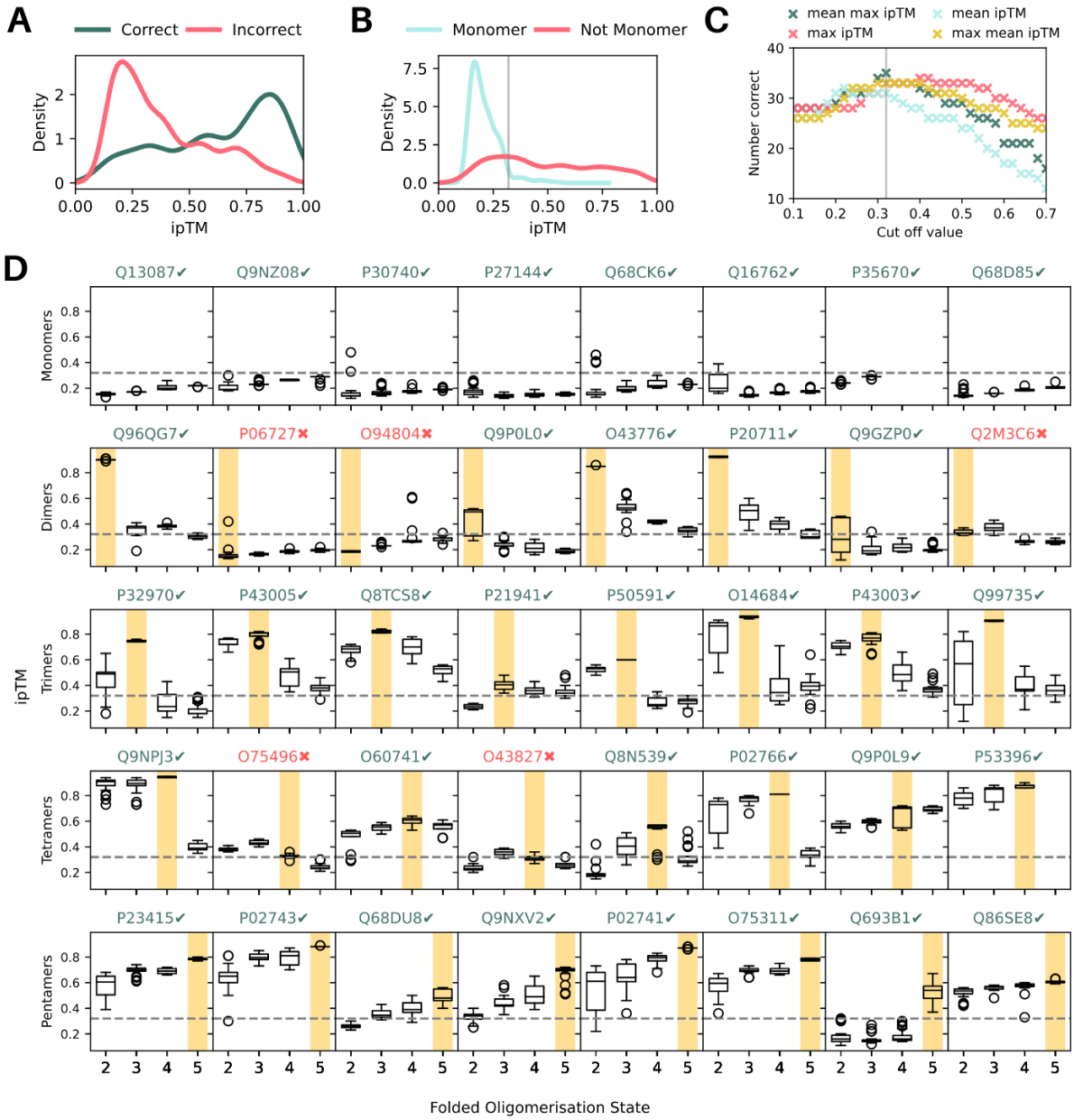

Figure 1. ipTM scores from AlphaFold2 Multimer correctly predict oligomeric states

(A) Probability distribution plot of ipTM scores for 32 known multimeric proteins predicted in the correct (green) and incorrect (pink) oligomeric states.

(B) Probability distribution plot of ipTM scores for 8 known monomeric proteins predicted with 2,3,4,5 subunits (blue) and 32 known multimeric proteins predicted with 2,3,4,5 subunits (pink). The grey line indicates the cut-off ipTM value of 0.32.
(C) Number of correctly assigned oligomeric states (out of 40) vs cut-off value selected for 4 different methods to distinguish between monomeric and multi-subunit proteins. The grey line indicates the optimal cut-off ipTM value of 0.32.
(D) Box plots showing individual ipTM scores for different folded oligomeric states across the test set of 40 proteins with the correct oligomeric state highlighted in yellow. The dashed line indicates the cut-off ipTM value of 0.32. UniProt IDs are noted in green for correct predictions and red for incorrect ones.

First, we asked whether the ipTM score reported by AF2-M, which measures the accuracy of the predicted relative positions of subunits within a protein-protein complex, can be used to correctly predict known protein oligomeric states.

Using the dataset curated from the UniProt Database by Deng et al. (2024), for their study on oligomeric state prediction using a deep learning approach,[6] we extracted a sample of 40 proteins with known oligomeric states (Table S1). This sample was selected to contain 8 examples each of proteins known to have oligomeric states of 1, 2, 3, 4 and 5 and contained a mix of soluble and membrane proteins (6 single-pass and 10 multi-pass). For each protein, we used AF2-M to predict their structures as homodimers, homotrimers, homotetramers and homopentamers. We used 3 recycles and predicted 20 structures in each oligomeric state for a total of 80 structures produced per protein.

For the 32 multimeric proteins in our dataset, the distribution of ipTM scores for all structures predicted in the correct oligomeric state (e.g., a known dimeric protein predicted with two subunits) differs considerably from the distribution found for all structures predicted in an incorrect oligomeric state (Fig. 1A). In general, higher ipTM scores are observed (peak at ipTM 0.75-1) for proteins predicted in their annotated oligomeric states compared to an incorrect state (peak at ipTM 0-0.32). While there exists some overlap between the distributions, the presence of distinct differences between these peaks suggests the ipTM score reported by AF2-M can distinguish between correctly and incorrectly predicted oligomeric states. For the 8 known monomeric proteins in our test set, predicting them as homodimers, homotrimers, homotetramers, or homopentamers consistently yielded low ipTM scores (0-0.4). The distribution of their ipTM scores differs substantially to that of the known 32 known multimeric proteins predicted in any oligomeric state (Fig. 1B), suggesting that AF2-M can also distinguish between proteins likely to be monomeric and those forming higher order protein assemblies.

These data indicate that we may be able to predict a protein's most likely oligomeric state by selecting the state which yields the highest mean or max ipTM score (calculated across the 20 structures predicted for each protein) if we can firstly define a cut-off value to assign a protein as monomeric, since monomeric predictions do not have an associated ipTM score. A cut-off value that is too low would lead to incorrect assignment of some monomeric proteins as multimeric and a too high cut-off value could lead to incorrect assignment of some multimeric proteins as monomeric. We optimised the cut-off value across four methods – taking the mean of all ipTM scores for a protein (mean), using the mean of the highest ipTM value for each oligomeric state (mean max), using the maximum of the mean ipTM value for each oligomeric state (max mean) and taking the maximum ipTM score for the protein (max) (Fig. 1C). We found that taking the mean of the maximum ipTM score of each oligomeric state for the protein lead to the highest number of correct states assigned when we optimise this cut-off value to 0.32 and use the max ipTM to assign the oligomeric state if it is not a monomer.

Using this method, we find that 35/40 proteins in our dataset are correctly assigned. Individual ipTM distributions across the 40 proteins highlights trends across the different types of oligomeric states (Fig. 1D). Across monomeric proteins, predictions at all oligomeric states yield low ipTM scores. In this set of proteins, we find that all monomeric, homotrimeric and homopentameric proteins were assigned to the correct oligomeric state. Three homodimers were incorrectly assigned – as a monomer (UniProt ID: P06727), tetramer (UniProt ID: O94804) and trimer (UniProt ID: Q2M3C6). Additionally, two homotetramers

were incorrectly assigned as trimers (UniProt IDs: O7549, O43827). Interestingly, all incorrectly assigned proteins had low ipTM scores across all oligomeric states, even those that were not ultimately assigned as monomeric, suggesting that the absolute values of the ipTM scores may be used to gauge confidence in the prediction if a protein is predicted as a homooligomer.

We note that the annotated oligomeric states for each protein typically have smaller ipTM ranges when compared to other oligomeric states (Fig. S1A). Additionally, the average pLDDT (a per-residue measure of local confidence) cannot be used to distinguish between annotated and other oligomeric states, as the monomeric form usually has the highest pLDDT score (Fig. S1B).

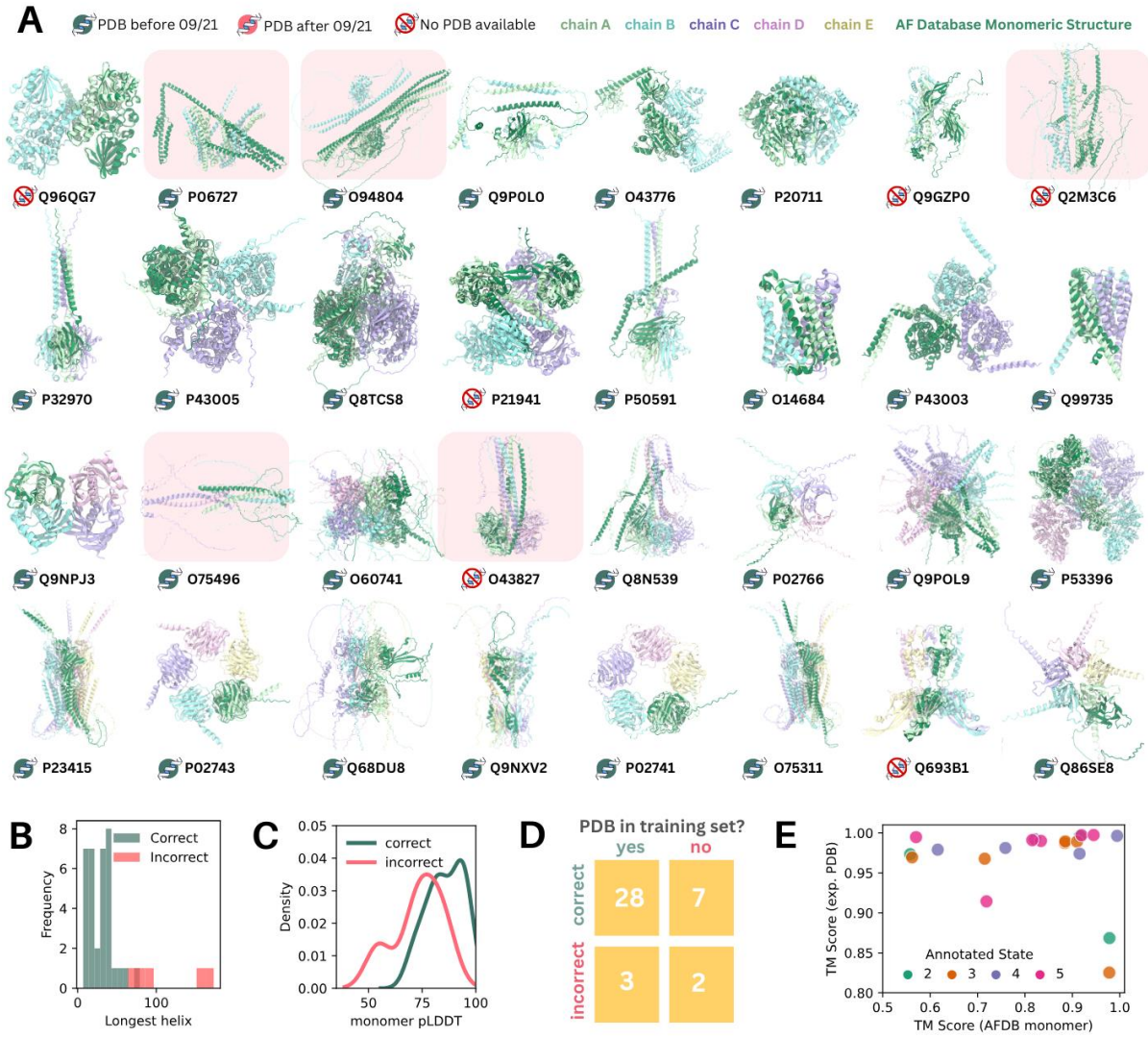

Figure 2. Multimeric structures predicted by AlphaFold2 resemble experimentally determined structures

(A) Snapshots of the top ranked structural predictions in the correct oligomeric state for each multimeric protein (ribbon, coloured by chain), aligned with the AFDB prediction of the protein as a monomer (ribbon, dark green). The presence of an experimentally resolved structure in the PDB or AlphaFold2 training set is indicated with an icon near each UniProt ID. The structures of proteins predicted in an incorrect oligomeric state are highlighted in red.

(B) Histogram showing the length of the longest helix in the corresponding AFDB monomer structure for proteins predicted in the correct (green) and in correct (pink) oligomeric states.

(C) Probability distribution plot of average monomeric pLDDT values (using the AFDB structure) for proteins predicted with the correct (green) and incorrect (pink) oligomeric states.

(D) Two-way table showing the distribution of proteins with oligomeric states predicted correctly and incorrectly based on whether an experimentally resolved structure exists within the AlphaFold2 training set.

(E) TM score between the top ranked AlphaFold2 predicted structure and experimentally resolved structure (if available) plotted against the TM score between the top ranked AlphaFold2 predicted structure and the monomeric structure from the AFDB, coloured by annotated oligomerisation state.

Next, we examined whether the structures of the oligomeric assemblies predicted resembled the monomeric structures in the AFDB and experimentally determined structures. Snapshots of the top ranked structural predictions in the correct oligomeric state were produced for each protein after aligning to the monomeric AFDB structure (Fig. 2A).

Proteins where the oligomeric state was incorrectly predicted by AF2-M all have abnormally long helices in both the monomeric and oligomeric structures (Fig. 2A-B), with the longest helices spanning 64-173 residues. We examined also whether the average monomeric pLDDT, calculated using the structure present in the AFDB indicates whether the correct oligomeric state will be predicted (Fig. 2C). In this sample set, if the pLDDT value was > 90 (14 structures), then the correct oligomeric state was assigned. If the pLDDT value was < 50 (1 structure), then an incorrect oligomeric state was predicted.

The presence of an experimentally resolved structure in the AlphaFold2.3 training set (prior to 30 September 2021) did not have a large impact on whether the correct oligomeric state was predicted in this set of proteins (Fig. 2D). 28/31 of structures present in the training set were correctly predicted compared to 7/9 structures absent from the training set being correctly predicted.

Next, we quantified the similarity between the predicted multimeric structures to both experimentally resolved structures and to AF2 monomeric predictions found in the AFDB in the cases for which the oligomeric state was correctly predicted and an experimentally resolved structure exists (Fig. 2E, Table S2). The TM score between the computationally predicted and experimentally determined structures are generally very high across all oligomerisation states, with most predicted structures having TM scores of greater than 0.95 to experimentally resolved structures. This suggests that predicted structures of higher order oligomers are likely to be accurate. In many cases, the predicted structures of subunits in the multimer showed significant differences to the AF2 predicted monomeric structure (Fig, 2E), suggesting changes to the individual subunit structure imposed by oligomer formation. Visual analysis of these cases (e.g. UniProt ID: P5O591) indicates that this is usually due to changes in the relative orientations of protein domains and highlights a limitation of using predicted monomeric structures from the AFDB.

Taken together, this suggests that AF2-M can predict the oligomeric states and structures of homomeric proteins accurately, independent of whether the protein structure is present in the training set. The oligomeric state prediction is likely to be accurate if the monomeric pLDDT is high and there are no abnormally long alpha helices in the monomeric structure.

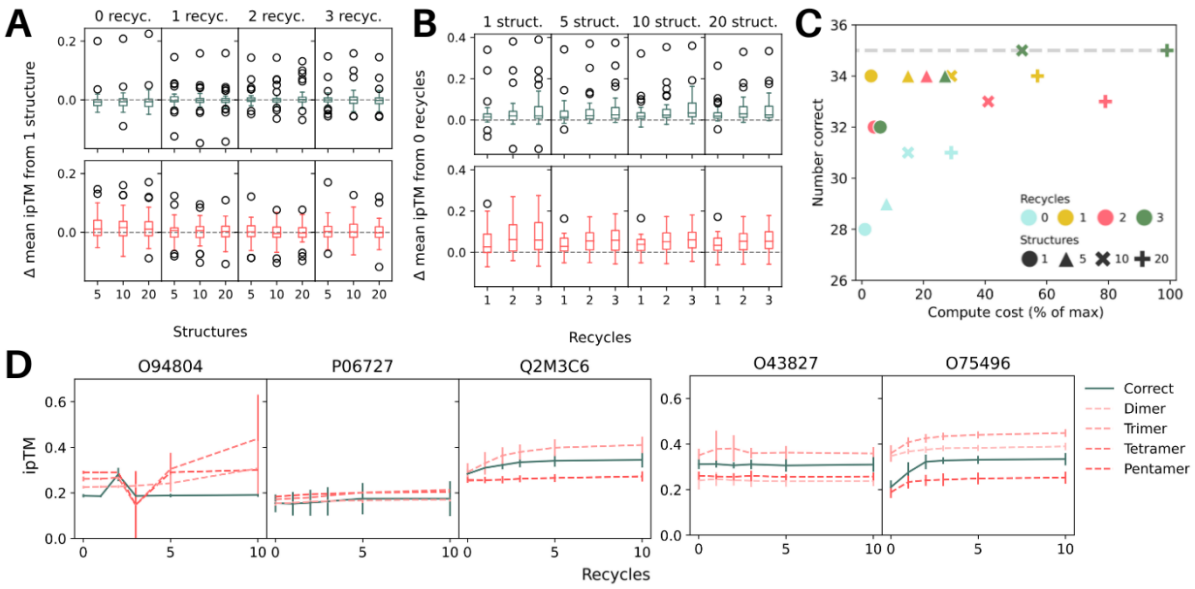

Figure 3. Optimising the computational cost

(A) Distribution of change in mean ipTM from 1 structure and either 0,1,2,3 recycles versus the number of predicted structures for proteins folded in the same oligomeric state as annotated (top, green) and in a different state than annotated (bottom, pink). Each datapoint represents one protein.

(B) Distribution of change in mean ipTM from 0 recycles and either 1,5,10 and 20 structures versus the number of recycles predicted for proteins folded in the same oligomeric state as annotated (top, green) and in a different state than annotated (bottom, pink). Each datapoint represents one protein.

(C) Number of correct oligomeric state predictions (using monomeric cut off value of 0.32) vs % of max compute cost across 0,1,2,3 recycles and 1,5,10,20 structures predicted. The dashed grey line indicates the highest number of correct predictions (35) across all parameters.

(D) ipTM scores of the annotated (green) and incorrect (red, various line styles) states across 0, 1, 2, 3, 5, and 10 recycles for incorrectly assigned dimers (left) and incorrectly assigned tetramers (right).

To optimise the computational cost associated with making these predictions, which can be considerable for larger proteins, we next assessed how the chosen number of recycles and number of predicted structures influences the accuracy of oligomeric state prediction. We carried out predictions using combinations of 0, 1, 2, 3 recycles and 1, 5, 10, 20 structures per condition using same set of proteins and assessed how these parameters altered the ipTM scores of the structures predicted in the annotated and incorrect oligomeric states.

Increasing the number of structures from 1 to 5, 10 or 20 does not generally increase the mean ipTM score per protein (Fig. 3A). This is true for predictions of proteins in both the annotated (green, top row) and other (red, bottom row) oligomeric states. However, for a few proteins (outlier values in box plots), increases in ipTM by up to 0.2 are observed when 5, 10 or 20 structures are predicted when compared to 1 structure. Increasing the number of recycles to 1, 2 or 3 generally increases the mean ipTM score compared to 0 recycles, particularly for structures predicted in their annotated (green, top) oligomeric states with 5, 10 or 20 structures (Fig. 3B). The mean ipTM score for proteins predicted in incorrect oligomeric states (red, bottom) also generally increases, however, for some proteins, additional recycles decreases the mean ipTM score instead.

The comparison between compute cost and number of proteins predicted in the correct oligomeric state suggests that 3 recycles and 10 structures is the most computationally efficient set of parameters, predicting 35/40 proteins in the correct oligomeric state (Fig. 3C). Producing 5 structures with 1,2, or 3 recycles also each yielded 34/40 correct predictions for a lower compute cost. The worst results are obtained using no recycles.

For the five proteins which were predicted incorrectly using 3 recycles and 20 structures, increasing the number of recycles to 5 or 10 does not correct the prediction and ipTM scores for structures predicted in both the incorrect and annotated oligomeric states do not further increase after 5 recycles (Fig. 3D).

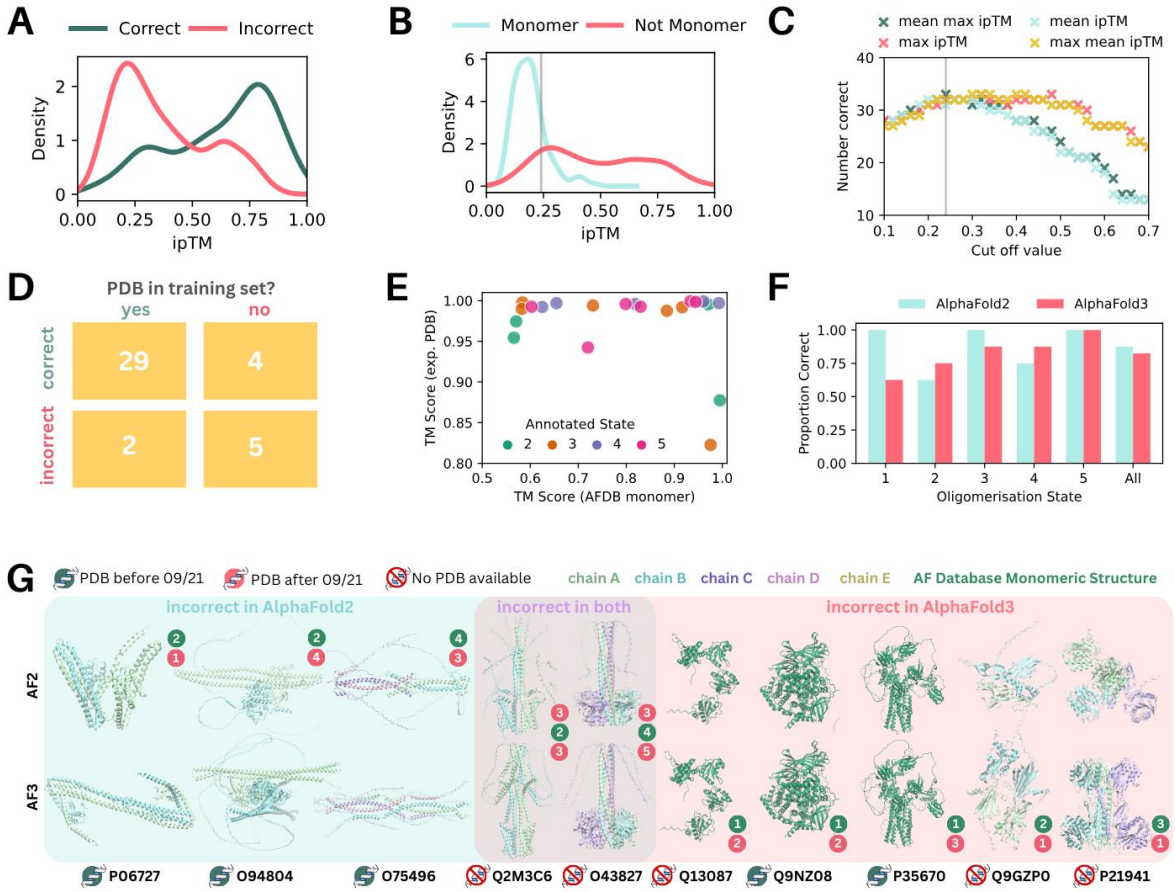

Figure 4. ipTM scores from the AlphaFold3 online server also accurately predict oligomeric states

(A) Probability distribution plot of ipTM scores for 32 known multimeric proteins predicted in the correct (pink) and incorrect (green) oligomeric states.

(B) Probability distribution plot of ipTM scores from AF3 for 8 known monomeric proteins predicted with 2,3,4,5 subunits (blue) and 32 known multimeric proteins predicted with 2,3,4,5 subunits (pink). The grey line indicates the cut-off ipTM value of 0.24.

(C) Number of correctly assigned oligomeric states (out of 40) vs cut-off value selected for 4 different methods to distinguish between monomeric and multi-subunit proteins. The grey line indicates the optimal cut-off ipTM value of 0.24.

(D) Two-way table showing the distribution of proteins with oligomeric states predicted correctly and incorrectly based on whether an experimentally resolved structure exists within the training set.

(E) TM score between the top ranked AlphaFold3 predicted structure and experimentally resolved structure (if available) plotted against the TM score between the top ranked AlphaFold3 predicted structure and the monomeric structure from the AFDB, coloured by annotated oligomerisation state.

(F) Comparison of AlphaFold2 and AlphaFold3 prediction accuracy by annotated oligomerisation state.

(G) Snapshots of the top ranked structural predictions generated by AF2-M (top row) and AF3 (bottom row) in the correct oligomeric state for proteins that were assigned to an incorrect oligomeric state by AF2-M, AF3, or both AF2-M and AF3 (ribbon, coloured by chain). For monomeric proteins, the AFDB monomeric structure is shown in dark green. The presence of an experimentally resolved structure in the PDB or AlphaFold2 training set is indicated with an icon near each UniProt ID. The correct oligomerisation state is indicated by each structure in green and the incorrect prediction by the associated method in red.

While availability of the source code for AF2-M allowed us to generate structural predictions for many proteins without restriction, we wondered whether the online AlphaFold3 server which provides an

accessible user interface to predict structural complexes could also be used to predict oligomeric states. To assess this, we carried out structural predictions of the 40 proteins as dimers, trimers, tetramers and pentamers as above using the default AF3 parameters of 10 recycles, 5 structures.

As with AF2-M, the ipTM scores of predicted oligomer structures can distinguish between proteins predicted in the annotated state versus an incorrect oligomeric state (Fig. 4A), as well as between monomeric and multimeric proteins (Fig. 4B). By optimising for the monomeric cutoff value to 0.24 as we did when using AF2-M, 33/40 of the proteins in our test set were correctly assigned (Fig. 4C, Fig. S2-3).

As AF3 uses PDB templates by default where available, we find most proteins (29/31) where a structure was present in the training set is predicted in the correct oligomeric state (Fig. 4D). In contrast, 5/9 structures which are not present in the training set were predicted in the incorrect oligomeric state.

Similar to AF2-M, the majority of predicted oligomeric structures had high TM scores to the experimentally resolved structures where available, but a lower correlation of TM scores to the predicted monomeric structures (Fig. 4E). Similar TM scores were obtained for the same proteins across AF2-M and AF3, and proteins which were predicted with lower TM scores using AF2-M were also less like the experimental structures when predicted using AF3.

Compared to AlphaFold2, AlphaFold3 assigned more dimeric and tetrameric proteins correctly (Fig. 4F), although the significance of this is hard to assess given the size of the data set. Structures predicted in an incorrect oligomeric state by AF2-M but had a PDB available were predicted in the correct oligomeric state by AF3 (Fig. 4G). Interestingly, 5/7 of the incorrect state assignments using AF3 was due to difficulty distinguishing between monomeric and homooligomeric proteins. Three monomeric proteins were predicted to form higher order oligomers (2 dimers, 1 trimer) and one dimeric and one trimeric protein was incorrectly assigned as monomeric.

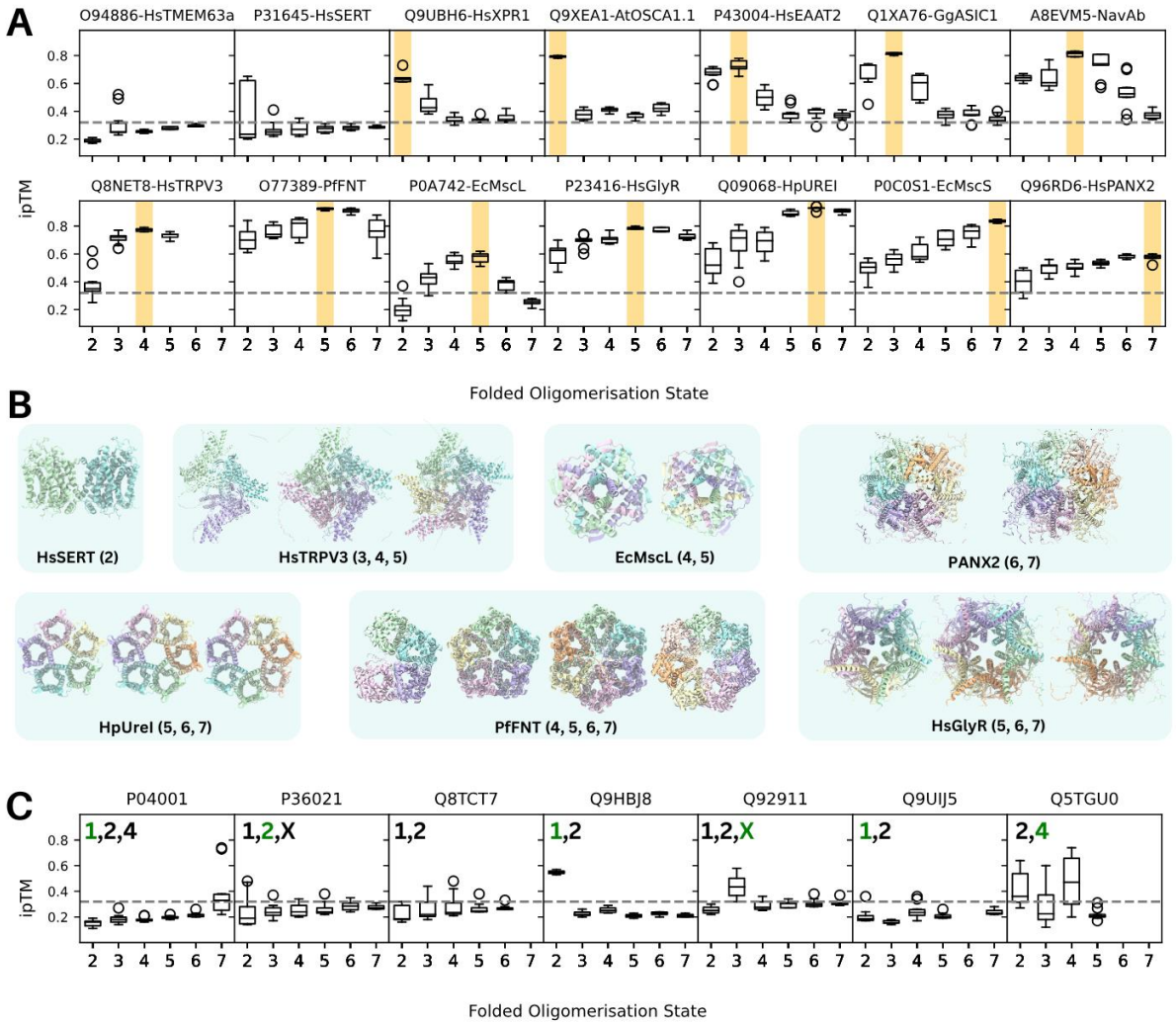

Figure 5. Application to membrane proteins

(a) Box plots showing individual ipTM scores for different folded oligomeric states (2-7) across a curated set of 14 membrane proteins with the correct oligomeric state highlighted in yellow. The dashed line indicates the previously determined AF2-M cut-off ipTM value of 0.32. UniProt IDs and common protein names are shown.

(b) Structural predictions of select membrane proteins in various oligomeric states. Protein names and oligomeric states shown are given below each set of structures.

(c) Boxplots showing IpTM distributions for 7 additional membrane proteins with multiple annotated oligomeric states (shown in each box) noted within each panel, with those correctly assigned by AF2-M shown in green. X denotes annotation as a 'homooligomer' in the UniProt Database

Next, we sought to apply this technique of oligomerisation state and structure prediction to membrane proteins. As many pore forming membrane proteins are multimeric, their oligomeric state is crucial for understanding their physiology. We curated a set of 14 membrane proteins with a diversity of known oligomeric states (monomeric to heptameric) of interest in our lab and used these to assess AF2-M's ability to correctly predict their oligomeric states (Fig. 5A, Table S3).

In this curated set, all proteins which form higher order oligomers (dimers and above) were predicted in the correct oligomeric states. However, two proteins which we expected to be assigned as monomeric (HsTMEM63a, UniProt ID: O94886 and HsSERT, UniProt ID: P31645) were assigned as trimers and dimers respectively using our monomeric cut-off value of 0.32. Interestingly, we note that for HsSERT, the top ranked structures of the dimeric form of the protein (as shown in Fig 5B) have the highest ipTM score (Fig 5A). Dimerisation and formation of higher order oligomers for HsSERT and other SLC6 members have been

previously reported, hinting that the oligomerisation state predicted here could be biological relevant.[12] Additionally, the top ranked predicted dimeric structure of HsSERT resembles one dimer conformation sampled in previously conducted molecular dynamics simulations probing HsSERT self-assembly.[13]

The transient receptor potential vanilloid-3 channel TRPV3 (UniProt ID: Q8NET8) was correctly predicted to be tetrameric. In addition, we noticed the generally high ipTM scores and low ipTM range of structures predicted in the trimeric and pentameric states, which both had ipTM scores only slightly lower than the tetramers. The predicted trimeric structures resembled the tetramers with one missing subunit (Fig 5B), highlighting a case where visual examination is necessary to understand whether a predicted oligomeric state is plausible. Interestingly, however, the pentameric structures resembled the cryo-EM structure of TRPV3 recently elucidated (after the AF2.3 training date cutoff) as a pentamer, thought to be a transient state of the protein with a dilated pore.[14]

Similarly, we observed that the mechanosensitive channel of large conductance MscL, whose structure has been determined in both the tetrameric and pentameric states, had similar ipTM score distributions for these two states. However, we note in this case that both structures were determined prior to the cut-off date for training AF2. Finally, structural predictions of Pannexin 2 suggest that the hexameric and heptameric oligomerisation states are both probable, as they have similar high ipTM scores. Our results are consistent with a recent report that the closely related pannexin 1 channels are active in both the hexameric and heptameric states.[15] This result was surprising given that all structures of pannexins thus far have been resolved as heptamers.[16–20]

Interestingly, the ipTM score distributions and structural predictions for HpUreI (a proton-gated urea channel in *H. pylori*), PfFNT (a lactate/proton transporter in *P. falciparum*) and HsGlyR (the human glycine receptor) suggest that these proteins may also oligomerise with a different number of subunits in addition to their currently known annotated states.

Given that these data suggest that AF2-M may be able to identify cases where there are multiple biologically relevant oligomeric states, we selected 7 additional membrane proteins with multiple annotated oligomeric states from the UniProt Database and carried out structural predictions (Fig. 5C). All but one of these proteins were annotated as forming both monomers and at least one higher order oligomer. For a protein (UniProt ID: Q9HBJ8) annotated to be both monomeric and dimeric, the low ipTM scores across multiple oligomeric state predictions led to its assignment as a monomer. However, it can be clearly seen that the dimeric form of the protein has the highest ipTM score compared to other oligomeric states. Another protein (UniProt ID: Q5TGU0) was annotated as both dimeric and tetrameric, which was well reflected in the ipTM score distributions, which predict the tetramer with a slightly higher mean ipTM score than the dimer. Mixed results were observed in the remaining cases, where 5/6 proteins (all but UniProt ID: Q8TCT7) were predicted to be in one of the annotated oligomeric states.

Discussion

Many proteins undergo oligomerisation to form their biologically relevant states. Predicting the stoichiometry and the structures they adopt is important for understanding their functional roles and interactions. Here, we show that both AlphaFold2-Multimer and AlphaFold3 can be used to accurately predict the oligomeric states of proteins and generate multimeric structures consistent with experimental techniques. The average pLDDT of the monomeric structure present in the AFDB is an indicator of the accuracy of oligomeric prediction while the presence of a structure within the AlphaFold2 training set does not impact the accuracy of the predictions for our test set.

We optimise the computational efficiency of our prediction method, showing that 3 structures predicted with 10 recycles using AF2-M and the default prediction parameters available on the AF3 online server are sufficient to derive correct oligomeric states for 35/40 and 33/40 of the proteins in our test set respectively.

The small number of proteins in our test set and differences in parameters used means that it is not possible to say here whether AF2-M or AF3 is more likely to predict the correct oligomeric state. However, it appears that within this set of proteins, proteins with generally high pLDDT scores and without abnormally long helices in the AFDB structure have their oligomeric states correctly predicted by AF2-M. It is less clear what determines whether a protein will be predicted in the correct oligomeric state by AF3, although it seems to perform less well for proteins without an available structure in the PDB and has some trouble distinguishing between monomeric and non-monomeric proteins. We note that in 30/31 cases, if both AF2-M and AF3 predict the same oligomeric state for a given protein, then this is the annotated state, suggesting that use of AF2-M and AF3 concurrently could increase confidence in the oligomeric state prediction.

The ipTM metric used to identify the most likely oligomeric state is reported by default for each structure in both AF2-M and AF3 and performs better than pLDDT values which are not expected to be highly influenced by the interchain contacts. However, it has been noted that the ipTM scores can be impacted by the presence of disordered protein regions.[21–23] The effect of these regions can be reduced by removing them if they are known not to be involved in the interaction, restricting calculations to only residues near the interaction site,[24] or to interchain residue pairs that have well predicted aligned error distances.[23] The accuracy of using ipTM to predict oligomerisation states may therefore be improved by using one of these approaches.

For several membrane proteins which can be physiologically found in different oligomeric states (TRPV3, MscL and Pannexins), this information can be captured in the ipTM distributions. This suggests the value of this approach for identifying proteins where multiple physiologically relevant oligomerization states may be plausible and warrant further investigation.

For two membrane proteins we examined (e.g. HsXPR1, NavAb), their oligomerisation states determined via structural biology techniques were not annotated within the UniProt Database. Additionally, for PANX1, which has recently been found to be functional as hexamers and heptamers, only the heptameric state is annotated, suggesting that the interactions sections on the UniProt Database may sometimes provide an incomplete picture of protein oligomerisation states.

The high computational cost of modelling four or more copies of larger proteins may prohibit this technique from being easily usable for proteins where one subunit contains more than 1000 amino acids. Additionally, due to computational limitations, we did not consider hexameric and higher order assemblies in our initial set of 40 proteins. However, this technique of oligomeric state prediction using the ipTM score is likely to be generalisable to higher order assemblies and we show that it is capable of correctly predicting several hexameric and heptameric assemblies in our curated membrane protein set. Additionally, it is possible that this technique could provide insight into stoichiometric ratios for heteromeric complexes in instances where the identity of interacting proteins is known but the number of each subunit is not.

One benefit of using AF to predict the oligomeric state is that it also yields structures of the protein in the predicted state which could provide functional insights. To increase confidence in these predictions, plausible structures could be combined with short, physics-based molecular dynamics simulations to ascertain their stability in solution or membrane environments in combination with experimental validation.

Ultimately, AF2-M and AF3 are useful computational tools for predicting yet unknown oligomeric states and structures of proteins.

Methods

Selection of oligomeric protein sequences

The DeepSub database (https://github.com/tibbdc/DeepSub/blob/main/DATA/Dataset_0724_new.csv) containing proteins from the UniProt database annotated with experimentally determined oligomeric

states was used to randomly select 40 human proteins (5 each of known monomeric, homodimeric, homotrimeric, homotetrameric and homopentameric proteins), detailed in Table S1.[6]

Structural prediction using AlphaFold2 Multimer

Selected sequences were extracted from the UniProt Database and used to generate multiple sequence alignments for structure prediction by ColabFold implementing AlphaFold-Multimer v.2.3.[25,26] 20 structures containing 2,3,4,5 copies of each protein were predicted with 3 recycles without subsequent relaxation nor modification to the multiple sequence alignment. To optimize computational efficiency, structural predictions were also carried out using all permutations of 1,5,10,20 structures and 0,1,2,3 recycles. For the 5 proteins with incorrect oligomeric state predictions, we carried out additional structural predictions using 5 and 10 recycles (generating 20 structures in each case). For two large proteins within our dataset (UniProt IDs: P35670, P533396), we were unable to predict their structures as tetramers/pentamers due to computational restrictions.

Structural prediction using AlphaFold3

The AlphaFold3 online server was used to predict multimeric complexes (2,3,4,5 copies) of each protein, using sequences extracted from the UniProt Database.[11] By default, 10 recycles were performed and 5 structures generated. As above, for two large proteins (UniProt IDs: P35670, P533396), we were unable to predict their structures as tetramers/pentamers due to computational restrictions.

Analysis of structures

Structures and output files produced by AlphaFold2-Multimer and AlphaFold3 were analysed using python scripts written using the pandas, scipy, matplotlib, MDAnalysis and MDTraj libraries.[27–31] Structures were visualised and figures produced using Visual Molecular Dynamics (VMD).[32]

For each protein, the predicted monomeric structure from the AFDB and the top ranked structure produced for the correct oligomeric state were oriented along the Z principal axis using the GMX editconf princ function within GROMACS (v.2023).[33] For oligomeric structures predicted using AlphaFold2-Multimer, the top ranked structure was chosen from those generated from 3 recycles and 20 structures. The α-carbon atoms of the monomeric structure and the first chain of the predicted oligomeric structure were aligned within VMD to generate the structural images.

The average monomeric pLDDT for each protein was calculated using each associated structure in the AFDB by averaging the per-residue pLDDT for all residues in each structure.

The secondary structure was calculated using the DSSP algorithm implemented in MDTraj.[31,34] For each structure, the length of the longest helix was assigned to be the longest continuous stretch of residues calculated to have a helical ('H') secondary structure.

The template modelling (TM) score for the structures of each protein predicted in the correct oligomeric state were calculated following sequence and structural alignment to the predicted monomeric structure in the AFDB and, where available, to valid experimentally resolved structures using US-align.[35] All experimentally resolved structures associated with the UniProt ID of each structure were assessed for validity. Monomeric structures were removed from comparison to our oligomeric predictions. Structures with < 95% sequence similarity between the UniProt sequence and PDB sequence were discarded, as were structures containing only fragments (< 50% of the UniProt sequence) of the protein. Using options -mm 1 and -ter 0, US-Align was used to calculate the TM score after alignment of the two multi-chain oligomeric structures. As some proteins of interest were resolved with accessory proteins and additional chains were present in some PDBs, structures were manually inspected to identify the correct chains for comparison where necessary. For the 5 proteins with a large number (> 10) of eligible PDBs (Table S3), if a PDB with the

correct number of chains yielded a TM score above 0.95, we did not calculate further TM scores for PDBs resolved with additional chains. TM scores to the monomeric structure were also calculated using US-Align, normalizing by the length of the monomeric protein.

References

1. Levy, E. D. & Teichmann, S. A. Chapter Two - Structural, Evolutionary, and Assembly Principles of Protein Oligomerization. in *Progress in Molecular Biology and Translational Science* (eds. Giraldo, J. & Ciruela, F.) vol. 117 25–51 (Academic Press, 2013).
2. Schweke, H. *et al.* An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999-1010.e15 (2024).
3. Chetri, P. B., Khan, H. & Tripathi, T. Chapter 5 - Methods to determine the oligomeric structure of proteins. in *Advances in Protein Molecular and Structural Biology Methods* (eds. Tripathi, T. & Dubey, V. K.) 49–76 (Academic Press, 2022). doi:10.1016/B978-0-323-90264-9.00005-2.
4. Baek, M., Park, T., Heo, L., Park, C. & Seok, C. GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic Acids Research* **45**, W320–W324 (2017).
5. Avraham, O., Tsaban, T., Ben-Aharon, Z., Tsaban, L. & Schueler-Furman, O. Protein language models can capture protein quaternary state. *BMC Bioinformatics* **24**, 433 (2023).
6. Deng, R. *et al.* DeepSub: Utilizing Deep Learning for Predicting the Number of Subunits in Homo-Oligomeric Protein Complexes. *International Journal of Molecular Sciences* **25**, 4803 (2024).
7. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research* **52**, D368–D375 (2024).
8. Homma, F., Huang, J. & van der Hoorn, R. A. L. AlphaFold-Multimer predicts cross-kingdom interactions at the plant-pathogen interface. *Nat Commun* **14**, 6040 (2023).
9. Humphreys, I. R. *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805.
10. Omidi, A., Møller, M. H., Malhis, N., Bui, J. M. & Gsponer, J. AlphaFold-Multimer accurately captures interactions and dynamics of intrinsically disordered protein regions. *Proceedings of the National Academy of Sciences* **121**, e2406407121 (2024).
11. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* (2024) doi:10.1038/s41586-024-07487-w.
12. Jayaraman, K. *et al.* SLC6 transporter oligomerization. *Journal of Neurochemistry* **157**, 919–929 (2021).
13. Periole, X., Zeppelin, T. & Schiøtt, B. Dimer Interface of the Human Serotonin Transporter and Effect of the Membrane Composition. *Scientific Reports* **8**, 5080 (2018).
14. Lansky, S. *et al.* A pentameric TRPV3 channel with a dilated pore. *Nature* **621**, 206–214 (2023).
15. Gupta, S. *et al.* Distinct properties and activation of hexameric and heptameric Pannexin 1 channel concatemers. *Journal of General Physiology* **157**, e202413676 (2024).
16. Jin, Q. *et al.* Cryo-EM structures of human pannexin 1 channel. *Cell Research* **30**, 449–451 (2020).
17. Mou, L. *et al.* Structural basis for gating mechanism of Pannexin 1 channel. *Cell Research* **30**, 452–454 (2020).
18. Qu, R. *et al.* Cryo-EM structure of human heptameric Pannexin 1 channel. *Cell Research* **30**, 446–448 (2020).
19. Deng, Z. *et al.* Cryo-EM structures of the ATP release channel pannexin 1. *Nature Structural & Molecular Biology* **27**, 373–381 (2020).
20. Ruan, Z., Orozco, I. J., Du, J. & Lü, W. Structures of human pannexin 1 reveal ion pathways and mechanism of gating. *Nature* **584**, 646–651 (2020).
21. Bret, H., Gao, J., Zea, D. J., Andreani, J. & Guerois, R. From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2. *Nature Communications* **15**, 597 (2024).
22. Martin, J. AlphaFold2 Predicts Whether Proteins Interact Amidst Confounding Structural Compatibility. *J. Chem. Inf. Model.* **64**, 1473–1480 (2024).
23. Dunbrack, R. L. Rēs ipSAE loquunt: What's wrong with AlphaFold's ipTM score and how to fix it. *bioRxiv* 2025.02.10.637595 (2025) doi:10.1101/2025.02.10.637595.
24. Varga, J. K., Ovchinnikov, S. & Schueler-Furman, O. actifpTM: a refined confidence metric of AlphaFold2 predictions involving flexible regions. Preprint at https://doi.org/10.48550/arXiv.2412.15970 (2024).
25. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679–682 (2022).
26. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
27. McKinney, W. Data Structures for Statistical Computing in Python. in 56–61 (Austin, Texas, 2010). doi:10.25080/Majora-92bf1922-00a.
28. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).
29. J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007).
30. Gowers, R. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. in 98–105 (Austin, Texas, 2016). doi:10.25080/Majora-629e541a-00e.
31. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109**, 1528–1532 (2015).
32. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).
33. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
34. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
35. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods* **19**, 1109–1115 (2022).

Supporting Information

Table S1 – Protein Set

UniProt ID	Annotated Oligomeric State	PDB Available?	PDB in Training Set?	Monomer pLDDT	Correct using AF2.3?	Correct using AF3?	Membrane Localisation	Membrane Protein Topology
Q13087	1	✗	✗	87.53	✓	✗	✗	N/A
Q9NZ08	1	✓	✓	92.76	✓	✗	✓	Single-pass
P30740	1	✓	✓	92.93	✓	✓	✗	N/A
P27144	1	✓	✓	90.22	✓	✓	✗	N/A
Q68CK6	1	✗	✗	91.42	✓	✓	✗	N/A
Q16762	1	✓	✗	96.43	✓	✓	✗	N/A
P35670	1	✓	✓	71.83	✓	✗	✓	Multi-pass
Q68D85	1	✓	✓	78.19	✓	✓	✓	Single-pass
Q96QG7	2	✗	✗	94.9	✓	✓	✓	N/A
P06727	2	✓	✓	80.45	✗	✓	✗	N/A
O94804	2	✓	✓	74.81	✗	✓	✓	N/A
Q9P0L0	2	✓	✓	79.7	✓	✓	✓	Single-pass
O43776	2	✓	✓	94.06	✓	✓	✗	N/A
P20711	2	✓	✓	96.94	✓	✓	✗	N/A
Q9GZP0	2	✗	✗	74.79	✓	✗	✗	N/A
Q2M3C6	2	✗	✗	54.62	✗	✗	✓	Multi-pass
P32970	3	✓	✓	83.43	✓	✓	✓	Single-pass
P43005	3	✓	✓	80.33	✓	✓	✓	Multi-pass
Q8TCS8	3	✓	✓	87.53	✓	✓	✓	N/A
P21941	3	✗	✗	83.72	✓	✗	✗	N/A
P50591	3	✓	✓	82.4	✓	✓	✓	Single-pass
O14684	3	✓	✓	96.25	✓	✓	✓	Multi-pass
P43003	3	✓	✓	81.85	✓	✓	✓	Multi-pass
Q99735	3	✓	✓	93.94	✓	✓	✓	Multi-pass
Q9NPJ3	4	✓	✓	96.16	✓	✓	✗	N/A
O75496	4	✓	✓	70.4	✗	✓	✗	N/A
O60741	4	✓	✓	70.56	✓	✓	✓	Multi-pass
O43827	4	✗	✗	86.23	✗	✗	✗	N/A
Q8N539	4	✓	✓	80.24	✓	✓	✓	Single-pass
P02766	4	✓	✓	88.93	✓	✓	✗	N/A
Q9P0L9	4	✓	✓	76.04	✓	✓	✓	Multi-pass

P53396	4	✓	✓	92.6	✓	✓	✗	N/A
P23415	5	✓	✓	84.63	✓	✓	✓	Multi-pass
P02743	5	✓	✓	94.41	✓	✓	✗	N/A
Q68DU8	5	✓	✓	69.39	✓	✓	✓	N/A
Q9NXV2	5	✓	✓	78.28	✓	✓	✗	N/A
P02741	5	✓	✓	94.23	✓	✓	✗	N/A
O75311	5	✓	✓	84.36	✓	✓	✓	Multi-pass
Q693B1	5	✗	✗	84.28	✓	✓	✗	N/A
Q86SE8	5	✓	✓	73.32	✓	✓	✗	N/A

Table S2 – Comparison of oligomer predictions to experimentally resolved structures and AF2 monomer

UniProt	# PDB	# Eligible PDB (seq similarity & length)	Eligible PDBs	Max TM Score (AF2)	Most similar PDB (AF2)	TM Score to Monomer (AF2)	Max TM Score (AF3)	Most similar PDB (AF3)	TM Score to Monomer (AF3)
Q96QG7	0	0		-	-	0.965	-	-	0.957
P06727	1	1	3S84	0.375	3S84	0.249	0.954	3S84	0.566
O94804	14	0		-	-	0.596	-	-	0.637
Q9P0L0	2	2	2RR3, 6TQR	0.973	6TQR	0.558	0.974	6TQR	0.571
O43776	6	5	5XIX, 8H53, 8TC7, 8TC8, 8TC9	0.996	8TC9	0.916	0.995	8TC7	0.971
P20711	5	3	3RBF, 3RBL, 3RCH	0.869	3RCH	0.978	0.877	3RCH	0.995
Q9GZP0	0	0		-	-	0.390	-	-	0.422
Q2M3C6	0	0		-	-	0.316	-	-	0.323
P32970	1	1	7KX0	0.968	7KX0	0.715	0.994	7KX0	0.731
P43005	14	9	6S3Q, 6X2L, 6X3E, 6X3F, 7NSG, 8CTC, 8CUA, 8CUI, 8CV2	0.988	8CV2	0.884	0.998	6X2L	0.584
Q8TCS8	2	1	3U1K	0.989	3U1K	0.885	0.987	3U1K	0.885
P21941	0	0		-	-	0.399	-	-	0.422
P50591	7	4	1D0G, 1DU3, 4N90, 5CIR	0.969	3U1K	0.562	0.990	5CIR	0.583
O14684	18	1	3DWW	0.825	3DWW	0.978	0.823	3DWW	0.976
P43003	5	1	7NPW	0.776	7NPW	0.907	0.772	7NPW	0.868
Q99735	4	4	6SSR, 6SSS, 6SSU, 6SSW	0.989	6SSS	0.910	0.992	6SSS	0.916
Q9NPJ3	3	2	2F0X, 3F5O	0.996	2F0X	0.994	0.997	2F0X	0.993
O75496	6	0		-	-	0.382	-	-	0.334
O60741	11	11	5U6O, 5U6P, 6UQF, 6UQG, 8T4M, 8T4Y, 8T50, 8UC7, 8UC8, 9BC6, 9BC7	0.979	8T4Y	0.616	0.997	8T4M	0.655
O43827	0	0		-	-	0.632	-	-	0.753
Q8N539	8	0		-	-	0.507	-	-	0.496
P02766	412	41	1G1O, 1GKO, 1ICT, 1QAB, 1RLB, 1SOQ, 1TZ8, 2QEL, 2WQA, 3BSZ, 3DGD, 3DID, 3DO4, 3GPS, 3GRB, 3GRG, 5EZP, 5LLV, 5TTR, 6SDZ, 7OB4, 7Y1I, 8ADE, 8E7D, 8E7E, 8E7H, 8E7I, 8E7J, 8G9R, 8GBR, 8I0O, 8PKE, 8PKF, 8PKG, 8TDN, 8TDO, 8VE0, 8VE1, 8VE2, 8VE3, 8VE4	0.992	8VE1	0.821	0.996	8VE1	0.818
Q9P0L9	3	1	6DU8	0.981	6DU8	0.759	0.992	6DU8	0.625
P53396	35	20	6HXH, 6O0H, 6POE, 6POF, 6QFB, 6UI9, 6UIA, 6UUW, 6UUZ, 6UV5, 7LIW, 7LJ9, 7LLA, 7RIG, 7RKZ, 7RMP, 8G1E, 8G1F, 8G5C, 8G5D	0.974	8G5C	0.915	0.999	6UUW	0.961

P23415	9	3	8DN2, 8DN4, 8DN5	0.990	8DN4	0.834	0.992	8DN4	0.830
P02743	13	13	1GYK, 1LGN, 1SAC, 2A3W, 2A3X, 2A3Y, 2W08, 3D5O, 3KQR, 4AVS, 4AVT, 4AVV, 4AYU	0.997	2W08	0.919	0.999	4AVV	0.934
Q68DU8	7	0		-	-	0.320	-	-	0.299
Q9NXV2	9	2	3DRX, 3DRY	0.914	3DRX	0.719	0.942	3DRX	0.720
P02741	16	16	1B09, 1GNH, 1LJ7, 3L2Y, 3PVN, 3PVO, 7PK9, 7PKB, 7PKD, 7PKE, 7PKF, 7PKG, 7PKH, 7TBA, 8WV4, 8WV5	0.997	7PKB	0.945	0.999	7PKE	0.945
O75311	5	5	5CFB, 5TIN, 5TIO, 5VDH, 5VDI	0.991	5TIO	0.815	0.996	5TIO	0.799
Q693B1	0	0		-	-	0.683	-	-	0.663
Q86SE8	1	1	3T30	0.995	3T30	0.570	0.992	3T30	0.603

Table S3 – Membrane Protein Set

UniProt ID	Annotated Oligomeric State	PDB Available?	PDB in Training Set?	Monomer pLDDT	Correct using AF2.3?	Membrane Localisation?	Membrane Protein Topology
O94886	1	✓	✗	75.3	✗	✓	Multi-pass
P31645	1	✓	✓	85.43	✗	✓	Multi-pass
Q9UBH6	2	✓	✓	83.73	✓	✓	Multi-pass
Q9XEA1	2	✓	✓	84.81	✓	✓	Multi-pass
P43004	3	✓	✗	77.74	✓	✓	Multi-pass
Q1XA76	3	✓	✓	83.42	✓	✓	Multi-pass
A8EVM5	4	✓	✓	93.73	✓	✓	Multi-pass
Q8NET8	4	✓	✓	76.93	✓	✓	Multi-pass
O77389	5	✓	✓	91.42	✓	✓	Multi-pass
P0A742	5	✓	✓	85.22	✓	✓	Multi-pass
P23416	5	✓	✓	84.44	✓	✓	Multi-pass
Q09068	6	✗	✗	90.27	✓	✓	Multi-pass
P0C0S1	7	✓	✓	92.62	✓	✓	Multi-pass
Q96RD6	7	✓	✗	58.15	✓	✓	Multi-pass
P04001	1,2,4	✗	✗	82.86	✓	✓	Multi-pass
P36021	1,2, X*	✗	✗	80.05	✓	✓	Multi-pass
Q8TCT7	1,2	✗	✗	76.42	✗	✓	Multi-pass
Q9HBJ8	1,2	✗	✗	77.39	✓	✓	Single-pass
Q92911	1,2, X*	✗	✗	81.58	✓	✓	Multi-pass
Q9UIJ5	1,2	✗	✗	83.08	✓	✓	Multi-pass
Q5TGU0	2,4	✗	✗	66.84	✓	✓	Multi-pass

* X denotes annotation as a 'homooligomer' in the UniProt Database

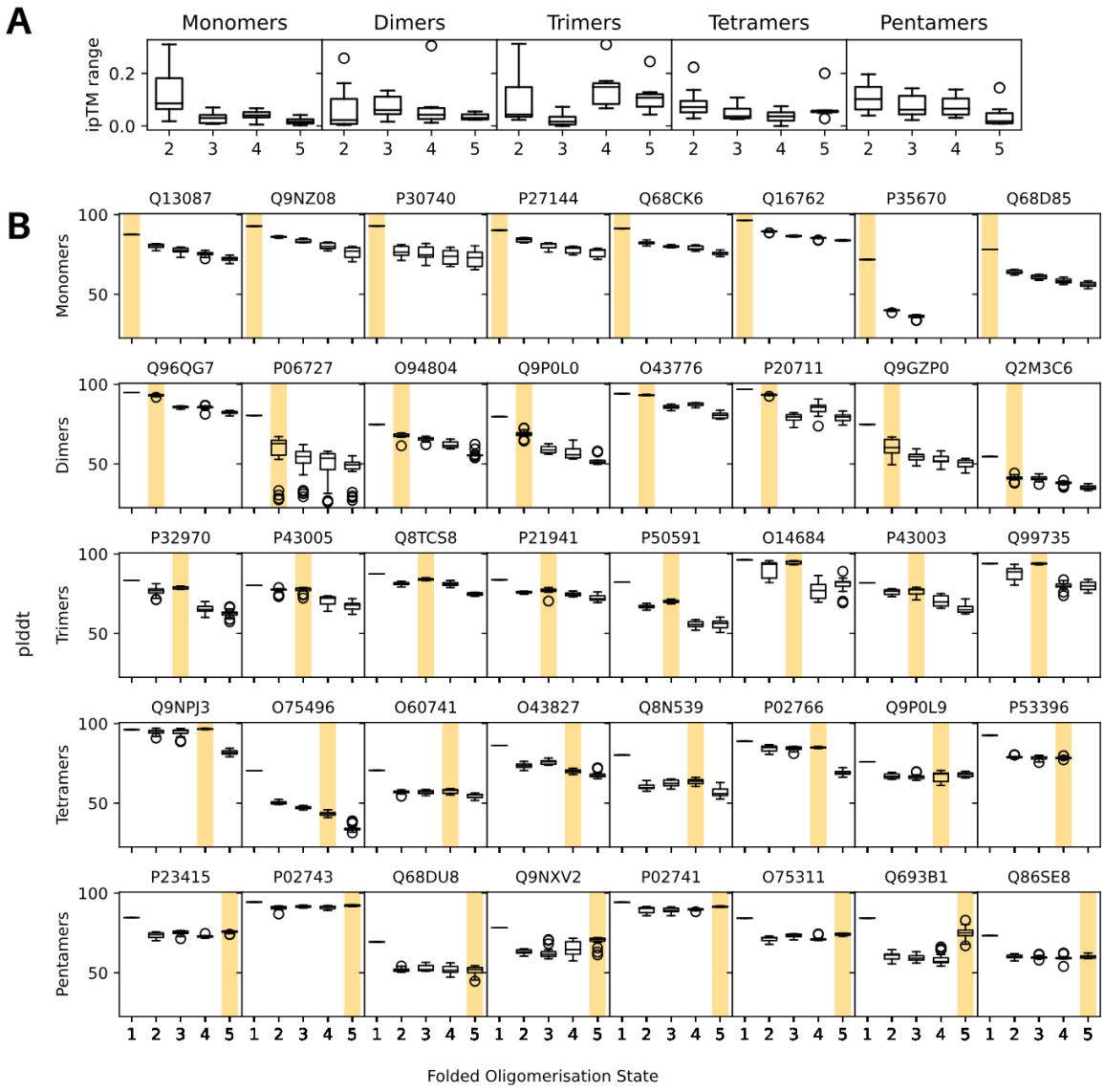

Figure S1 – ipTM range and pLDDT plots for AF2-M predictions

(a) ipTM range distributions for monomeric, dimeric, trimeric, tetrameric and pentameric proteins in different folded oligomeric states

(b) Box plots showing pLDDT score distributions for each protein in different folded oligomeric states across the 40 test set proteins. The annotated oligomeric state highlighted in yellow.

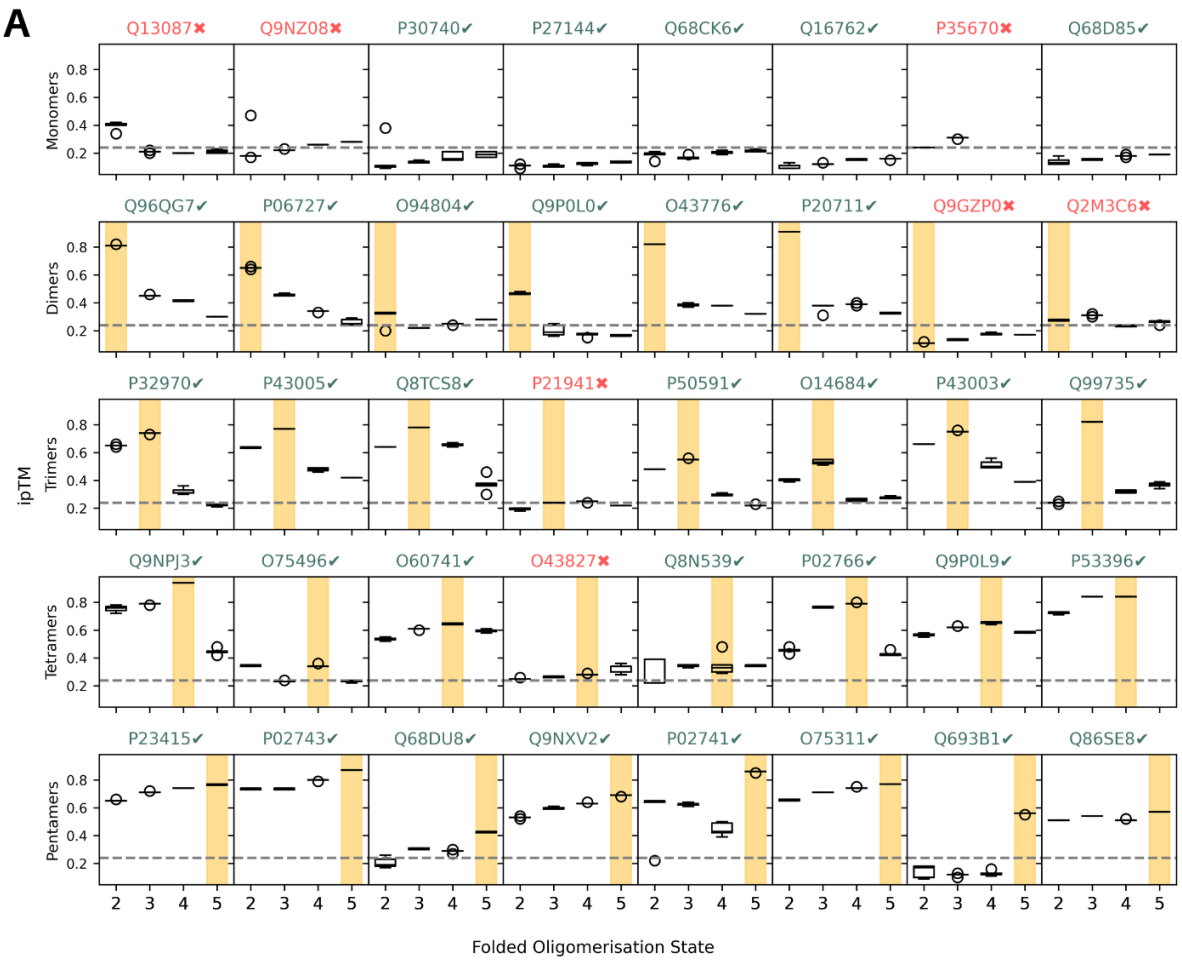

Figure S2 – ipTM distribution plots for AF3 predictions

(a) Box plots showing ipTM score distributions for each protein in different folded oligomeric states across the 40 test set proteins. The annotated oligomeric state highlighted in yellow. UniProt IDs are noted in green for correct predictions and red for incorrect ones.

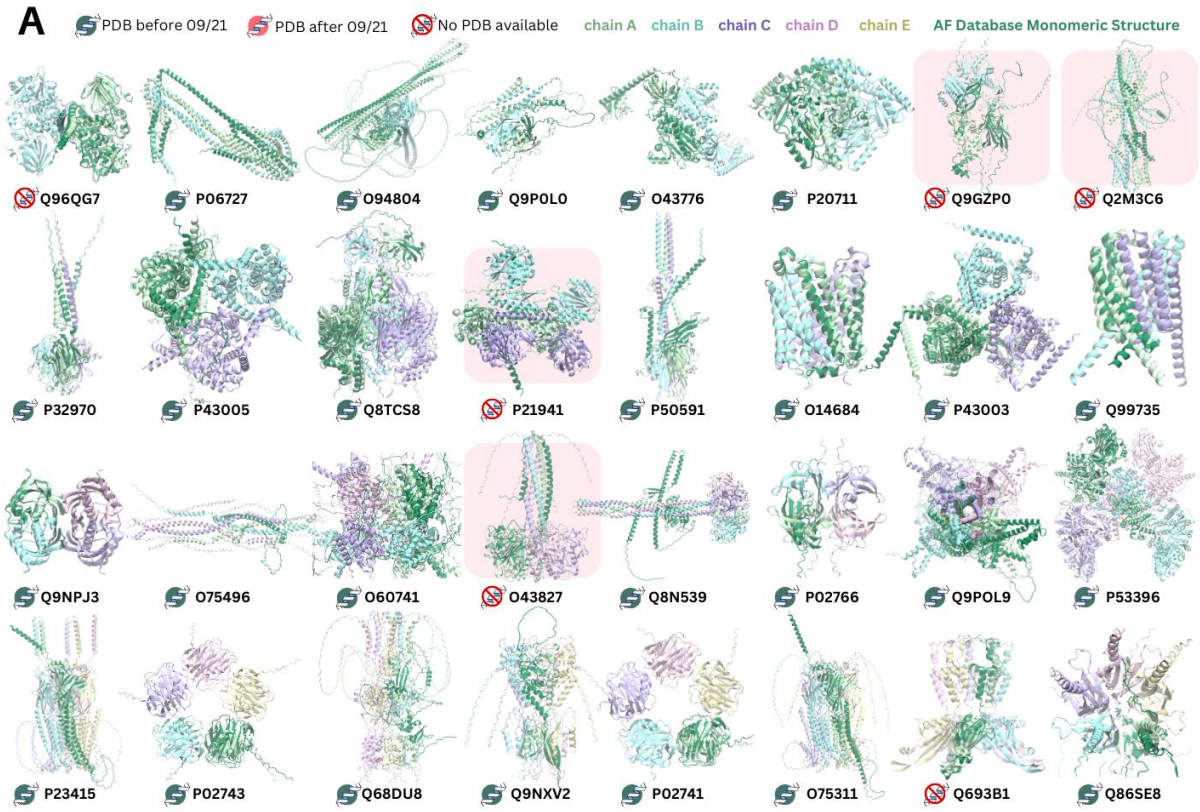

Figure S3 – structural images of AF3 predictions

(a) Snapshots of the top ranked structural predictions by AF3 in the correct oligomeric state for each multimeric protein (ribbon, coloured by chain), aligned with the AFDB prediction of the protein as a monomer (ribbon, dark green). The presence of an experimentally resolved structure in the PDB or AlphaFold3 training set is indicated with an icon near each UniProt ID. The structures of proteins predicted in an incorrect oligomeric state are highlighted in red.