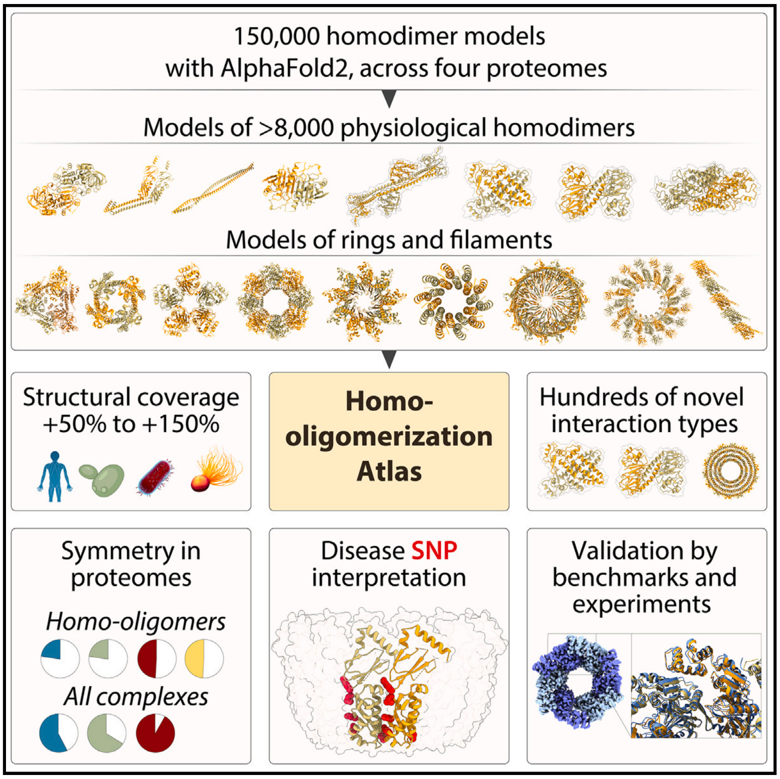
An atlas of protein homo-oligomerization across domains of life

Cell

Resource

An atlas of protein homo-oligomerization across domains of life

Graphical abstract


150,000 homodimer models
with AlphaFold2, across four proteomes
Models of >8,000 physiological homodimers
Models of rings and filaments
Structural coverage
+50% to +150%
Homo-oligomerization Atlas
Hundreds of novel interaction types
Symmetry in proteomes
Homo-oligomers
All complexes
Disease SNP interpretation
Validation by benchmarks and experiments

Authors

Hugo Schweke, Martin Pacesa,
Tal Levin, ..., Bruno E. Correia,
Sucharita Dey, Emmanuel D. Levy

Correspondence

d.n.woolfson@bristol.ac.uk (D.N.W.),
bruno.correia@epfl.ch (B.E.C.),
sdey@iitj.ac.in (S.D.),
emmanuel.levy@gmail.com (E.D.L.)

In brief

A strategy for predicting homo-oligomers yields structural models for protein complexes, offering insights into quaternary structure organization across proteomes.

Highlights

- Models of >8,000 homomer structures across four proteomes, including rings and filaments
- The models increase the structural coverage of homomers by 50%–150% per proteome
- Hundreds of interface types were identified, with three being experimentally validated
- The models form a basis for proteome-wide structuromics analyses

Schweke et al., 2024, Cell *187*, 999–1010
February 15, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2024.01.022

CellPress

Cell

CellPress
OPEN ACCESS

Resource

An atlas of protein homo-oligomerization across domains of life

Hugo Schweke,[1] Martin Pacesa,[2] Tal Levin,[1] Casper A. Goverde,[2] Prasun Kumar,[3,4,5,6] Yoan Duhoo,[7] Lars J. Dornfeld,[2] Benjamin Dubreuil,[1] Sandrine Georgeon,[2] Sergey Ovchinnikov,[8] Derek N. Woolfson,[3,4,5,6,*] Bruno E. Correia,[2,*] Sucharita Dey,[9,*] and Emmanuel D. Levy[1,10,*]

[1]Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot, Israel
[2]Laboratory of Protein Design and Immunoengineering, École Polytechnique Fédérale de Lausanne and Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3]School of Chemistry, University of Bristol, Bristol BS8 1TS, UK
[4]School of Biochemistry, University of Bristol, Bristol BS8 1TD, UK
[5]Bristol BioDesign Institute, University of Bristol, Life Sciences Building, Bristol BS8 1TQ, UK
[6]Max Planck-Bristol Centre for Minimal Biology, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK
[7]Protein Production and Structure Characterization Core Facility (PTPSP), School of Life Sciences, École polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[8]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA, USA
[9]Department of Bioscience and Bioengineering, Indian Institute of Technology Jodhpur, Rajasthan, India
[10]Lead contact
*Correspondence: d.n.woolfson@bristol.ac.uk (D.N.W.), bruno.correia@epfl.ch (B.E.C.), sdey@iitj.ac.in (S.D.), emmanuel.levy@gmail.com (E.D.L.)

https://doi.org/10.1016/j.cell.2024.01.022

SUMMARY

Protein structures are essential to understanding cellular processes in molecular detail. While advances in artificial intelligence revealed the tertiary structure of proteins at scale, their quaternary structure remains mostly unknown. We devise a scalable strategy based on AlphaFold2 to predict homo-oligomeric assemblies across four proteomes spanning the tree of life. Our results suggest that approximately 45% of an archaeal proteome and a bacterial proteome and 20% of two eukaryotic proteomes form homomers. Our predictions accurately capture protein homo-oligomerization, recapitulate megadalton complexes, and unveil hundreds of homo-oligomer types, including three confirmed experimentally by structure determination. Integrating these datasets with omics information suggests that a majority of known protein complexes are symmetric. Finally, these datasets provide a structural context for interpreting disease mutations and reveal coiled-coil regions as major enablers of quaternary structure evolution in human. Our strategy is applicable to any organism and provides a comprehensive view of homo-oligomerization in proteomes.

INTRODUCTION

The organization of proteins into complexes and biomolecular networks underlies cellular processes and functions. At the most fundamental level, protein assembly occurs by homo-oligomerization, whereby identical copies of a protein interact symmetrically to form higher-order structures.[1,2] These so-called homomers possess unique structural and functional properties[1] (Figure 1A). They enable the formation of repetitive structural elements, as in the cytoskeleton, and can create shapes like rings, barrels, or cages.[3] More broadly, the repetition of protein chains in homomers provides multivalence, a parameter critical to protein binding, notably in the formation of biomolecular condensates.[4] Functionally, the conformation of their subunits can be coupled to mediate allosteric transitions, and their formation can be modulated by environmental cues, such as pH or post-translational modifications.[5] As such, comprehensive knowledge of homomer structures provides a foundational layer of information to analyze and interpret protein structure and function. In particular, it would allow modeling and predicting the underlying molecular basis of a wide variety of human diseases and associated mutations that occur at, or close to, interfaces. For example, aquaporin forms ring-like channels, allowing water to flow through membranes, and mutations impairing the ring assembly are associated with nephrogenic diabetes insipidus disease.[6] Beyond their functional importance, homomers are shaping the evolution of protein complexes and networks; they represent the ancestral state of myriad key macromolecular complexes such as histones, proteasomes, or chaperones, which diversified through gene duplication.[7]

The central role of homomers in biology motivates their comprehensive characterization. Recent advances in machine learning have revolutionized the accuracy with which protein tertiary structure is predicted.[8,9] These advances have been scaled

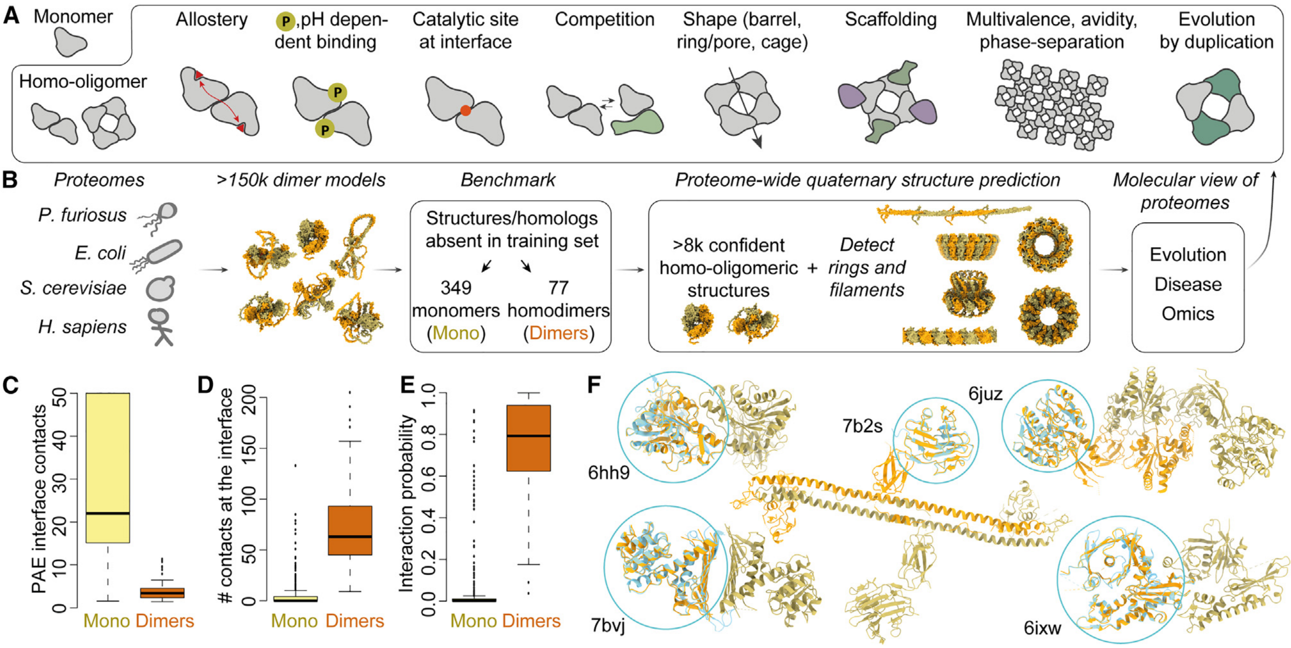


Figure 1. AlphaFold2 predicts the structure of homodimers with high accuracy

(A) Homo-oligomeric proteins possess unique functional, shape, and evolutionary properties.

(B) Overview of the data flow in this work. Dimer structures were predicted for proteins across four species' proteomes, yielding 156,065 models. The models were subsequently scored based on a benchmark, yielding over 8,000 high-confidence dimer structures. The dimer structures were used to predict higher-order biologically relevant macromolecular assemblies into rings and filaments, yielding proteome-wide homo-oligomerization information.

(C) Average Predicted Aligned Error (PAE) of contacting residues for monomer (yellow) and dimers (orange) from the benchmark dataset.

(D) The number of residue-residue contacts between subunits also discriminates dimers from monomers.

(E) The average PAE of interface contacts and the number of contacts at the interface (C and D) were used together to fit the benchmark by logistic regression, yielding an interaction probability (STAR Methods). Benchmark data are detailed in Table S1.

(F) Examples of discrepancies where the experimental structure is a monomer (blue and circled, PDB code indicated), while our predictions suggest a dimer (orange/green).

See also Figure S1.

up, making the structure of monomeric proteins available across entire proteomes.[10,11] Machine learning approaches can also be employed to predict the structure of protein complexes[12,13] and serve to predict heteromeric complexes in yeast[14] and human.[15] However, two major challenges make it difficult to predict homo-oligomers systematically on a proteome-wide scale. While AlphaFold2 has been the method of choice, it requires knowledge of the number of protein copies present in a complex, and this number is typically unknown. In addition, computation and memory requirements scale exponentially with the number of copies modeled by AlphaFold2, making it difficult to predict large complexes at scale.

Here, we addressed these challenges to predict the structure of homo-oligomers on a proteome scale. We systematically generated structures for putative homodimers and analyzed them to identify those with physiological relevance. The latter were subsequently processed independently of AlphaFold2 to predict higher-order structures, including rings and filaments. We computed homo-oligomeric structures for four species: *Pyrococcus furiosus*, Escherichia *coli*, Saccharomyces *cerevisiae*, and *Homo sapiens*. The resulting datasets comprise 872, 2,181, 1,196, and 3,946 homo-oligomers, covering 20%–45% of the analyzed proteomes. This emphasizes that a considerable fraction of the proteome undergoes homo-oligomerization, highlighting once more the importance of this phenomenon for understanding protein structure, function, and evolution. A number of these models recapitulate large structures including a hexameric ring that we validated experimentally by cryoelectron microscopy (cryo-EM) or a megadalton macrophage pore-forming complex, which consists of a ring with 16 protein copies.[16] These datasets add a quaternary structure dimension to proteomes and will bolster our molecular understanding of their function and evolution (Figure 1B). We illustrate such biological insights in three analyses showing that (1) coiled-coil regions are major enablers of quaternary structure evolution in the human proteome, (2) interaction interfaces in homo-oligomers across the human proteome are 70% more likely to contain disease mutations than protein surfaces, and (3) strikingly large fractions of homo- and hetero-oligomeric protein complexes in prokaryotes and eukaryotes appear to be symmetric.

RESULTS

Predicting homodimers with AlphaFold2

We first assessed the accuracy of AlphaFold2 at identifying homodimers and correctly predicting their structure. We used the initial AlphaFold2 weights rather than the "multimer" weights because the gain in accuracy for homo-oligomers appeared limited.[12] In addition, those weights were trained on single chains, thus avoiding overfitting when predicting multi-chain interactions in homo-oligomers. We compiled a non-redundant dataset from the PDB,[17] consisting of 349 monomers and 77 homodimers with a structure deposited after May 2018 (Figure 1B; STAR Methods). These structures were therefore absent from the Alphafold2 training set. Predictions of multiple metrics showed an excellent agreement with the X-ray crystallography-derived dataset of monomers and dimers (Figure S1; Table S1). Two metrics were particularly informative in discriminating physiologically relevant homodimers from monomers: the first is the average Predicted Aligned Error (PAE) of amino acids in contacts (Figure 1C), and the second is the number of contacts between amino acids (Figure 1D). We combined both metrics in a logistic regression model, which predicts the probability of an AlphaFold2 dimer to be physiologically relevant (Figure 1E). This simple two-parameter model accurately captured the oligomeric state of experimental crystal structures, with an area under the receiver operator curve (AUC) of 0.978. Manually inspecting cases where the predictions differed from the experimental structure revealed cases where the AlphaFold2 model appeared physiologically relevant despite contradicting experimental data (Figure 1F). In the case of a two-domain esterase (PDB: 6HH9[18]), the predicted dimer was, in fact, observed experimentally and existed in the crystal lattice, hinting at its possible existence in solution. In several instances (e.g., PDB: 7B2S and 6JUZ[19]), the structure solved by X-ray crystallography was truncated and did not include the dimerization domain, which explained the apparent inconsistency. In another example (PDB: 7BVJ[20]), the primary reference provided evidence for the formation of a homodimer. In the last example, an actin-like protein appeared in the PDB as a monomer due to point mutations, but the interface driving filament assembly was still detected by AlphaFold2 despite these mutations.

Overall, this analysis shows that AlphaFold2 accurately predicts the structure of homodimers and that we can efficiently discriminate between physiological homodimers and artifactual complexes, which is consistent with a recent report.[21] These results motivate the generalization of its use to discover homo-oligomers across proteomes.

Proteome-wide discovery of homodimers

Protein structure inference is computationally expensive and hardly applicable to large complexes on a proteome-wide scale. To address this limitation, we adopted a hierarchical approach where we initially predicted homodimers and subsequently analyzed whether they form larger structures based on the dimer's internal symmetry. We generated a total of 156,065 homodimer models altogether covering 99.8%, 98.2%, 94.7%, and 89.7% of reference proteomes[22] for *P. furiosus*, *E. coli*, *S. cerevisiae*, and *H. sapiens*, respectively. The incomplete coverage was mostly due to proteins exceeding 1,200 residues (2,400 in the dimer) because their prediction required excessive resources (STAR Methods).

We analyzed the inter-subunit contacts of these models, their consistency across the five AlphaFold2 networks, and their PAE statistics. We then used these metrics (Figures 1 and S1) to calculate confidence probabilities based on the benchmark set. The scoring process (STAR Methods) yielded 872, 2,181, 1,196, and 3,946 homodimers that covered 43%, 44%, 21%, and 21% of the four proteomes, respectively (Figure 2A; detailed information about reference sets is provided in Table S2). A significant fraction of these predictions closely matched sequences with an experimentally solved structure. Because we employed a version of AlphaFold2 trained on single chains, we evaluated whether these models recapitulated known homo-oligomeric structures. This comparison revealed an excellent agreement, with 95.3%, 97.7%, 98.9%, and 98.7% of models recapitulating the known interaction interface in *P. furiosus*, *E. coli*, *S. cerevisiae*, and *H. sapiens*, respectively (Figures 2C and S2). We did not find structural homologs with similar subunit interaction geometry for 15%–20% of the models (STAR Methods), which thereby represent hundreds of potentially new quaternary structure types (Figure 2B). While the pace of new protein fold discovery is relatively slow, likely due to the extensive coverage of existing structures, the number of quaternary structure types that we discovered indicates that they cover a vast structural landscape, much larger than that of tertiary structures (Figure 2B).

Next, we focused on obligate homo-oligomers, which are expected to be unstable as monomers.[24] We reasoned that predicting monomers instead of dimers could reveal such complexes because we assumed that their structure would change between both states due to their instability as monomers. We generated monomer models for the human proteome and compared the resulting structures with those of individual chains in the dimer models. We assessed structural similarity by the template-modeling score (TM-score),[25] where values close to 1 indicate high structural similarity, and values close to 0 reflect a lack thereof. Surprisingly, we observed a high similarity in structure between chains in either state, with less than 4% of homodimer chains showing a structure highly different from that of the monomer (TM-score < 0.7) (Figure 2D). Such a degree of similarity was unexpected because a large fraction of our dataset corresponds to structures with dozens of residues stabilized by intermolecular contacts (Figure 2E). This observation shows that AlphaFold2 almost systematically identifies native-like structures of protein chains forming obligate homo-oligomeric complexes, even in the absence of their partner.

A large number of proteins exhibit extensive intermolecular interactions, emphasizing the importance of representing these models in their quaternary dimension. This representation will therefore be essential to fully leverage the recent explosion of structural information and gain biological insights. For example, the tripartite motif (TRIM)-containing protein 77 is stabilized by considerable inter-subunit contacts (Figure 2F); however, the monomer chain exhibits a similar structure as in the dimer. Moreover, the dimer information is key to visualizing the multivalent and spatial organization of the RING and SPRY domains in this protein—information that is absent from the monomeric structure. In a different example, the transcription factor AP-2-α (Figure 2G) exhibits extensive intermolecular contacts. This family of helix-span-helix transcription factors has no experimentally determined structure, and accordingly, this quaternary structure

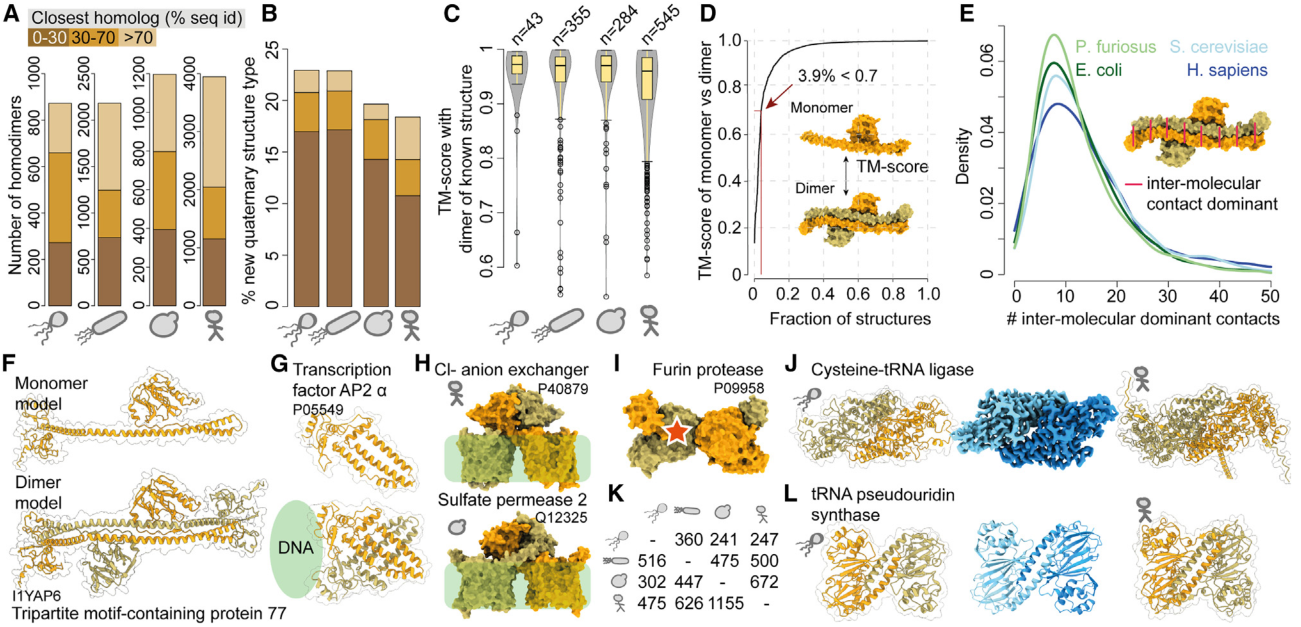


Figure 2. Proteome-wide discovery of homodimers

(A) Numbers of high-scoring dimers in each species' dataset and their similarity to known structures.

(B) New quaternary structure types correspond to dimer models for which no matching homologous dimer was detected in the PDB, and the percentage of such dimers is given for each species' dataset.

(C) Distribution of TM-scores of the structural superposition between dimer models and their matching experimental structure when available, shown as violins. Boxes show the interquartile range, whiskers extend to the 25th and 75th percentiles. Outliers are detailed in Figure S2.

(D) Cumulative distribution of TM-scores between the structure of chains predicted independently as homodimers or as monomers.

(E) Distribution of the number of intermolecular dominant contacts in each species' dataset. Intermolecular dominant contacts are likely stabilizing the structure through dimer formation, as illustrated in the structure (inset).

(F) The monomer (top) and dimer (bottom) models of the tripartite-motif-containing protein 77. Extensive intermolecular contacts in this dimer highlight the necessity to consider oligomerization information for interpreting the wealth of structural data.

(G) The transcription factor AP-2 α functions as a dimer,[23] but its structure is unresolved. Our model represents a new quaternary structure type (i.e., not observed in the PDB) and is compatible with biochemical data.[23]

(H) Membrane transporters that adopt a quaternary structure type that is shared between human and yeast and is absent from the training set.

(I) The dimer model of the furin protease shows a pro-form that trans-inactivates. The red star indicates the catalytic site obstructed by binding of the partner.

(J) A cysteine tRNA ligase shows a novel dimer quaternary structure type that is conserved between *P. furiosus* and *H. sapiens* and is absent from the PDB. We solved the structure of the protein from *P. furiosus* by electron microscopy. The density map (blue) highlights the same dimer interaction geometry (TM-score = 0.99).

(K) Number of dimers from each species (line) sharing structural homology with dimers from the other species (columns).

(L) A tRNA pseudouridine synthase shows a novel and similar dimer quaternary structure type between *P. furiosus* and *H. sapiens*. We solved the structure of the protein from *P. furiosus* (Q8U2C1) by X-ray crystallography (blue), which revealed the same dimer structure (TM-score = 0.99).

See also Figures S2 and S3 and Tables S6 and S7.

type is novel with no homodimer homolog detected across the PDB. Moreover, this dimer matches existing biochemical and mutational data,[23] providing further validation of its accuracy. We also identified novel quaternary structure types among membrane proteins. For example, the chloride/bicarbonate anion exchanger S26A3 shows an interface geometry that is absent from the training dataset but is substantiated by the recent characterization of a homodimer homolog.[26] Interestingly, the proteome-wide nature of our predictions enabled the comparison of these structures across organisms and revealed a homologous sulfate transporter in *S. cerevisiae* (Figure 2H).

These models also pinpoint potential regulatory features. For example, the furin protease is a key enzyme that processes cellular precursor proteins and viral factors essential for the function of HIV, influenza, and SARS-CoV-2. This protease must remain inactive in its intra-cellular form to avoid mis-cleavage events, but the structure of its pro-form is unknown. Our models suggest that furin and other family members, including those in *P. furiosus*, trans-inactivate as dimers whereby each chain binds and obstructs the catalytic site of its partner (Figure 2I).

In a different example, we identified a cysteine tRNA ligase as exhibiting a new quaternary structure type conserved in *P. furiosus* and *H. sapiens*. We used cryo-EM to solve the structure of the *P. furiosus* protein, which revealed a homodimer closely matching the predicted model (TM-score = 0.99; Figures 2J, S3A, and S3B). The comprehensive nature of our datasets renders them suitable to analyze homo-oligomerization conservation and evolution. We found that 247 and 500 dimer structures from *P. furiosus* and *E. coli*, respectively, shared structural homology with dimers from the human proteome.

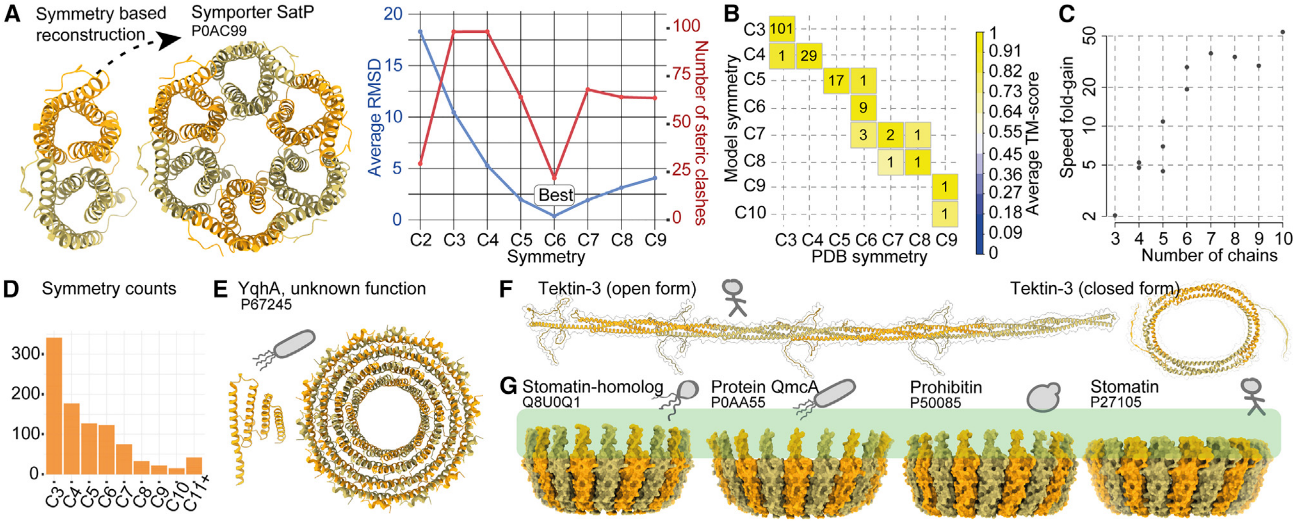


Figure 3. Proteome-wide discovery of ring complexes and filaments

(A) The symmetry information contained in a dimer model is used to find the best-compatible ring symmetry.

(B) The structure of cyclic complexes so obtained was compared with known experimental structures, showing that 95% (160/168) of cyclic symmetries are inferred correctly. The average TM-score is above 0.93 for all bins on the diagonal and is above 0.7 for off-diagonal predictions, except for the C8-C7 bin (0.69).

(C) Speed fold change observed in the prediction of complexes containing three to ten chains.

(D) Numbers and types of ring symmetries reconstructed across the four proteomes.

(E) A monomer of Yqha is shown next to its ring structure, which represents a novel quaternary structure type.

(F) Tektin-3 is predicted as forming a filamentous structure identical to that recently proposed[28] and a novel closed form is also predicted.

(G) Stomatin and prohibitin proteins are membrane associated. They assemble as rings through an interface geometry conserved across the four proteomes studied.

See also Figure S4B.

Conversely, 475 and 626 dimers in the human proteome shared structural homology to a dimer from *P. furiosus* or *E. coli*, respectively (Figure 2K). These data can also serve to identify cases of divergence and will provide a basis for comprehensive analyses of interface and oligomeric state evolution. One notable example is the tRNA pseudo-uridine synthase. This protein adopts a novel quaternary structure type observed in the proteomes of *P. furiosus*, yeast, and human. We solved the structure of the *P. furiosus* protein by X-ray crystallography, which revealed a dimer almost identical to the model (TM-score = 0.99; Figure 2L). Interestingly, in *E. coli*, the interaction interface occurs at a similar surface site but is mediated by extended loop regions absent in other species (Figure S3C).

Taken together, these analyses show that these structure models are reliable, that they increase the structural coverage of homo-oligomer information by ~50% in human to >100% in *P. furiosus*, and that they contain hundreds of quaternary structure types, thereby providing a rich resource for functional and evolutionary analyses.

Proteome-wide discovery of ring- and filament-forming homo-oligomers

A majority of homo-oligomers form homotypic or "head-to-head" interfaces, resulting in dimers with C2 symmetry. A different type of assembly involves heterotypic or "head-to-tail" interactions, which create ring structures and filaments. These rings are difficult to predict due to the uncertainty in the number of subunits and due to their large size. However, we reasoned that the symmetry information contained within a dimer could suffice to reconstruct ring-like and filament-forming complexes. This concept is illustrated with the synporter SatP, where the predicted dimer model interacts head-to-tail. The rotation information contained within the dimer is best compatible with C6 symmetry, which we identified through an analytical method[27] (Figure 3A). This strategy yields a model of SatP closely matching the experimental structure (TM-score = 0.99). Comparing the symmetries derived with this approach to their matching experimental structures also reveals an excellent agreement, with 95% (160/168) of cyclic symmetries being inferred correctly (Figure 3B). This strategy thus tackles both limits of Alphafold2, first by inferring the number of subunits given the symmetry of a dimer and second by keeping manageable the resources required for predicting these complexes. Indeed, we compared time and memory requirements for predictions with AlphaFold2 multimer.[12] Complexes with 4 to 10 subunits required 5- to 50-fold more time (Figures 3C and S4A) and 1.3- to 6.5-fold more GPU memory than that required for their respective dimer predictions.

One drawback of the symmetry-based reconstruction of ring complexes is that loops could be intertwined, and flexible regions sometimes clashed extensively with the ring structure. To overcome this problem, we developed an AlphaFold2 protocol that makes use of the backbone of the complete symmetry-generated structure to produce final models. We initially

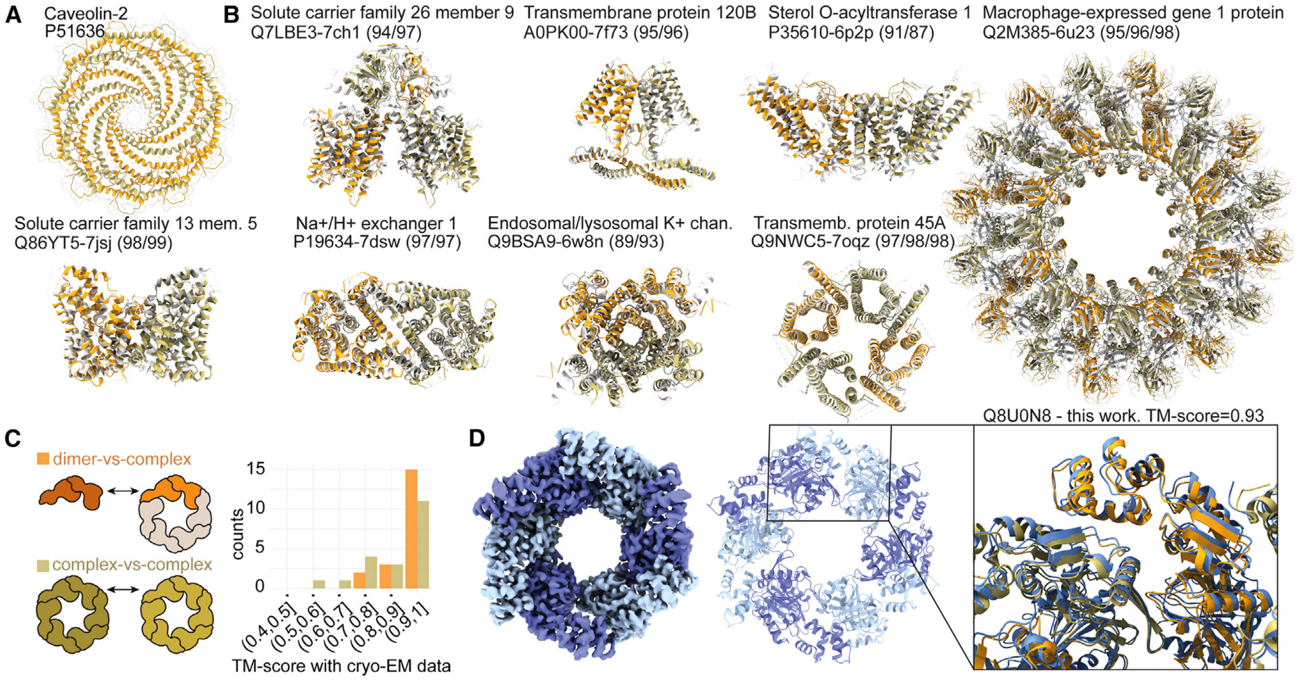


Figure 4. Assessing the accuracy of predictions based on recent cryo-EM structures

(A) The model of caveolin-2 is similar to a recently published structure of caveolin-1, absent from the training set.[29]

(B) Models of cyclic complexes (orange-green) superposed to cryo-EM structures (white) released after May 2018 and with no close homolog in the training set. The name and UniProt ID of the proteins are given along with the PDB code of the matching structure. Numbers in parenthesis indicate the TM-score × 100 for the monomer, dimer, and complex (for those with more than two subunits) superposition.

(C) Histogram of TM-score between the dimer (orange) or ring (green) models against their matching EM structures.

(D) Using cryo-EM, we solved the hexameric structure of a protein of unknown function from *P. furiosus*. Our model captured the quaternary structure of this complex with high accuracy. Left shows cryo-EM density, middle the corresponding atomic model, and right the overlay between the experimental structure (blue) and the predicted model (orange).

See also Figures S5 and S6 and Table S6.

supplied the structural information to AlphaFold2 as a template of which the side-chain information had been masked, limiting a too heavy bias toward the generated structure. However, we found that many recycles were needed and were sometimes not sufficient for large structures. We hypothesized that this was due to the "black hole" initialization of the structure module, where atomic coordinates are all initialized at zero. Hence, we implemented what we call a "big bang" initialization, where instead of initializing the coordinates at zero, we initialized them to the input structure, resulting in faster and consistent convergence to the final model (STAR Methods). This protocol allowed us to reconstruct the ring complexes with up to 6,500 residues in total while resolving the clashes introduced by the symmetry-based model generation.

This strategy enabled us to reconstruct hundreds of ring complexes (Figures 3D and S4B), many of which represent novel quaternary structure types. One example is Yqha, a protein of unknown function from *E. coli*. The monomer structure of this protein consists of four helices interacting laterally, which appears highly unstable due to the absence of a protein core. By contrast, our model shows how the four helices pack with additional copies to form a ring structure containing 14 subunits (Figure 3E). In a different example, we noticed an unusual structure of intertwined α helices for caveolin-2 (Figure 4A), closely resembling that of caveolin-1 solved by cryo-EM[29] and absent from AlphaFold2 training set. This example motivated us to evaluate the accuracy of the models against human homo-oligomers specifically solved by cryo-EM. The models matched these structures closely, as illustrated for the transmembrane protein 45A (TM-score = 0.95; Figure 4B), or the megadalton complex of macrophage-expressed gene 1 (TM-score = 0.98; Figure 4B). Overall, out of 20 complexes compared (Table S3), the median TM-score was 0.92, reflecting an excellent agreement (Figures 4C and S5). Most of the mismatches between models and experimental structures were caused by differences in the cyclic symmetry that we inferred (e.g., C12 instead of C11) in the case of the human calcium homeostasis modulator 5 (PDB: 7d60[30]). Such symmetry changes involve minute structural differences in the dimer interaction geometry, and proteins can frequently adopt multiple states.[31] This is also observed among viruses adopting a quasi-symmetry.[32] Indeed, focusing the benchmark on the dimer interaction geometry, our prediction accuracy for the same structures increases significantly (Figure 4C, orange).

To validate our predictions, we selected several ring-forming proteins for structural determination. One of them, which was predicted to form a hexamer, could be expressed and was amenable to analysis by cryo-EM. This protein from *P. furiosus* (Q8U0N8) is uncharacterized and does not have a PFAM domain

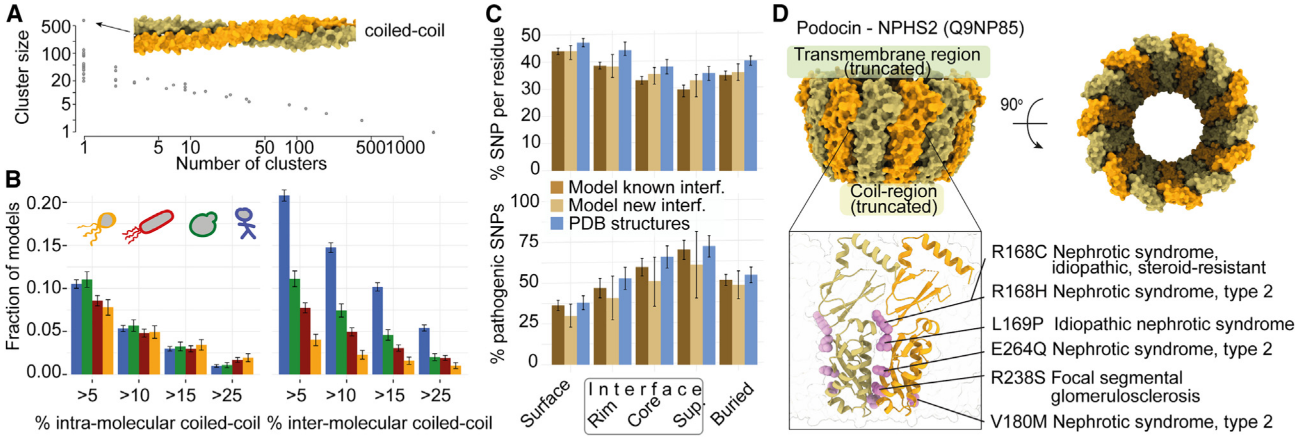


Figure 5. Analyzing the quaternary structure datasets provides insights into proteome evolution, disease mutations, and structure

(A) Clustering all dimer models by structural similarity yields 2,991 clusters. The scatter plot depicts the number of clusters having a particular size. At the extremes, there are 1,957 singletons, and the largest cluster contains over 500 structures and corresponds to coiled-coil-mediated dimerization.

(B) Barplot of the fraction of models containing intra (left) or intermolecular (right) coiled coils. Frequencies are comparable across kingdoms for intramolecular coiled coils, whereas intermolecular coiled coils are more prominent in eukaryotes. Bars show two standard errors of the estimated mean.

(C) Barplot depicting the percentage of SNPs normalized by residues (top) or pathogenic SNPs normalized by benign and pathogenic (bottom) across protein regions as defined by Levy.[43] Three types of structures are compared: human structures from the PDB (blue), models with quaternary structure types that are previously observed (dark brown), and those that are novel (light brown). Error bars show the 95% confidence interval on the median (top) or mean (bottom).

(D) Predicted ring structure of podocin, which contains fourteen subunits and is bound to the plasma membrane through an α helix absent from the model. Two podocin chains are highlighted, and interface residues with known pathogenic mutations from ClinVar[44] are shown (purple) along with the name of the associated condition.

See also Figure S7 and Table S4.

assignment.[33] We solved its structure to a resolution of 2.8 Å, which revealed a close agreement with our prediction. The global TM-score between our model and the determined structure was 0.93 (Figure 4D). Interestingly, the protein contains an N-terminal domain that exhibits varying degrees of flexibility, as indicated by the reduced local resolution (Figure S6). Not considering this domain, our model showed higher agreement with the experimental data (TM-score = 0.96, Cα-root-mean-square deviation [RMSD] = 2.4 Å).

Beyond ring complexes, we also identified 179 models expected to form filamentous assemblies (Figure S4B; Table S2). One such example is Tektin-3, a component of dynein-decorated doublet microtubules. Here, the filament structure is identical to that recently proposed.[28] Interestingly, one of the five models of Tektin-3 converged toward a different conformation corresponding to a homodimer (Figure 3F). We speculate that such a closed structure could be adopted after synthesis to facilitate delivery to doublet microtubules. Finally, we also identified one novel quaternary structure type conserved in all kingdoms. The proteins forming these structures were annotated with an “Ambiguous” or “*trans*” (translational) symmetry (Figure S4B) because they contain a flexible coil conflicting with the symmetry search procedure. However, upon truncating that region, the procedure predicts ring-shaped assemblies containing about 20 chains and compatible with negative-stain images of yeast prohibitin[34] (Figure 3G). In human, Stomatin and homologous proteins such as podocin associate with lipid rafts and diverse ion channels to regulate their activities.[35,36] The molecular basis of these interactions remains unknown despite their association with numerous diseases,[37] and the ring-shaped assembly characterized here provides a structure with which they can be interpreted, as we will show in the next section.

Evolutionary and structural insights from proteome-wide homo-oligomerization

The proteome-wide characterization of quaternary structures paves the way to a molecular description of proteomes, both in health and disease. We employed the newly characterized datasets to investigate three general molecular properties of proteomes.

First, we clustered the dimer models by structural similarity to identify the most frequent type of structure associated with homo-oligomerization. The largest cluster involved intermolecular coiled coils (Figure 5A), motivating an analysis of their representation across proteomes. While coiled-coil regions can be detected from sequence alone,[38] such predictions show higher false-positive rates and lower sensitivity than structure-based assignment methods.[39–41] Moreover, such sequence-based approaches cannot distinguish inter and intramolecular coiled coils, whereas our data enable comparing both types. Therefore, we used the structure-based method, SOCKET,[40] to identify coiled coils in our quaternary structure models. Our analysis revealed that intramolecular coiled coils exist at comparable frequencies across the four proteomes: 10.5% of proteins in the human dataset contained >5% of intramolecular coiled coils, versus 11% in yeast, 8.6% in *E. coli*, and 7.8% in *P. furiosus* (Figure 5B). In contrast, intermolecular coiled coils showed a marked increase in the human proteome, with 20.7% of homo-oligomers containing >5% of intermolecular coiled coils, versus 11.1%, 7.8%, and 4% for yeast, *E. coli*, and *P. furiosus*, respectively

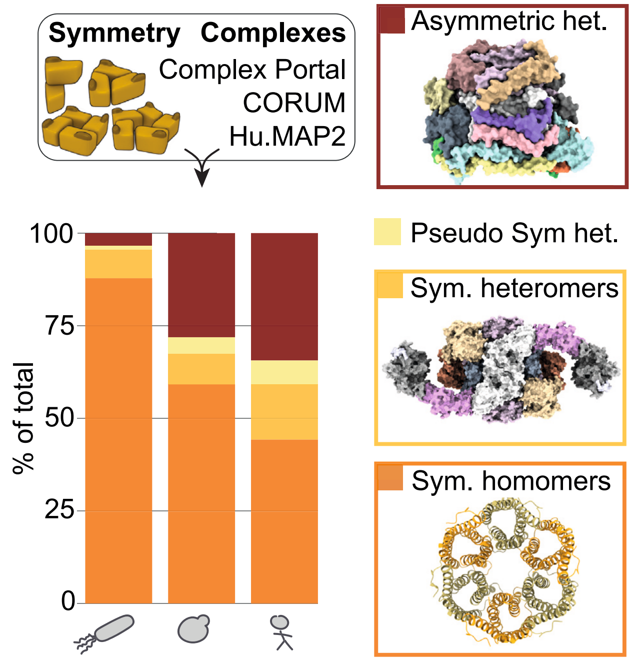


Figure 6. Integrating our datasets with omics data reveals a ubiquity of symmetry in cellular protein complexes

The proteome-wide quaternary structure information was integrated with omics data on protein complexes. Each complex (left) was assigned one of four types based on its protein composition: homo-oligomeric, symmetric heteromer, pseudo-symmetric heteromer, and asymmetric (STAR Methods). This analysis reveals a striking ubiquity of symmetry among protein complexes.

(Figure 5B). This finding implies that coiled-coil regions have been major enablers of quaternary structure evolution in the human lineage. Given the ability to design coiled-coil interfaces,[42] this finding opens prospects for designing coiled-coil peptides and proteins to probe and intervene in many cellular processes.

Second, we projected 662,413 non-synonymous single-nucleotide polymorphisms (SNPs) onto quaternary structures of the human dataset (STAR Methods). We considered separately known structures and models and further separated the models into those involving a known interface type or a novel one. We found that interfaces contained lower SNP frequencies than non-interface regions with equivalent solvent exposure in the monomer (Figures 5C and S7), consistent with the former being under stronger purifying selection than the latter (interface-core versus surface: $p < 0.0001$; other comparisons and effect sizes are detailed in Table S4). Furthermore, interfaces showed significant enrichment in disease-associated SNPs when compared with non-interface regions (interface-core versus surface: +71.6%, $p = 0.0027$; Table S4). The enrichment of disease-associated SNPs at interaction interfaces generalizes previous observations based on existing structures.[45,46] Importantly, the enrichment is of a similar magnitude among experimentally characterized structures (+74.4%), supporting the idea that the predicted quaternary structure types are as likely to be involved in diseases via interface mutations. A notable example is podocin, a protein expressed in podocyte cells, which act as filters in the blood-urine barrier. Several mutations associated with nephrotic syndromes and renal failure appear at the interface of podocin (Figure 5D), suggesting that they impair its assembly into rings.

Third, we estimated the prevalence of symmetry among all protein complexes characterized to date by proteomics experiments. We gathered information on protein complex composition from multiple sources (STAR Methods) and assigned each complex to one of four categories: symmetric homo-oligomer, symmetric heteromer, pseudo-symmetric heteromer, and asymmetric heteromer. Each assignment was made depending on the complex composition in homo-oligomer-forming proteins and paralogous sequences (STAR Methods). We found that a majority of protein complexes form symmetrical assemblies (Figure 6; Table S5). This is especially striking in *E. coli*, where more than 90% of the complexes form symmetrical homo- or hetero-oligomers. In eukaryotes, we found that 60%–65% of complexes are symmetric, and these numbers increase to 65%–70% when including pseudo-symmetries. These numbers highlight the ubiquity of symmetry in proteomes, which is a key point to consider when analyzing protein complex evolution and assembly.[47–49]

DISCUSSION

We have characterized protein quaternary structures across four proteomes with very high accuracy. Importantly, inferring quaternary structure information is challenging even when an experimental structure is characterized by X-ray crystallography. This is due to the difficulty in distinguishing fortuitous crystal contacts from physiological ones. These difficulties mean that upward of 10% of biological assemblies available in the PDB are estimated to be non-physiological.[50] Consequently, the accuracy of our predictions can be compared with quaternary structure information originating from X-ray crystallography. In addition, a significant advantage of our approach is the inclusion of full-length proteins, which can reveal homo-oligomerization modes that are missing or altered in truncated proteins. The strategy devised in this work can be readily scaled up to cover a larger number of organisms relevant to fundamental and applied biology.

Here, we have focused on four proteomes—one archaeon, one bacterium, and two eukaryotes—and increased the quaternary structure coverage of these organisms by 50%–100%. Notably, only a handful of folds have been discovered among ∼600,000 monomeric structures predicted by AlphaFold2.[51] By contrast, we have identified hundreds of novel quaternary structure types, implying that this space is much larger than that of tertiary structures. Remarkably, some of these novel quaternary structure types are conserved across the tree of life, implying their functional importance.

These proteome-wide datasets will aid biologists in gaining insights into specific proteins, and excitingly, it will provide a basis for a structure-guided understanding of proteome assembly. Here, we use these data to conduct three general analyses of proteomes' quaternary structures. We observed that coiled coils are important mediators of quaternary structure evolution. We found that interaction interfaces in homo-oligomers are hotspots of disease-associated polymorphisms. Finally, we revealed that 60% of known protein complexes (both homo- and hetero-oligomeric) appear symmetric in yeast and human, and this number increased to over 90% in *E. coli*.

By expanding our knowledge of the protein quaternary structure space, this research opens up exciting possibilities for interpreting network, omics, disease, and evolutionary data through a structural lens, ultimately aiding our understanding of the fundamental principles of proteome assembly and evolution.

Limitations of the study

This work focused on the detection of homo-oligomers with cyclic symmetry and will serve as a basis for the detection of dihedral and cubic groups, which is more computationally demanding to model. The evolution of GPUs with increased speed and memory will render the modeling of these groups feasible on a proteome-wide scale. Another computational challenge in our work was the cost of quaternary structure superposition, which we used for identifying novel quaternary structure types. Comprehensive structure-based searches will require the development of methods such as Foldseek[52] that are applicable to multi-chain complexes. Such developments will enable us to explore the space of quaternary structures with higher confidence to refine the numbers presented in this work.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Generating homodimer models
  - ○ Protein structure processing
  - ○ Selecting a representative model per protein
  - ○ Evaluating the confidence of homodimer models
  - ○ Quaternary structure searches against the PDB
  - ○ Clustering quaternary structures
  - ○ Defining novel quaternary structure types
  - ○ Symmetry analysis and reconstruction
  - ○ Comparing monomer and homodimer models
  - ○ Comparing the models against cryo-EM data
  - ○ Detection of coiled coil domains
  - ○ Generation of final models
  - ○ Analysis of SNPs
  - ○ Prevalence of symmetry analysis
  - ○ Protein purification
  - ○ Molecular weight determination
  - ○ Structure determination by cryo-EM
  - ○ Structure determination by crystallography
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.01.022.

ACKNOWLEDGMENTS

We thank Sergei Grudinin for help with AnAnaS as well as Harry Greenblatt and the HPC Weizmann team for help with the computer infrastructure. We thank the Protein Production and Structure Characterization Core Facility, EPFL, Switzerland for help with protein purification. We thank the Dubochet Center for Imaging (DCI), EPFL-UNIL-UNIGE, Switzerland, for cryo-EM data collection. We thank Sreenath Nair, Mihaly Varadi, and Sameer Velankar for help with integration of our data into 3DBeacon. We thank Anil Pasupulati, John Jumper, Moran Shalev Benami, Patrick Barth, Pierre Gönczy, Robert Jefferson, Sven Dahms, and the members of E.D.L. and B.E.C. laboratories for helpful discussions. We thank the anonymous reviewers for helpful and constructive feedback. M.P. was supported by the Peter und Traudl Engelhorn Stiftung. P.K. was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) grant to D.N.W. (BB/R00661X/1). D.N.W. was also supported by the BrisSynBio, a BBSRC/Engineering and Physical Sciences Research Council-funded Synthetic Biology Research Centre (BB/L01386X/1). B.E.C. was supported by the Swiss National Science Foundation, the NCCR in Chemical Biology, the NCCR in Molecular Systems Engineering, and the Swiss Cancer League (grant number KFS-5032-02-2020). E.D.L. acknowledges support by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 819318), by a research grant from HFSP (ref.-no: RGP0016/2022), by the Israel Science Foundation (grant no. 1452/18), and by the Abisch-Frenkel Foundation. S.O. and C.A.G. were supported by Amgen for this project. S.D. acknowledges the Ramalingaswami research grant from the Department of Biotechnology, Govt. India (RLS: BT/RLF/Re-entry/10/2020, sanction order serial number 145).

AUTHOR CONTRIBUTIONS

H.S., S.D., and E.D.L. carried out structure processing and analyses; T.L. carried out the SNP analysis with help from B.D.; P.K. and H.S. performed the coiled-coil analysis; C.A.G. refined the structures and implemented the big bang protocol with S.O. to generate final models; M.P., Y.D., L.J.D., and S.G. carried out protein expression and purification assays; M.P. processed cryo-EM and crystallographic data and structures; D.N.W., B.E.C., S.D., and E.D.L. oversaw and funded the research; H.S. and E.D.L. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 8, 2023
Revised: November 3, 2023
Accepted: January 15, 2024
Published: February 6, 2024

REFERENCES

1. Goodsell, D.S., and Olson, A.J. (2000). Structural symmetry and protein function. Annu. Rev. Biophys. Biomol. Struct. *29*, 105–153.
2. Levy, E.D., and Teichmann, S. (2013). Structural, evolutionary, and assembly principles of protein oligomerization. Prog. Mol. Biol. Transl. Sci. *117*, 25–51.
3. Yeates, T.O., Liu, Y., and Laniado, J. (2016). The design of symmetric protein nanomaterials comes of age in theory and practice. Curr. Opin. Struct. Biol. *39*, 134–143.
4. Marzahn, M.R., Marada, S., Lee, J., Nourse, A., Kenrick, S., Zhao, H., Ben-Nissan, G., Kolaitis, R.-M., Peters, J.L., Pounds, S., et al. (2016).

Higher-order oligomerization promotes localization of SPOP to liquid nuclear speckles. EMBO J. *35*, 1254–1275.

5. Marianayagam, N.J., Sunde, M., and Matthews, J.M. (2004). The power of two: protein dimerization in biology. Trends Biochem. Sci. *29*, 618–625.
6. Calvanese, L., D'Auria, G., Vangone, A., Falcigno, L., and Oliva, R. (2018). Structural Basis for Mutations of Human Aquaporins Associated to Genetic Diseases. Int. J. Mol. Sci. *19*, 1577.
7. Pereira-Leal, J.B., Levy, E.D., Kamp, C., and Teichmann, S.A. (2007). Evolution of protein complexes by duplication of homomeric interactions. Genome Biol. *8*, R51.
8. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.
9. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876.
10. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444.
11. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596.
12. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. (2021). Protein complex prediction with AlphaFold-Multimer. https://doi.org/10.1101/2021.10.04.463034.
13. Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. Nat. Commun. *13*, 1265.
14. Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T.J., Banjade, S., Bagde, S.R., et al. (2021). Computed structures of core eukaryotic protein complexes. Science *374*, eabm4805.
15. Burke, D.F., Bryant, P., Barrio-Hernandez, I., Memon, D., Pozzati, G., Shenoy, A., Zhu, W., Dunham, A.S., Albanese, P., Keller, A., et al. (2023). Towards a structurally resolved human protein interaction network. Nat. Struct. Mol. Biol. *30*, 216–225.
16. Pang, S.S., Bayly-Jones, C., Radjainia, M., Spicer, B.A., Law, R.H.P., Hodel, A.W., Parsons, E.S., Ekkel, S.M., Conroy, P.J., Ramm, G., et al. (2019). The cryo-EM structure of the acid activatable pore-forming immune effector Macrophage-expressed gene 1. Nat. Commun. *10*, 4288.
17. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
18. Michalak, L., La Rosa, S.L., Leivers, S., Lindstad, L.J., Røhr, Å.K., Lillelund Aachmann, F., and Westereng, B. (2020). A pair of esterases from a commensal gut bacterium remove acetylations from all positions on complex β-mannans. Proc. Natl. Acad. Sci. USA *117*, 7122–7130.
19. Zhuang, N., Zhang, H., Li, L., Wu, X., Yang, C., and Zhang, Y. (2020). Crystal structures and biochemical analyses of the bacterial arginine dihydrolase ArgZ suggests a "bond rotation" catalytic mechanism. J. Biol. Chem. *295*, 2113–2124.
20. Manissorn, J., Sitthiyotha, T., Montalban, J.R.E., Chunsrivirot, S., Thongnuek, P., and Wangkanont, K. (2020). Biochemical and Structural Investigation of GnnA in the Lipopolysaccharide Biosynthesis Pathway of Acidithiobacillus ferrooxidans. ACS Chem. Biol. *15*, 3235–3243.
21. Schweke, H., Xu, Q., Tauriello, G., Pantolini, L., Schwede, T., Cazals, F., Lhéritier, A., Fernandez-Recio, J., Rodríguez-Lumbreras, L.A., Schueler-Furman, O., et al. (2023). Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. Proteomics *23*, e2200323.
22. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489.
23. Williams, T., and Tjian, R. (1991). Characterization of a dimerization motif in AP-2 and its function in heterologous DNA-binding proteins. Science *251*, 1067–1071.
24. Nooren, I.M.A., and Thornton, J.M. (2003). Diversity of protein-protein interactions. EMBO J. *22*, 3486–3492.
25. Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins *57*, 702–710.
26. Walter, J.D., Sawicka, M., and Dutzler, R. (2019). Cryo-EM structures and functional characterization of murine Slc26a9 reveal mechanism of uncoupled chloride transport. eLife *8*, e46986.
27. Pagès, G., Kinzina, E., and Grudinin, S. (2018). Analytical symmetry detection in protein assemblies. I. Cyclic symmetries. J. Struct. Biol. *203*, 142–148.
28. Gui, M., Farley, H., Anujan, P., Anderson, J.R., Maxwell, D.W., Whitchurch, J.B., Botsch, J.J., Qiu, T., Meleppattu, S., Singh, S.K., et al. (2021). *De novo* identification of mammalian ciliary motility proteins using cryo-EM. Cell *184*, 5791–5806.e19.
29. Porta, J.C., Han, B., Gulsevin, A., Chung, J.M., Peskova, Y., Connolly, S., Mchaourab, H.S., Meiler, J., Karakas, E., Kenworthy, A.K., et al. (2022). Molecular architecture of the human caveolin-1 complex. Sci. Adv. *8*, eabn7232.
30. Liu, J., Wan, F., Jin, Q., Li, X., Bhat, E.A., Guo, J., Lei, M., Guan, F., Wu, J., and Ye, S. (2020). Cryo-EM structures of human calcium homeostasis modulator 5. Cell Discov. *6*, 81.
31. Marciano, S., Dey, D., Listov, D., Fleishman, S.J., Sonn-Segev, A., Mertens, H., Busch, F., Kim, Y., Harvey, S.R., Wysocki, V.H., et al. (2022). Protein quaternary structures in solution are a mixture of multiple forms. Chem. Sci. *13*, 11680–11695.
32. Caspar, D.L., and Klug, A. (1962). Physical principles in the construction of regular viruses. Cold Spring Harb. Symp. Quant. Biol. *27*, 1–24.
33. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427–D432.
34. Tatsuta, T., Model, K., and Langer, T. (2005). Formation of membrane-bound ring complexes by prohibitins in mitochondria. Mol. Biol. Cell *16*, 248–259.
35. Huber, T.B., Schermer, B., Müller, R.U., Höhne, M., Bartram, M., Calixto, A., Hagmann, H., Reinhardt, C., Koos, F., Kunzelmann, K., et al. (2006). Podocin and MEC-2 bind cholesterol to regulate the activity of associated ion channels. Proc. Natl. Acad. Sci. USA *103*, 17079–17086.
36. Montel-Hagen, A., Kinet, S., Manel, N., Mongellaz, C., Prohaska, R., Battini, J.L., Delaunay, J., Sitbon, M., and Taylor, N. (2008). Erythrocyte Glut1 triggers dehydroascorbic acid uptake in mammals unable to synthesize vitamin C. Cell *132*, 1039–1048.
37. Browman, D.T., Hoegg, M.B., and Robbins, S.M. (2007). The SPFH domain-containing proteins: more than lipid raft markers. Trends Cell Biol. *17*, 394–402.
38. Rackham, O.J.L., Madera, M., Armstrong, C.T., Vincent, T.L., Woolfson, D.N., and Gough, J. (2010). The evolution and structure prediction of coiled coils across all genomes. J. Mol. Biol. *403*, 480–493.
39. Walshaw, J., and Woolfson, D.N. (2001). Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J. Mol. Biol. *307*, 1427–1450.
40. Kumar, P., and Woolfson, D.N. (2021). Socket2: A Program for Locating, Visualising, and Analysing Coiled-coil Interfaces in Protein Structures. Bioinformatics *37*, 4575–4577.

41. Simm, D., Hatje, K., Waack, S., and Kollmar, M. (2021). Critical assessment of coiled-coil predictions based on protein structure data. Sci. Rep. *11*, 12439.

42. Woolfson, D.N. (2023). Understanding a protein fold: the physics, chemistry, and biology of α-helical coiled coils. J. Biol. Chem. *299*, 104579.

43. Levy, E.D. (2010). A simple definition of structural regions in proteins and its use in analyzing interface evolution. J. Mol. Biol. *403*, 660–670.

44. Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. Nucleic Acids Res. *48*, D835–D844.

45. Livesey, B.J., and Marsh, J.A. (2022). The properties of human disease mutations at protein interfaces. PLoS Comput. Biol. *18*, e1009858.

46. David, A., Razali, R., Wass, M.N., and Sternberg, M.J.E. (2012). Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum. Mutat. *33*, 359–363.

47. Marsh, J.A., Hernández, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V., and Teichmann, S.A. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell *153*, 461–470.

48. Ahnert, S.E., Marsh, J.A., Hernández, H., Robinson, C.V., and Teichmann, S.A. (2015). Principles of assembly reveal a periodic table of protein complexes. Science *350*, aaa2245.

49. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N., and Levy, E.D. (2017). Proteins evolve on the edge of supramolecular self-assembly. Nature *548*, 244–247.

50. Dey, S., Ritchie, D.W., and Levy, E.D. (2018). PDB-wide identification of biological assemblies from conserved quaternary structure geometry. Nat. Methods *15*, 67–72.

51. Bordin, N., Sillitoe, I., Nallapareddy, V., Rauer, C., Lam, S.D., Waman, V.P., Sen, N., Heinzinger, M., Littmann, M., Kim, S., et al. (2023). AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. Commun. Biol. *6*, 160.

52. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-01773-0.

53. Varadi, M., Nair, S., Sillitoe, I., Tauriello, G., Anyango, S., Bienert, S., Borges, C., Deshpande, M., Green, T., Hassabis, D., et al. (2022). 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. GigaScience *11*, giac118.

54. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

55. Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T., and Bricogne, G. (2011). Data processing and analysis with the autoPROC toolbox. Acta Crystallogr. D Biol. Crystallogr. *67*, 293–302.

56. Kabsch, W. (2010). XDS. Acta Crystallogr. D Biol. Crystallogr. *66*, 125–132.

57. Liebschner, D., Afonine, P.V., Baker, M.L., Bunkóczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. D Struct. Biol. *75*, 861–877.

58. Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501.

59. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. *30*, 70–82.

60. Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., et al. (2018). MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. *27*, 293–315.

61. Mukherjee, S., and Zhang, Y. (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res. *37*, e83.

62. Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309.

63. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.

64. Ritchie, D.W., Ghoorah, A.W., Mavridis, L., and Venkatraman, V. (2012). Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. Bioinformatics *28*, 3274–3281.

65. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682.

66. Drew, K., Wallingford, J.B., and Marcotte, E.M. (2021). hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Mol. Syst. Biol. *17*, e10016.

67. Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S.J. (2009). Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res. *37*, 825–831.

68. Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V., et al. (2008). CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. *36*, D646–D650.

69. Meldal, B.H.M., Bye-A-Jee, H., Gajdoš, L., Hammerová, Z., Horácková, A., Melicher, F., Perfetto, L., Pokorný, D., Lopez, M.R., Türková, A., et al. (2019). Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. Nucleic Acids Res. *47*, D550–D558.

70. Karp, P.D., Billington, R., Caspi, R., Fulcher, C.A., Latendresse, M., Kothari, A., Keseler, I.M., Krummenacker, M., Midford, P.E., Ong, Q., et al. (2019). The BioCyc collection of microbial genomes and metabolic pathways. Brief. Bioinform. *20*, 1085–1093.

71. Keseler, I.M., Gama-Castro, S., Mackie, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Kothari, A., Krummenacker, M., Midford, P.E., Muñiz-Rascado, L., et al. (2021). The EcoCyc Database in 2021. Front. Microbiol. *12*, 711077.

72. Levy, E.D., Pereira-Leal, J.B., Chothia, C., and Teichmann, S.A. (2006). 3D complex: a structural classification of protein complexes. PLoS Comput. Biol. *2*, e155.

73. Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Res. *5*, 189.

74. Pearson, W.R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. *183*, 63–98.

75. Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A.N., and Alva, V. (2020). Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr. Protoc. Bioinformatics *72*, e108.

76. Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.-P., Simmonett, A.C., Harrigan, M.P., Stern, C.D., et al. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput. Biol. *13*, e1005659.

77. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

78. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067.

79. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., and Gama-Castro, S. (2002). The EcoCyc Database. Nucleic Acids Res. *30*, 56–58.

80. Punjani, A., Rubinstein, J.L., Fleet, D.J., and Brubaker, M.A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods *14*, 290–296.

81. Punjani, A., Zhang, H., and Fleet, D.J. (2020). Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. Nat. Methods *17*, 1214–1221.

82. Guardia, C.M., Tan, X.-F.X.F., Lian, T., Rana, M.S., Zhou, W., Christenson, E.T., Lowry, A.J., Faraldo-Gómez, J.D., Bonifacino, J.S., Jiang, J., et al. (2020). Structure of Human ATG9A, the Only Transmembrane Protein of the Core Autophagy Machinery. Cell Rep. *31*, 107837.

83. Crawshaw, S.G., Cross, B.C.S., Wilson, C.M., and High, S. (2007). The oligomeric state of Derlin-1 is modulated by endoplasmic reticulum stress. Mol. Membr. Biol. *24*, 113–120.

84. Wu, X., Siggel, M., Ovchinnikov, S., Mi, W., Svetlov, V., Nudler, E., Liao, M., Hummer, G., and Rapoport, T.A. (2020). Structural basis of ER-associated protein degradation mediated by the Hrd1 ubiquitin ligase complex. Science *368*, eaaz2449.

85. Dey, S., and Levy, E.D. (2018). Inferring and Using Protein Quaternary Structure Information from Crystallographic Data. In Protein Complex Assembly: Methods and Protocols, J.A. Marsh, ed. (Springer), pp. 357–375.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
E. coli BL21 (DE3)	NEB	C2527H
Chemicals, peptides, and recombinant proteins		
Potassium chloride	Sigma-Aldrich	P3911
IPTG	VWR International Gmbh	A1008.0050
Ampicillin	Chemie Brunschwig AG	FIBBP1760-25
L-arginine	Roth Ag	1689.3
HEPES	Chemie Brunschwig AG	FIBBP310-1
Magnesium chloride	Sigma-Aldrich	208337
PEG 8000	Chemie Brunschwig AG	FIBBP233-1
PEG 1000	Sigma-Aldrich	1546489
Sodium chloride	Chemie Brunschwig AG	FSHS/3105/70
Deposited data		
Atomic coordinates and structure factors of the tRNA pseudouridine synthase A homodimer (Q8U2C1)	This study	PDB: 8Q70
Atomic coordinates and cryoEM map of the cysteine tRNA ligase homodimer (Q8U227)	This study	PDB: 8QHP EMDB: 18415
Atomic coordinates and cryoEM map of the uncharacterized Q8U0N8 protein from *Pyrococcus furiosus*	This study	PDB: 8P49 EMDB: 17402
PDB files of the models of the dataset provided by this study	This study	https://figshare.com/s/af3c1d5969f7468f2caa
PDB files of the models of the dataset provided by this study	This study[53]	https://www.ebi.ac.uk/pdbe/pdbe-kb/3dbeacons/
Software and algorithms		
gnomeAD V2.1.1 lifted to GRCh38	Karczewski et al.[54]	https://gnomad.broadinstitute.org/
Clinvar 2023-01-15	NIH	https://www.ncbi.nlm.nih.gov/clinvar/
autoPROC	Vonrhein et al.[55]	https://www.globalphasing.com/autoproc/
XDS	Kabsch[56]	http://xds.mpimf-heidelberg.mpg.de/
Phenix	Liebschner et al.[57]	https://phenix-online.org/
Coot	Emsley et al.[58]	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/
ChimeraX	Pettersen et al.[59]	https://www.cgl.ucsf.edu/chimerax/
MolProbity	Williams et al.[60]	https://github.com/rlabduke/MolProbity
AlphaFold2	Jumper et al.[8]	https://www.deepmind.com/research/highlighted-research/alphafold
MM-align	Mukherjee and Zhang[61]	https://zhanggroup.org/MM-align/
TM-align	Zhang and Skolnick[62]	https://zhanggroup.org/TM-align/
MMseqs2	Steinegger and Söding[63]	https://github.com/soedinglab/MMseqs2
Kpax	Ritchie et al.[64]	https://kpax.loria.fr/
ColabFold	Mirdita et al.[65]	https://github.com/sokrypton/ColabFold
QSproteome protocol on GitHub	This study	https://github.com/HugoSchweke/QSproteome_protocol
QSproteome protocol on Zenodo	This study	https://zenodo.org/records/10450934

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
AlphaFold2 bigbang	This study	https://colab.research.google.com/github/sokrypton/ColabDesign/blob/gamma/af/examples/predict_bb.ipynb
Other		
hu.MAP 2.0	Drew et al.[66]	http://humap2.proteincomplexes.org/
CYC2008	Pu et al.[67]	https://wodaklab.org/cyc2008/
CORUM	Ruepp et al.[68]	http://mips.helmholtz-muenchen.de/corum/
Complex Portal	Meldal et al.[69]	https://www.ebi.ac.uk/complexportal/home
YHTP2008	Pu et al.[67]	https://wodaklab.org/cyc2008/
YeastCyc	Karp et al.[70]	https://yeast.biocyc.org/
EcoCyc	Keseler et al.[71]	https://ecocyc.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Emmanuel Levy (emmanuel.levy@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Experimental structures have been deposited in the PDB and are publicly available as of the date of publication. All predicted structures are available on FigShare and 3DBeacons[53], as listed in the Key resources table.
- All original code has been deposited on GitHub or Google Colab. It is public and a snapshot is available at Zenodo, as described in the Key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the Lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

For protein expression we used *E. coli* BL21 DE3 cells and expressed overnight with 0.5 mM IPTG at 18°C in LB medium supplemented with 50 μg/ml ampicillin.

METHOD DETAILS

Generating homodimer models

The proteome sequences were downloaded from Uniprot for *P. furiosus* (proteome id: UP000001013), *E. coli* strain O157:H7 (proteome id UP000000558), *S. cerevisiae* (proteome id: UP000002311, reviewed entries) and *H. sapiens* (proteome id UP000005640, reviewed entries). The sequences of the proteomes were submitted to AlphaFold2 using a local implementation of ColabFold.[65] MSAs were generated with the MMseqs2 online service[63] and saved locally for later reuse. Sequences longer than 1200 amino acids were not processed as the effective length of dimers for such proteins exceeds 2400 amino acids, for which predictions became too GPU and memory expensive. Additionally, some sequences with a size below this cut-off were not predicted when their corresponding jobs repeatedly failed to run successfully. For all other predictions, five models were generated for each protein, and information relating to these structures was stored in a MySQL relational database. The same protocol and saved MSAs were used to generate monomeric structural models of the *H. sapiens* proteome.

Protein structure processing

Intra- and intermolecular residue contacts were identified using the same definitions as in 3DComplex.[72] Contacts between atoms were recorded when their center of mass was closer than their Van Der Waals radii plus 0.5 Å, and a contact between two residues was counted when they showed at least one atom pair in contact. We quantified clashes in two ways: the first was atom pairs closer than 2 Å. In addition, we calculated a clashscore using the Phenix implementation of MolProbity.[60] Solvent accessibility and interface sizes were determined using FreeSASA.[73] The residues stabilized in the dimer structure (Figure 2E) are defined as those exhibiting at least two inter-chain contacts and at most one intra-chain contact. All residues were considered for inter-chain contacts, and residues from *i*-4 to *i*+4 were excluded in calculating intra-chain contacts of residue *i*.

Structural models are computed on full-length protein sequences. As a result, they often contain flexible, disordered regions. Several analyses require comparing structures, and the presence of such flexible regions or domains can mask an otherwise structurally conserved core. We used a three steps procedure (Figure S1A) to trim these regions and define the "core structure": (i) we discarded residues with a predicted local distance difference test (pLDDT) score below 40; (ii) of the remaining residues, a median pLDDT score was computed, and residues with a pLDDT score below 75 and below the median value are discarded; (iii) we applied single linkage clustering on the contact matrix of the remaining residues and retained the largest cluster, thus eliminating disconnected structural parts. Unless stated otherwise, subsequent analyses are applied on the core structures of the models.

Selecting a representative model per protein

Out of the five models generated for each protein sequence, we selected one representative per protein. We performed a pairwise structural superposition of the models' structure using Kpax.[64] Structures were matched when they showed a TM-score above 0.75. We note that these matches are not necessarily reciprocal because structures can have different lengths after trimming flexible regions. Within a group, the structure with the highest dimer probability (pae4-con3 metric, defined later) was kept as a reference. Structures showing more than 10% of atomic clashes at the interface, or more than 10% of residues with clashes were discarded (clashes here are defined based on atoms whose centers of mass are closer than 2 Å). If all models in a group exceeded these clash thresholds, the selected representative structure was the one with the fewest clashes at the interface.

Evaluating the confidence of homodimer models

First, we assembled a dataset to assess the power of specific metrics returned by AlphaFold2 (e.g., inter-chain PAE) or computed on the models (e.g., size of the interface) to discriminate between monomers and dimers. The dataset was derived from the PDB[17] and consisted of 349 monomers and 77 homodimers non-redundant at a level of 30% sequence identity and elucidated by X-ray crystallography. The structures of this benchmark dataset were deposited after May 2018 and thus were not part of the AlphaFold2 training set, and no structure prior to that date showed >35 % identity and >50% overlap. As crystallographic structures can contain fortuitous interactions mediating the crystal assembly, we excluded low-confidence monomers and homodimers (error probability >25%), as evaluated by the QSBIO resource.[50] We computed AlphaFold2 predictions for these structures, using the corresponding Uniprot ID and sequence. Since our proteome predictions are performed on full-length proteins, we also employed the full-length proteins in this benchmark. We derived several metrics on the predicted model structures, and all were highly informative for discriminating between correct homodimer models (TM-score > 0.8 with experimental structure) and monomers in the dataset (Figures 1A and S1B).

We used the following metrics individually:

(i) The number of residue-residue contacts in the core structures (metric identified as "*con3*");
(ii) The mean inter-chain PAE values of interface residues in the core structure (metric identified as "*pae3*");
(iii) The mean PAE values of contact (i.e., pixels) in the core structures (metric identified as "*pae4*");
(iv) The consistency of the five models as determined from the structural superposition of their core structure before the single linkage clustering was applied. Each model was assigned a score ranging from 0 (where no other model was structurally similar) to 8 when a model reciprocally matched all the other four models (metric identified as "*repre2*"). The name in parentheses is used to refer to each metric later in the text.

We converted these four metrics into probabilities by fitting a logistic regression model based on the benchmark data, and also added the combination of the *pae4* and *con3* metrics, henceforth *pae4con3*. The logistic regression models subsequently enabled scoring any dimer model to yield a probability ranging between 0 (likely monomer) and 1 (likely dimer).

Five probabilities were calculated for the 156,065 models. As each metric showed a strong discriminatory power, we were inclusive and initially retained all models with any of the five probabilities above 0.5. Additionally, to maximize coverage, models showing a probability below 0.5 but sharing homology in quaternary structure geometry with an experimental structure (TM-score>0.7) were retained (representing 567 models in the final set), because such conservation was shown to reliably indicate physiological relevance.[50] In the final dataset we did not consider structures where the number of interface clashes represented more than 10% of the contacts, where the number of residues involving clashes represented more than 5% of the protein length, or where the number of contacts in the core structure or full-length protein was less than 10 or 30 respectively.

Quaternary structure searches against the PDB

The sequences of the four reference proteomes were searched against known structures released up until March 1st 2022. We used FASTA[74] as well as HHblits[75] to perform pairwise alignments and stored information about the hits, their corresponding sequence identity, and sequence coverage in two separate MySQL tables. Each model was assigned to one of three categories of sequence identity: 0-30%, 30-70%, and 70-100%. We imposed an overlap of at least 50% of the core structure to accept a match. We estimate the increase in structural coverage of a proteome as the number of proteins in the first category (0-30%) divided by the number in the third (70-100%).

Clustering quaternary structures

The core structure of each dimer model from the datasets was structurally superposed with KPAX[64] to all other models sharing one or more PFAM domains.[33] A binary matrix was derived where all structure pairs showing a TM-score above 0.7 were set to 1, and the rest to 0. Single-linkage clustering was applied to this matrix, yielding 2991 quaternary structure distinct clusters or "types".

Defining novel quaternary structure types

Dimer models were superposed to all putative homologs identified with FASTA[74] and HHblits.[75] We employed two superposition methods: KPAX[64] and MMalign[61] and kept the highest TM-score calculated on the smallest chain and based on the dimer model's core structure. We initially excluded as novel any dimer model for which a TM-score higher than 0.65 was identified. We note that this cut-off has to be different from the threshold classically used for monomers (0.5) because the same protein forming two distinct dimer structures will show a minimum TM-score of 0.5. We previously found the value of 0.65 to be optimal.[50] Additionally, we imposed that at least 50% of residues in each chain of the dimer core structure should be matched. Finally, we only considered a quaternary structure type as novel when all those in the same cluster were novel themselves (see section "clustering quaternary structures from the reference set"). Thus, a novel quaternary structure type was inferred for a model when we did not identify an experimentally determined structure where subunits interacted with a similar geometry. All the searches and structural alignments were carried out using structures from 3DComplex updated to March 1st 2021. Importantly, the current computational cost of quaternary structure superposition makes it difficult to run multi-chain PDB-wide structural searches. As a result, the homologs we found were prefiltered based on sequence. This approach can lack sensitivity and can be confounded by specific thresholds (e.g., on sequence coverage). While we expect that a majority of these new quaternary structure types do not exist in the PDB, a number of homologs may have not been detected. Future endeavors to develop multi-chain superposition applicable to vast amounts of data will therefore be needed to refine the numbers presented in this work.

Symmetry analysis and reconstruction

We used AnAnaS[27] to identify symmetries in the homodimer models. AnAnaS uses an analytical approach to detect axes of rotational symmetry in protein complexes to infer their symmetry group and type. An RMSD of 0 indicates a perfectly symmetric complex, while a higher RMSD reflects imperfect symmetry either due to chains' position relative to each other, or due to local structural changes between chains. Importantly, AnAnaS can infer symmetry axes from partial complexes and we used this feature to infer C3 and higher-order cyclic symmetries from homodimer models. The symmetry detection workflow is performed in several steps. First, we detected C2 symmetries and assigned it to models with a RMSD (as calculated by AnAnaS) below 4Å and a clash score below 200. When a C2 symmetry was not detected, we searched for higher-order symmetries, from C3 to C12. The symmetry with the lowest RMSD was then retained, provided it showed a RMSD value below 4 Å for C3, 3.5 Å for C4, 3 Å for C5, and 2.5 Å for C6-C12, all with a clash score below 200. If the best symmetry detected was C12, higher symmetries were searched further, from C13 to C24. The symmetry with the best RMSD was retained, providing it had a better RMSD than the C12 symmetry, and an RMSD below 2 Å for C13-14, and 1.5 Å for C15-23 with a clash score below 200.

At this stage, structures with no detected symmetry encompass monomers, dimers where both subunits are flexible and where a symmetry axis cannot be reliably defined, and proteins that form infinite assemblies, such as actin filaments.

To distinguish between those three types, we relied on the symmetry of the contact matrix that enables distinguishing homotypic from heterotypic interactions. We listed all residue pairs exhibiting more than one atom in contact (e.g., number 10 in chain A with number 40 in chain B), and recorded the fraction of those showing a reciprocal contact (i.e., number 40 in chain A with number 10 in chain B), including reciprocal pairs with a single atomic contact. If all residues in contact are reciprocal, the interface is necessarily homotypic. However, C2 symmetry may not be detected due to structural flexibility. Conversely, a homotypic score of 0 means the dimer is compatible with translational, helical or cyclic symmetry. Thus, we combined this residue-based symmetry information with global symmetry information and defined three additional categories:

- *Trans*: dimers that show no point-group symmetry (identified with AnAnaS) and a homotypic score of 0, which implies the formation of filaments with translational or helical symmetry.
- *C2-flex*: dimers with no global symmetry but pronounced local symmetry (homotypic score >= 0.4). These typically result from flexible structures with local C2 symmetry axes that are not aligned globally across the protein.
- *Ambiguous*: dimers with no global symmetry and limited local symmetry (homotypic score >0 and <0.4). In these structures, different regions can display incompatible symmetries, e.g., one domain exhibiting C2 symmetry and another adopting a translational symmetry.

Comparing monomer and homodimer models

To assess the structural similarity between the monomers and dimers predicted by AlphaFold2, the core structure of the five monomeric models was superposed onto both subunits of the corresponding dimer core structure from the reference set using TMalign.[62] The core structure of the dimer was trimmed to remove residues not present in the core structure of the monomer (in order to not artificially decrease the TM score). The TM-scores were normalized by the length of the chain of the dimer and the final TM-score was the highest of the 10 scores.

Comparing the models against cryo-EM data

We selected 23 structures of human proteins solved by electron microscopy, released after May 2018, and for which no structure released prior to that date shared >35% sequence identity with a sequence coverage of 50% or more. After manual curation, three structures were discarded from the set because the quaternary structure state was ambiguous (detailed in Figure S5 and Table S3). The core models were further trimmed to remove residues not present in the EM structure and were superposed using MMalign[61] or TMalign[62] to superpose monomers.

Detection of coiled coil domains

Coiled coils were detected structurally using the software SOCKET2,[40] with a packing cutoff of 7 Å. This identifies the knobs-into-holes packing signature of coiled coil structures rather than sequence-based signatures such as heptad repeats.[39] Coiled coils were classified into two categories: intra and intermolecular coiled coils, which are defined as those involving residues belonging to the same or different chains, respectively.

Generation of final models

All dimer models were relaxed using openMM v.7.3.1[76] in a constrained AMBER force field,[76] identical to the protocol previously published with AlphaFold2.[8] We provide the dimer structures of all models, with both original and relaxed coordinates. Ring complexes were generated with AnAnaS[27] based on the dimer core-structure. Importantly, flexible segments absent from the core-structure could occupy the space of a ring subunit adjacent to the dimer, leading to extensive steric clashes in the full-length symmetry-based reconstruction. Thus, to include the flexible segments while eliminating steric clashes, the assembly had to be repredicted in the context of the complete ring. However, such predictions are very demanding both GPU and memory-wise, and we also noticed that AlphaFold2 converged less efficiently for large complexes. For these reasons, we developed a protocol tailored to reconstruct large assemblies given a known template (ring or filament) with missing segments. First, the AnAnaS[27] reconstructed model containing only amino-acids defined in the core-structure was used as a template input to AlphaFold2 to facilitate its convergence while avoiding the need for memory heavy multiple sequence alignments. To give the AlphaFold2 predictions more flexibility, we masked all sequence information in the template by changing the sequence to all gap-tokens and masking out all sidechain atoms except for C-beta, given that the C-beta distance matrix is one of the inputs. In the case of glycine, a virtual C-beta atom was added to the template. However, the sole use of a template was often not sufficient to enable AlphaFold2 to predict the correct structure, and we overcame this issue by initializing the coordinates using the template. We call this protocol “big bang” initialization, which contrasts with the default where coordinates are initialized at zero. We therefore used the template to initialize the backbone coordinates for the first iteration in the structure module. In the case of discontinuous segments in the template, we initialized the coordinates as an interpolation between the start and end residue of the missing segment. After this process, the final models were aligned with the initial ring (predicted based on core-dimer symmetry) using MMalign[61] to ensure they were similar. Only four structures out of 955 displayed a TM-score < 0.5 and these were discarded from the final set.

Analysis of SNPs

Protein sequences were aligned to genomic locations using Ensembl Variant Predictor (VEP).[77] Human SNP data were obtained from gnomeAD v2.1.1 mapped to GRCh38 and matched to the protein sequence.[54] Information on disease-associated SNPs was extracted from ClinVar (version 20230115).[78] We counted as disease SNPs those that contained the strings ‘pathogen’ or ‘risk’ in their ClinVar description. Benign SNPs were those containing the string ‘benign’. Other descriptions were discarded.

We analyzed the frequency of SNPs and pathogenic SNPs in different protein structural regions defined by Levy.[43] The SNP frequency in a region was calculated per protein as the number of SNPs in the region divided by the number of residues. The distributions of these frequencies per region are described in Figure S7. Figure 5C summarizes the median of each distribution along with a 95% confidence interval estimated from 10,000 bootstraps. When comparing regions we computed p-values under a null hypothesis where medians were equal. The p-values were inferred from the bootstrap data, as the fraction of iterations where the median value for the interface rim|core|support region was greater than the median for the surface|surface|interior respectively (Table S4). For disease-associated SNPs we followed the same process, except that we focused on the mean of the distributions because they included mostly discrete values (0/1) due to the small numbers of pathogenic and benign SNPs. The distribution of pathogenic SNP frequencies is shown in Figure S7B, and Figure 5C summarizes the mean of the distributions along with a 95% confidence interval.

Prevalence of symmetry analysis

Analyzing the prevalence of symmetry at the complexome level required comprehensive information on protein complexes. We therefore identified protein complexes using several databases:

- *H. sapiens* complexes were collected from Humap2[66] using high-confidence entries (confidence score <= 3), CORUM[68] and Complex Portal.[69]
- *S. cerevisiae* complexes were collected from YeastCyc,[70] CYC2008,[67] YHTP2008[67] and Complex Portal.[69]
- *E. coli* complexes were collected from EcoCyc[71,79] and Complex Portal.[69]
- *P. furiosus* was not analyzed due to the lack of information on complexes.

For each organism, we concatenated information from the databases and then removed redundancies. Two complexes sharing at least 80% of subunits (identified through UNIPROT identifiers) were considered redundant, and only the largest was kept. These filters resulted in 3489, 590 and 265 non-redundant complexes for *H. sapiens*, *S. cerevisiae* and *E. coli* involving respectively 6217, 2149, and 759 unique proteins. Combining these data with our datasets gave 6234, 1445, and 2167 homo or hetero complexes corresponding to 8511, 2957 and 2665 proteins, respectively.

Next, we categorized each complex according to its composition in homo-oligomer-forming proteins, and based on the presence of paralogous chains. Two proteins were considered paralogs if they showed any degree of sequence similarity and an alignment overlap of at least 60%, or shared the same PFAM domain architecture. Based on this information four categories were defined:

- Symmetric homomers contain a single subunit type.
- Symmetric heterocomplexes contain over 30% of homo-oligomerizing subunits and none of them has a paralogous chain in the complex.
- Pseudo-symmetric heterocomplexes contain over 30% of homo-oligomerizing subunits and at least one of them has a paralog in the complex.
- Asymmetric heterocomplexes contain less than 30% of homo-oligomerizing subunits.

Protein purification

Nucleotide sequences encoding the proteins identified by UNIPROT IDs Q8U0N8, Q8U2C1, and Q8U227 were ordered from Twist Biosciences as synthetic genes cloned into the pET21 vector with C-terminal His$_6$-tag. The plasmids were transformed into *E. coli* BL21 DE3 cells and expressed overnight with 0.5 mM IPTG at 18 °C in LB medium supplemented with 50 μg/ml ampicillin. The pellets were resuspended and sonicated in 20 mM HEPES pH 7.5, 500 mM NaCl, 50 mM L-arginine, 10% glycerol buffer supplemented with 1 mM PMSF, and 125 μg/ml DNase. Cell lysates were clarified using ultracentrifugation and loaded on a 5 ml Ni-NTA Superflow column (QIAGEN) and washed with 20 mM HEPES pH 7.5, 500 mM NaCl, 50 mM L-arginine buffer with imidazole ranging from 10-30 mM, and subsequently eluted with 300 mM imidazole. Main protein fractions were concentrated and injected onto a Superose S6 10/300 gel filtration column (GE Healthcare) in 20 mM HEPES pH 7.5, 300 mM NaCl. Protein fractions were concentrated, flash-frozen in liquid nitrogen, and stored at -80 °C.

Molecular weight determination

Mass photometry experiments were conducted using a Refeyn TwoMP system (Refeyn Ltd., Oxford, UK) equipped with the AcquireMP and DiscoverMP software packages for data acquisition and analysis, respectively, utilizing standard settings. For the experiments, microscope coverslips of high precision, sourced from Refeyn, were utilized on a one-time basis. To maintain the droplet shape of the sample, self-adhesive silicone culture wells (Grace Bio-Labs reusable CultureWell™ gaskets) were employed. For contrast-to-mass calibration, Bovine Serum Albumin Fraction V low Heavy Metals (Millipore) oligomers with molecular weights of 66, 132, 198, and 264 kDa were employed. Prior to the measurements, protein stocks were diluted in stock buffers containing 20 mM HEPES pH 7.5 and 300 mM NaCl. Specifically, 2 μL of the protein solution was combined with 18 μL of analysis buffer, resulting in a final drop volume of 20 μL with a concentration of approximately 1 μg/mL.

Structure determination by cryo-EM

Protein concentrations ranging from 1-4 mg/ml were applied to a glow discharged 300-mesh holey carbon grid (Au 1.2/1.3 Quantifoil Micro Tools), blotted for 1.5–2.5 s at 95% humidity, 10 °C, plunge frozen in liquid ethane (Vitrobot, FEI) and stored in liquid nitrogen. Data collection (Table S6) was performed on a 300 kV FEI Titan Krios G4 microscope equipped with a FEI Falcon IV detector. Micrographs were recorded at a calibrated magnification of 168,674× with a pixel size of 0.83 Å and a nominal defocus ranging from −0.8 μm to −2 μm for Q8U0N8 and −1.0 μm to −2.6 μm for Q8U227.

Acquired cryo-EM data was processed using cryoSPARC.[80] Gain-corrected micrographs were imported, and micrographs with a resolution estimation worse than 5 Å were discarded after patch CTF estimation. Initial particles were picked using blob picker with 100–140 Å particle size for Q8U0N8 and 90-120 for Q8U227. Particles were extracted with a box size of 360 × 360 pixels, down-sampled to 120 × 120 pixels for Q8U0N8 and 300 × 300 pixels, down-sampled to 130 × 130 pixels for Q8U227. After 2D classification, clean particles were used for *ab initio* 3D reconstruction. After several rounds of 3D classification, the class with most detailed features was re-extracted using full box size and subjected to non-uniform and local refinement to generate high-resolution reconstructions.[81] The local resolution was calculated and visualized using ChimeraX.[59]

For structure building, the *in silico* models generated in our study were split into segments and docked into density using ChimeraX. Subsequent manual model adjustment and refinement was completed using COOT.[58] Atomic model refinement was performed using Phenix.real_space_refine.[57] The quality of the refined model was assessed using MolProbity.[60] Structural figures were generated using ChimeraX. The refined atomic models and corresponding cryoEM maps were deposited under PDB: 8P49 and EMDB: 17402 for Q8U0N8 and PDB: 8QHP and EMDB: 18415 for Q8U227.

Structure determination by crystallography

The Q8U2C1 protein (13 mg/ml) was crystallized in the $P6_522$ space group using the sitting drop vapor diffusion setup at 18 °C in 0.1 M Tris 8.5, 0.2 M $MgCl_2$, 10% w/v PEG 8000, and 10% w/v PEG 1000 buffer. Crystals were cryoprotected with 20% glycerol and flash-cooled in liquid nitrogen. Diffraction data (Table S7) was collected at the beamline PXI (X06SA) of the Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland) at a temperature of 100 K. Raw data were processed and scaled with XDS.[56] Data was processed using the autoPROC package,[55] and phases were obtained by molecular replacement using the Phaser module of the Phenix package.[57] Atomic model adjustment and refinement was completed using COOT[58] and Phenix.refine. The quality of the refined model was assessed using MolProbity.[60]

QUANTIFICATION AND STATISTICAL ANALYSIS

Our analysis is mostly descriptive. It contains statistical quantities and analyses in Figure 4B, where bars represent the standard error calculated for each binary category (respectively the fraction of proteins presenting a fraction of at least 0.05, 0.1, 0.15, 0.2 and 0.25 of inter or intra coiled coils in their full-length structures). The error bars were plotted by adding and subtracting one standard error from the proportion. We also assess the significance of differences between groups in Figure 4C and use bootstrapping as detailed in the STAR Methods section.

Supplemental figures

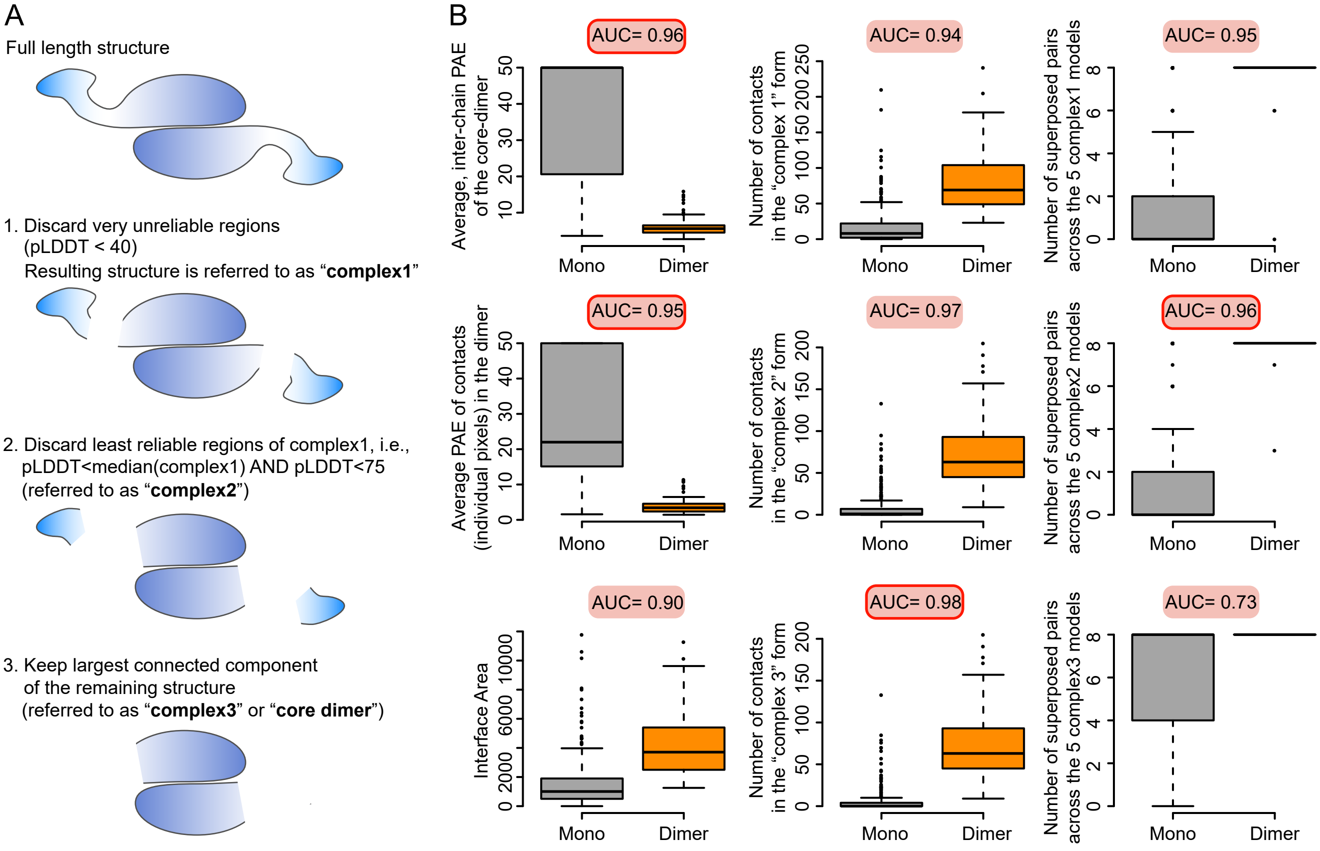


Figure S1. Structure processing to remove low-confidence regions and accuracy of several metrics to discriminate monomer and dimer states from the predicted structure models, related to Figure 1

(A) Structure models predicted by AlphaFold2 frequently contain flexible regions. When comparing structures, these regions artificially lower the structural similarity due to their conformational flexibility. Therefore, several of the analyses in this work focused on structures where these regions were truncated. Here, the procedure to truncate these regions is highlighted. First, all residues with a pLDDT value below 40 were discarded. Second, we calculated the median pLDDT score of the remaining residues and discarded those with a pLDDT below this median score, provided their value is not above 75. Third, we discarded residues and segments disconnected from the core structure, with the latter being defined as the largest connected component in the residue contact matrix.

(B) We analyzed several features for their ability to discriminate homodimers (orange) from monomers (gray) in our benchmark dataset. Several of the metrics that we tested proved reliable. Four metrics were eventually used to derive the reference sets of quaternary structures: the average interchain predicted aligned error (PAE) of the residues in contact (referred to as *pae3*); the average interchain PAE of the individual contacts (referred to as *pae4*); the number of contacts at the interface of the core structure (referred to as *con3*); and the number of superposed pairs across the five models generated by AlphaFold2, considering the truncated "complex2" structures (referred to as *repre2*).

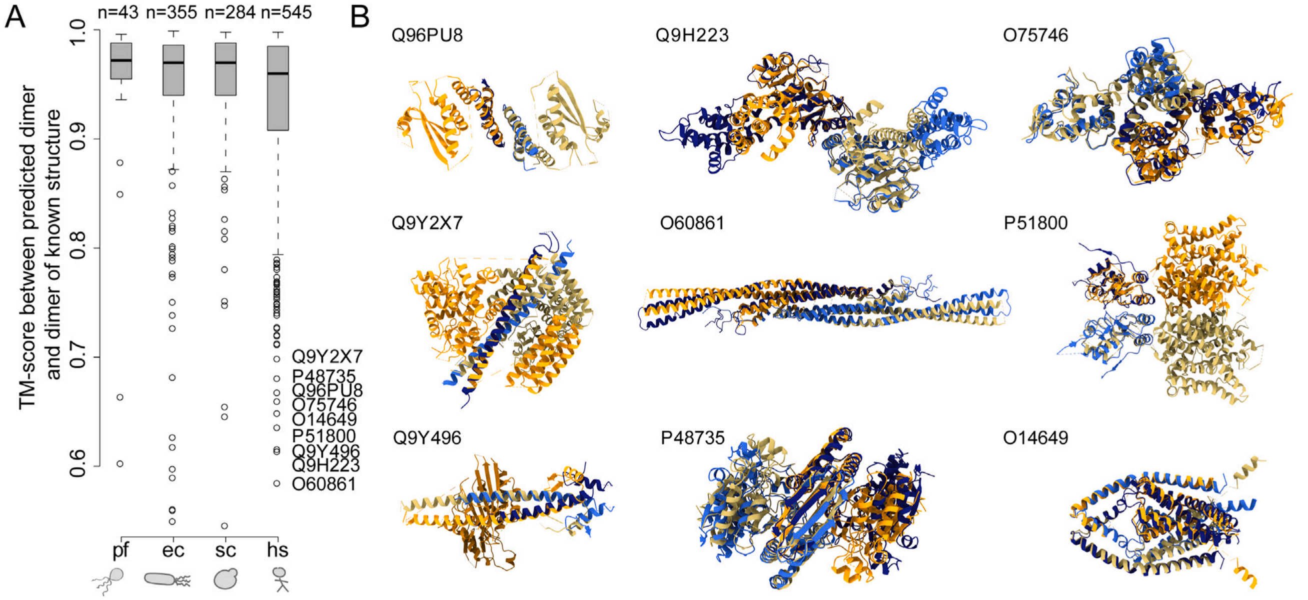


Figure S2. Dimers models show close agreement with their matching experimental structure, related to Figure 2

(A) The dimer structures from our dataset were compared with those of closely related structures in the PDB (sequence identity > 90%) for which the QSBIO error probability was below 25%.[50] The distribution of TM-scores of dimer structure pairs is shown for each species. In the human dataset, only nine structures exhibit a TM-score below 0.7, and we provide the matching UniProt identifiers.

(B) Overall, these discrepancies are not caused by incorrect predictions. Instead, they originate in the conformational flexibility of the monomers (7 cases) or in experimental artifacts (2 cases). The nine cases are depicted with the experimental structure in blue and models in orange-brown colors. Each case is detailed as follows: (1) PDB: 4DNN and Q96PU8. The low TM-score is due to 4dnn containing many selenomethionine residues, which were not considered in the structural superposition. Inspection of the two structures reveals an excellent agreement between the two interface geometries. (2) PDB: 5MTV and Q9H223. The low TM-score is due to the conformational flexibility of each monomer, but the dimeric interface is highly similar between the PDB structure and the dimer model from our reference set. (3) PDB: 4P5X and O75746. The low TM-score is due to conformational flexibility of the monomers, while the interaction geometry is similar. (4) PDB: 2W6A and Q9Y2X7. The low TM-score is due to the core structure and the X-ray structure not showing a large enough overlap. The core structure of the model misses regions from the PDB structure, but the parts present in both interact in the same manner. (5) PDB: 6IKO and O60861. The low TM-score is due to conformational flexibility of each monomer, and the dimeric interface is highly similar between the PDB structure and the dimer model. (6) PDB: 2PFI and P51800. The low TM-score is due to a poor overlap between regions seen in the PDB structure and regions present in the core structure. However, the interface region is similar. (7) PDB: 5JX1 and Q9Y496. The low structural similarity is due to 5jx1 being a chimeric protein with only a small region matching Q9Y496. (8) PDB: 4JA8 and P48735. The low TM-score is due to the conformational flexibility of each monomer, and the dimeric interface is similar between the PDB structure and the dimer model. (9) PDB: 6RV2 and O14649. Both quaternary structures are very similar, but the TM-score is low due to a poor overlap between the missing regions of the PDB structure and the missing regions of the dimer core structure.

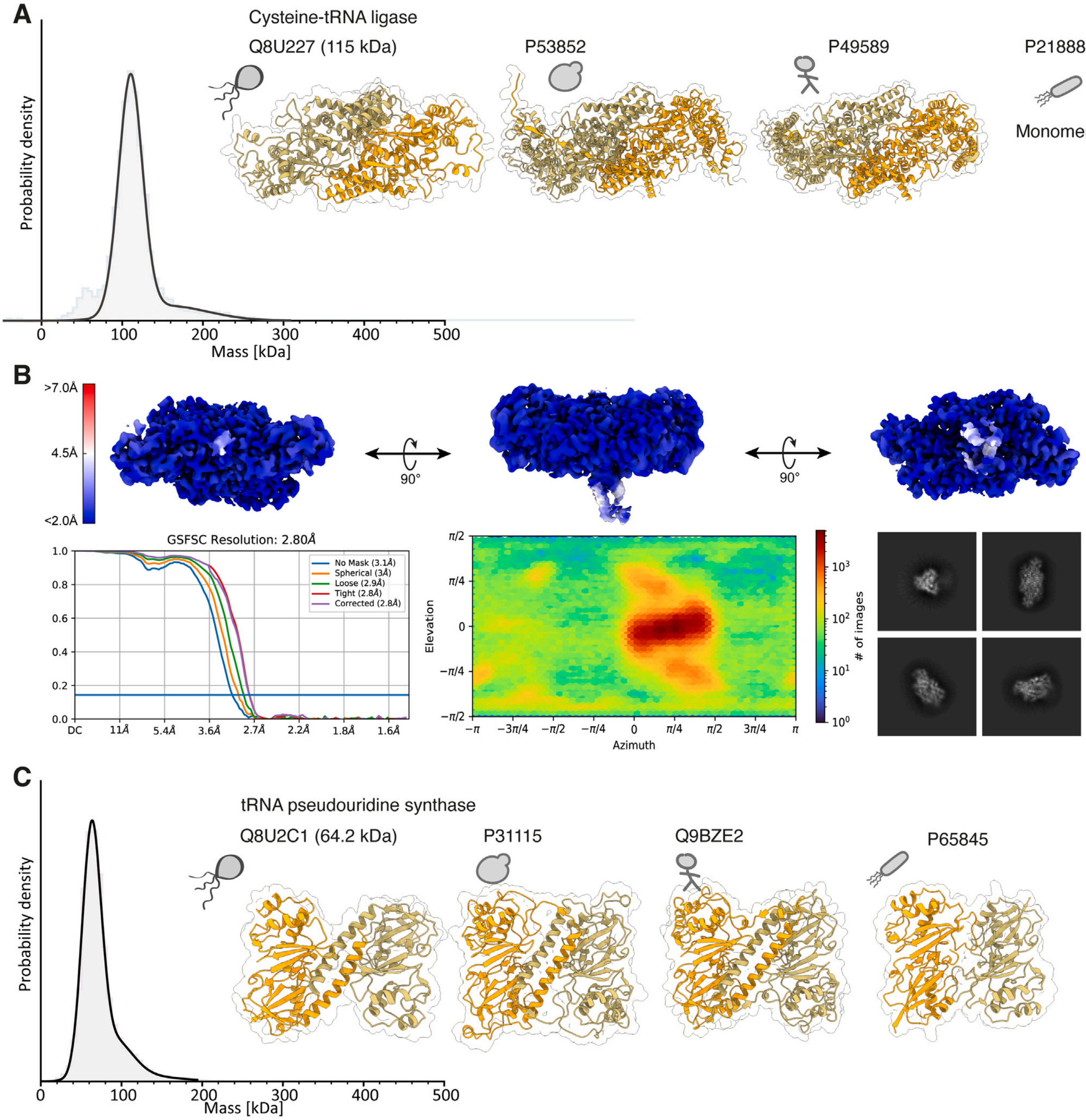


Figure S3. Experimental validation of the dimeric state corresponding to novel interface types, related to Figure 2

(A) Cysteine tRNA ligase from *P. furiosus* (Q8U227) was predicted as forming a homodimer. Similar dimers were also predicted in yeast (P53852) and human (P49589). This dimer form was absent from PDB and represents a novel quaternary structure type. The crystal structure of a homolog from *E. coli* (P21888) reflects a monomeric state. The oligomeric state of the purified protein from *P. furiosus* appears dimeric, with a peak at 115 kDa seen by mass photometry.

(B) We solved the structure of purified Q8U227 by single-particle cryo-EM and showed the details of the maps used for model building. Top row: views of the unsharpened cryo-EM density maps colored by local resolution. Bottom row: gold-standard Fourier Shell Correlation (FSC) curve with resolution cutoff indicated at 0.143, particle distribution heatmap of the final reconstruction, and example 2D class averages of different particle views.

(C) A tRNA pseudouridine synthase from *P. furiosus* (Q8U2C1) was predicted as forming a homodimer. Similar dimers were also predicted in yeast (P31115) and human (Q9BZE2). This dimer form was absent from PDB and represents a novel quaternary structure type. Interestingly, a crystal structure of a homolog from *E. coli* (P65845) also forms a homodimer. Although the interaction geometry between the two chains is similar between *P. furiosus* and *E. coli* structures, the interface is entirely different. The protein from *P. furiosus* appears dimeric, with a peak at 69 kDa seen by mass photometry.

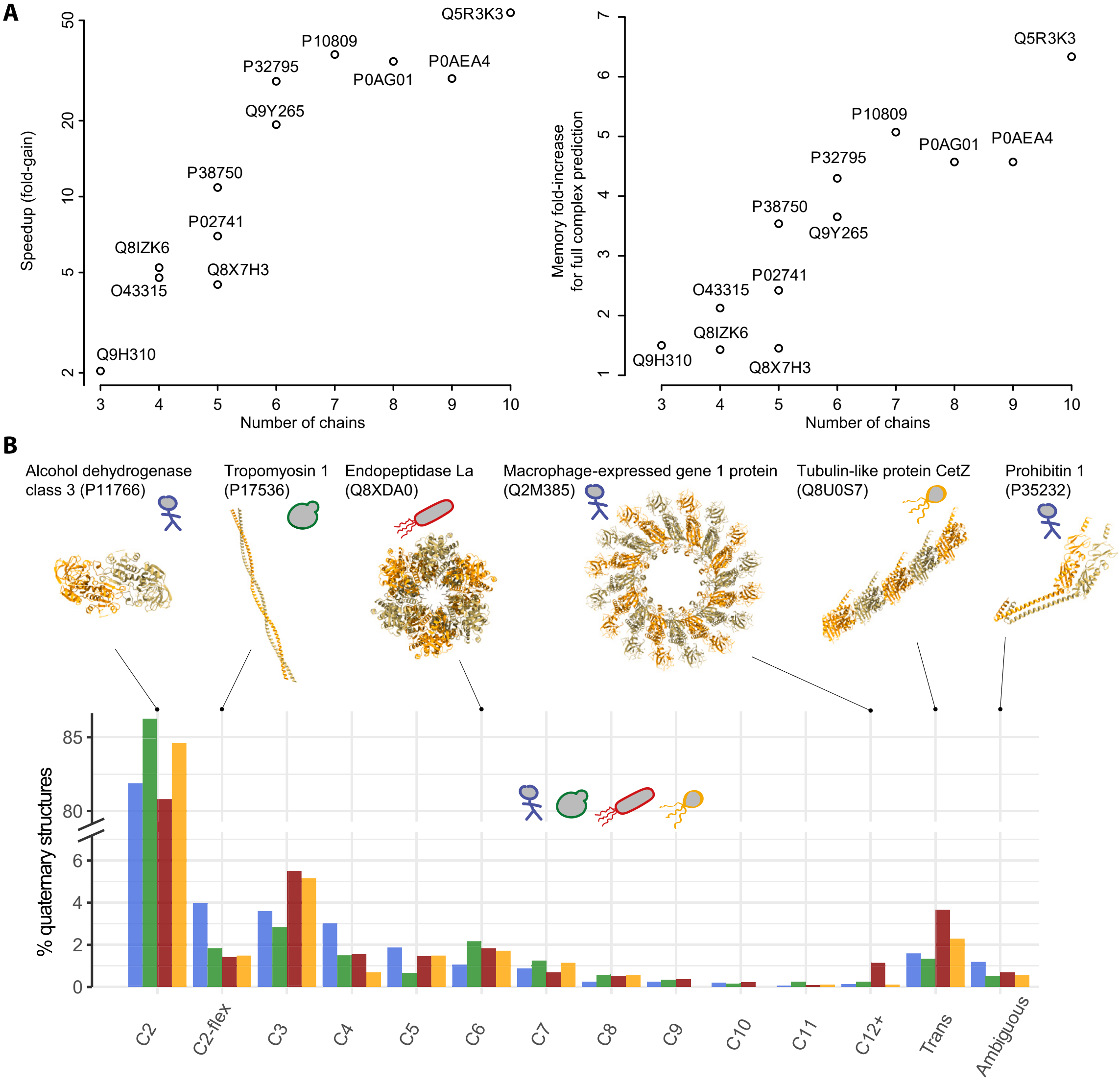
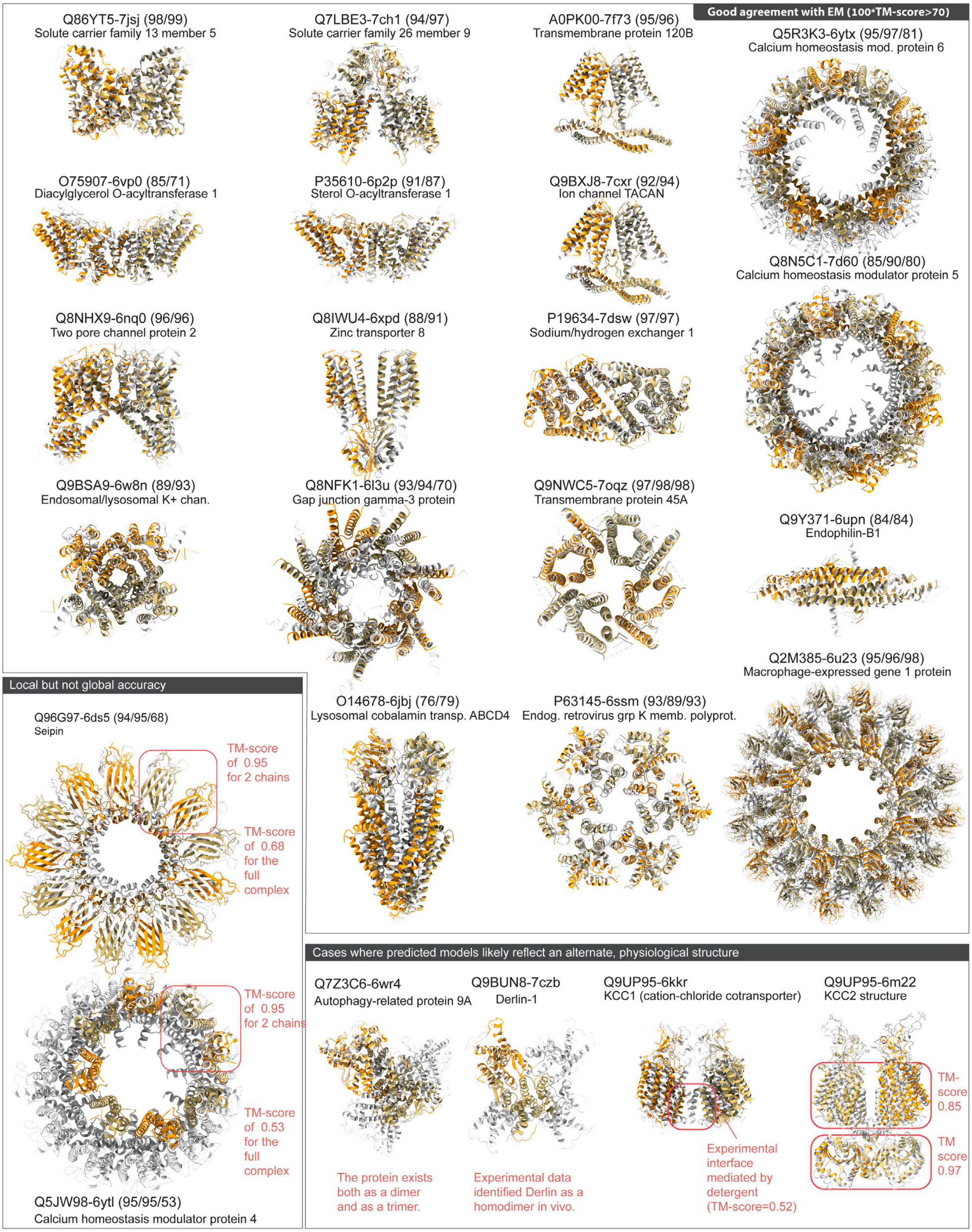


Figure S4. Time and memory requirements in predicting the structure of dimer models or the respective full-size complex and distribution of homo-oligomer symmetry across organisms, related to Figure 3

(A) Models and full complexes were predicted on the same machine and GPU (RTX6000 ADA, 48GB). The first three models were generated with three recycles each. The time and memory requirements were averaged for models 2 and 3, and the ratio of "complex/dimer" is displayed for each structure. We observe a 50-fold speedup for a complex with 10 subunits (left) and a >6-fold gain in memory requirements (right).

(B) The percentage of each symmetry type is shown. Dimers represent the largest class (note the axis break). *Cn* represents point-group cyclic symmetries and those involving 12 or more subunits were grouped in one class. The categories C2-flex, *trans*, and Ambiguous are defined as described in the STAR Methods section "Symmetry detection, assignment, and reconstruction." The alcohol dehydrogenase class 1 (UniProt ID Q2M385) is an example of a C2 complex. The tropomyosin 1 (UniProt ID P17536) is an example of a flexible C2 complex. The endopeptidase La (UniProt ID Q8XDA0) is a C6 complex. The macrophage-expressed gene 1 protein (UniProt ID Q2M385) is an example of a large cyclic complex comprising 16 subunits. The tubulin-like protein CetZ (UniProt ID Q8U0S7) is an example of a complex with translational (*trans*) symmetry and forming filaments. The prohibitin 1 (UniProt ID P35232) was classified as Ambiguous as no symmetry could be detected due to the flexible C-ter α helix forming a coiled-coil interface between the two chains of the model.


Good agreement with EM (100*TM-score>70)
Q86YT5-7jsj (98/99)
Solute carrier family 13 member 5
Q7LBE3-7ch1 (94/97)
Solute carrier family 26 member 9
A0PK00-7f73 (95/96)
Transmembrane protein 120B
Q5R3K3-6ytx (95/97/81)
Calcium homeostasis mod. protein 6
O75907-6vp0 (85/71)
Diacylglycerol O-acyltransferase 1
P35610-6p2p (91/87)
Sterol O-acyltransferase 1
Q9BXJ8-7cxr (92/94)
Ion channel TACAN
Q8N5C1-7d60 (85/90/80)
Calcium homeostasis modulator protein 5
Q8NHX9-6nq0 (96/96)
Two pore channel protein 2
Q8IWU4-6xpd (88/91)
Zinc transporter 8
P19634-7dsw (97/97)
Sodium/hydrogen exchanger 1
Q9BSA9-6w8n (89/93)
Endosomal/lysosomal K+ chan.
Q8NFK1-6l3u (93/94/70)
Gap junction gamma-3 protein
Q9NWC5-7oqz (97/98/98)
Transmembrane protein 45A
Q9Y371-6upn (84/84)
Endophilin-B1
Q2M385-6u23 (95/96/98)
Macrophage-expressed gene 1 protein
Local but not global accuracy
Q96G97-6ds5 (94/95/68)
Seipin
TM-score of 0.95 for 2 chains
TM-score of 0.68 for the full complex
O14678-6jbj (76/79)
Lysosomal cobalamin transp. ABCD4
P63145-6ssm (93/89/93)
Endog. retrovirus grp K memb. polyprot.
TM-score of 0.95 for 2 chains
TM-score of 0.53 for the full complex
Q5JW98-6ytl (95/95/53)
Calcium homeostasis modulator protein 4
Cases where predicted models likely reflect an alternate, physiological structure
Q7Z3C6-6wr4
Autophagy-related protein 9A
Q9BUN8-7czb
Derlin-1
Q9UP95-6kkr
KCC1 (cation-chloride cotransporter)
Q9UP95-6m22
KCC2 structure
TM-score 0.85
TM score 0.97
Experimental interface mediated by detergent (TM-score=0.52)
The protein exists both as a dimer and as a trimer.
Experimental data identified Derlin as a homodimer in vivo.

(legend on next page)

Figure S5. Quaternary structure models accurately recapitulate recent structures solved by electron microscopy, related to Figure 4

We identified structures of human proteins solved by electron microscopy with no close homolog in the PDB before May 2018 (>35% sequence identity). We then superposed the predicted models onto these structures. Each superposed pair is shown, with the model in orange-green and the experimental structure in white. We provide the UniProt and PDB codes above each pair along with scores (TM-score × 100). The scores represent the structural similarity of the monomer, dimer, and full complex superposition. Most models agree with the experimental structure (score > 70). Two models show excellent chain-chain interaction geometry (dimer score = 95) but poor global structure similarity due to inconsistent numbers of chains. In three cases where the prediction and experimental structure differ significantly, we highlight the ambiguous nature of the experimental data. The structure identified as PDB: 6WR4 exists both as a dimer and trimer, as noted in the original publication.[82] Interestingly, the dimer structure that we modeled involves the same interfaces as those seen in the experimentally characterized trimer. Therefore, our model may capture the alternative dimer state observed experimentally. In the second example, the structure PDB: 7CZB forms a tetramer, while our model is dimeric. Derlin-1 has been observed to form homodimers *in vivo*[83] and was also observed to be part of the hetero-oligomeric Hrd1 ubiquitin ligase complex, within which it exists as a monomer.[84] These observations imply that it can adopt multiple oligomeric states, and our model may capture the experimentally observed homodimer state. In the last example, the structure PDB: 6KKR consists of the transmembrane domain of the cation-chloride co-transporter KCC1. The structure shows a dimeric assembly mediated by detergent molecules. Our model shows a different interaction mode where dimerization is mediated by the cytosolic domain (which is absent from the structure 6KKR). Interestingly, the assembly seen in our model is similar to that of a homolog characterized more recently (KCC3, PDB: 6m22). Such similarity supports the validity of our model and suggests that the detergent-mediated interface represents an alternative interaction mode or could result from using a truncated construct. Promiscuous or non-physiological protein-protein interfaces are highly evolvable[49] and are pervasive in X-ray crystallography experiments.[50,85] The lower protein concentrations required for cryo-EM experiments make such interactions unlikely; however, certain contexts (truncations and concentration in membranes) might nevertheless promote their formation.

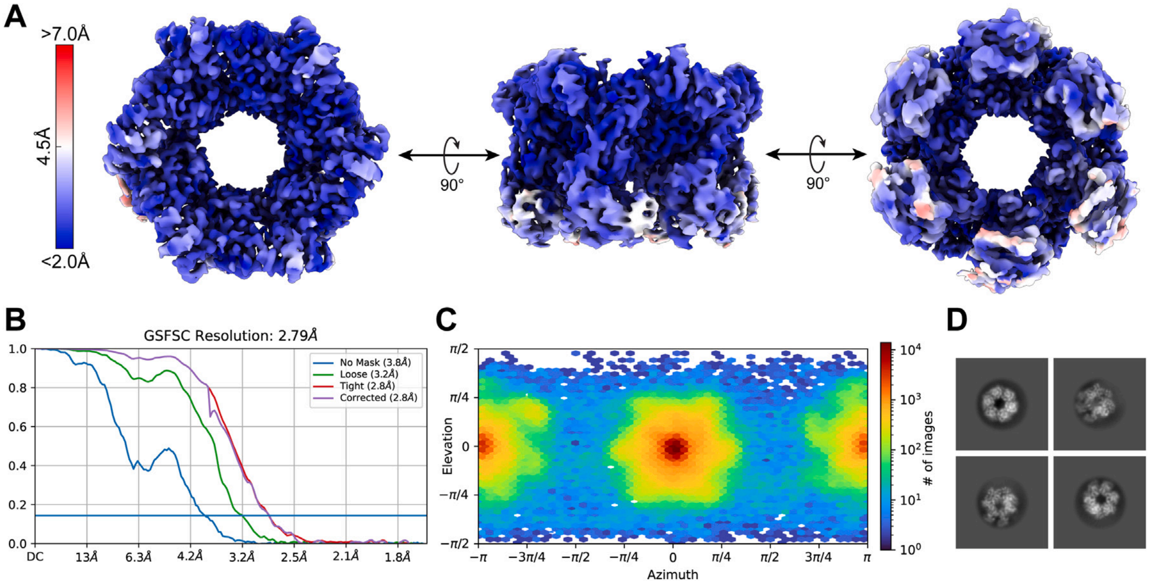


Figure S6. Cryo-EM data processing, related to Figure 4

(A) Q8U0N8 hexamer cryo-EM map used for model building. Views of the unsharpened cryo-EM density maps colored by local resolution.

(B) Gold-standard FSC curve with resolution cutoff indicated at 0.143.

(C) Particle distribution heatmap of the final reconstruction.

(D) Example 2D class averages of different particle views.

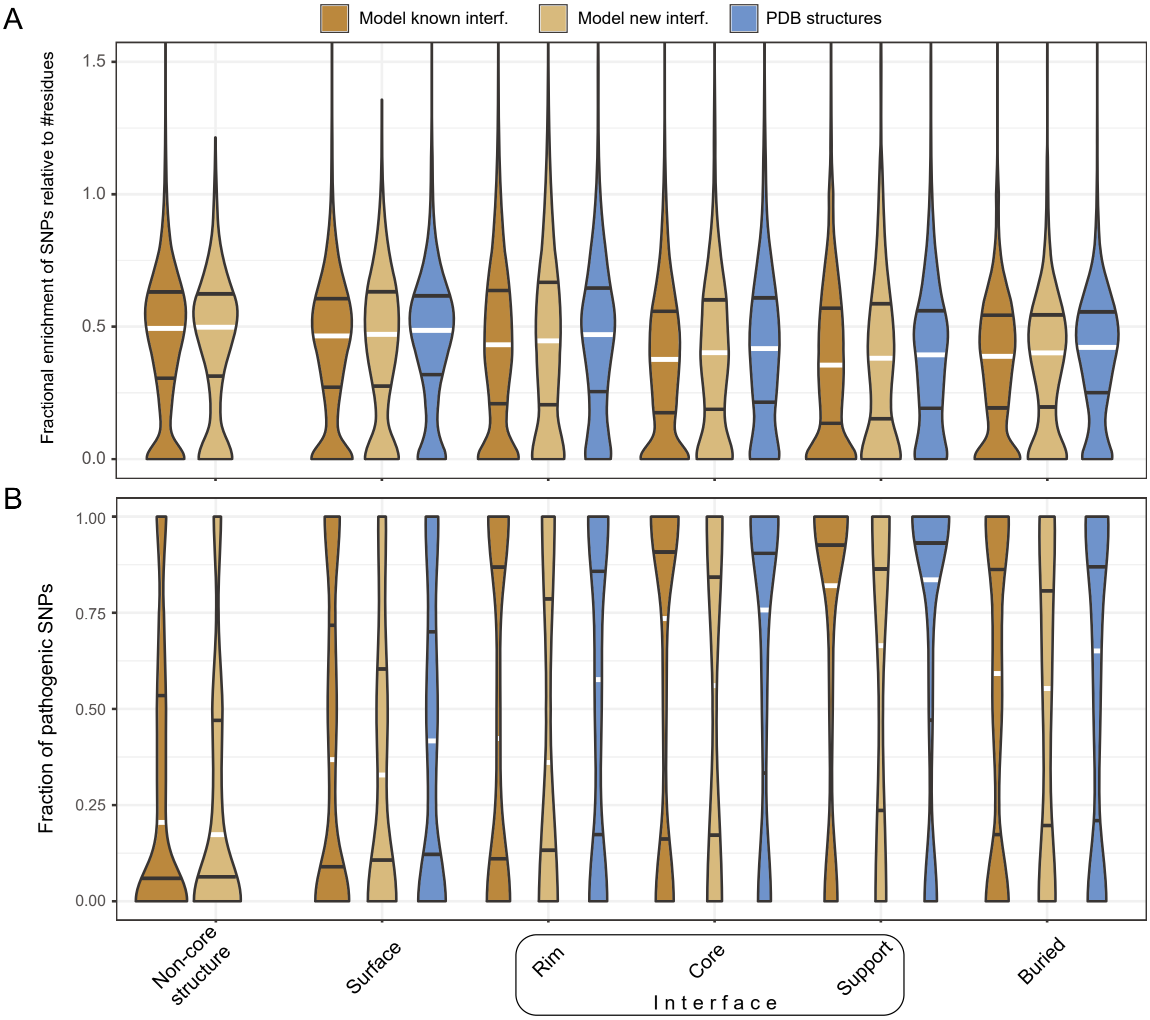


Figure S7. Distributions of SNPs and pathogenic SNPs in human proteins, related to Figure 5

(A) Violin plot depicting the fraction of SNPs[54] in each region relative to its size (number of residues). Three types of structures are compared: structures from the PDB (blue), models with quaternary structure types that were previously observed (dark brown), and those that were novel (light brown). We calculated this ratio for each type and across protein regions as defined in Levy.[43] An additional region, the "non-core structure," consists of residues absent from the core structure (STAR Methods), which are enriched in residues with low pLDDT values. We note that one residue may contain up to eight missense SNPs as different nucleotides of a codon can yield several types of missense mutations. Certain regions can therefore contain more SNPs than residues; therefore, the density extends beyond 1. We only show the range 0 to 1.5 for clarity.

(B) Violin plot showing the fraction of pathogenic SNPs in the same complex types and across the same regions. Clinical annotations are derived from ClinVar.[78] The number of pathogenic SNPs in each region is normalized by the number of benign and pathogenic SNPs (STAR Methods). White and black horizontal bars show the median and 25/75th quantiles of each distribution, respectively.