

AlphaFold Database expands to proteome-scale quaternary structures

Yewon Han^{1,2*}, Maxim I. Tsenkov^{3,*}, Niccolò A. E. Venanzi^{4,*}, Damian Bertoni³, Sooyoung Cha^{1,2}, Alejandro Chacón⁴, Nick Dietrich⁵, Boris Fomitchev⁴, Yonathan Goldtzvik³, Darren Hsu⁴, Jeannie Austin³, Joseph Ellaway³, Kieran Didi^{4,6}, Oleg Kovalevskiy⁵, Dariusz Lasecki⁵, Agata Laydon⁵, Micha Livne⁴, Paulyna Magaña³, Maciej Majewski⁷, Sreenath Nair³, Urmila Paramval³, Nilkanth Patel⁴, Risha Patel⁵, Ivanna Pidruchna³, Brianda Santini Lopez⁴, Prashant Sohani⁴, Ahsan Tanweer³, Duc Tran⁴, Kyle Tretina⁴, Melanie Vollmar³, Quan Vu⁴, Augustin Žídek⁵, Sameer Velankar^{3,#}, Martin Steinegger^{1,2,8,9,#}, Jennifer Fleming^{3,#}, Milot Mirdita^{1,10,11,#}, Christian Dallago^{4,12,13,#}

¹School of Biological Sciences, Seoul National University, Seoul, Korea. ²Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea. ³European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. ⁴NVIDIA, Santa Clara, CA, USA. ⁵Google DeepMind, London, UK. ⁶Department of Computer Science, Oxford University, UK. ⁷Independent. ⁸Artificial Intelligence Institute, Seoul National University, Seoul, Korea. ⁹Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Korea. ¹⁰Department of Precision Medicine, Sungkyunkwan University School of Medicine, Suwon, Korea. ¹¹Biomedical Institute for Convergence at Sungkyunkwan University, Sungkyunkwan University, Suwon, Korea. ¹²Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA. ¹³Department of Cell Biology, Duke University, Durham, NC, USA. *These authors contributed equally. #Correspondence to: sameer@ebi.ac.uk, martin.steinegger@snu.ac.kr, jfleming@ebi.ac.uk, milot@skku.edu, cdallago@nvidia.com

Abstract

Protein function is governed by molecular interactions, yet structural coverage of these interactions remains sparse. The AlphaFold Protein Structure Database (AFDB) transformed access to accurate monomeric protein structures at scale. Here, we expand AFDB with 1.8M *high-confidence* protein complexes by conducting a large-scale study of over 31M predicted homo- and heteromeric protein complexes compiled from 4,777 proteomes, including model- and global health organisms, and using STRING physical-interaction annotations. We calibrate confidence metrics to assess the quality of complex predictions, and propose confidence cutoffs. These enabled the discovery of emergent structure and topologies in complex structure prediction that is not present with monomeric predictions. Clustering of *high-confidence* complexes showed that the largest 1% of non-singleton representatives account for ~25% of all complexes, and that ~9% of clusters are conserved across superkingdoms. In summary, large-scale structural predictions of the interactome serve as a foundational resource to facilitate functional and mechanistic discovery across biology.

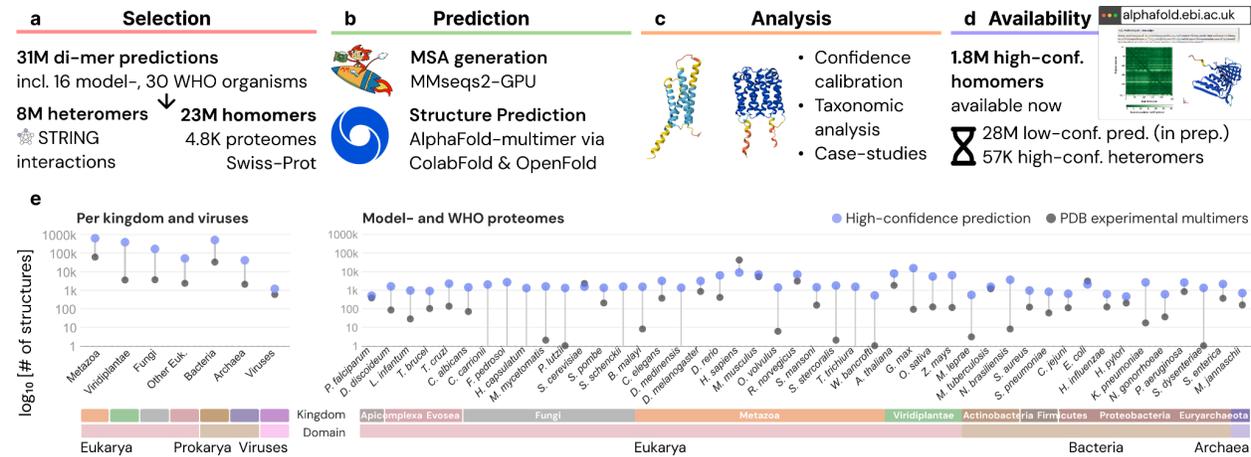


Figure 1: High-confidence protein complex 3D predictions extend the AlphaFold Database. (a) ~31 million protein homodimer and heterodimer complexes are retrieved from UniProt¹ and STRING^{1,2}. (b) MSAs are constructed using MMseqs2-GPU³, and 3D structures are predicted with AlphaFold-multimer⁴ through ColabFold⁵ or OpenFold⁶. (c) We determine reliability threshold and present biologically relevant case studies to highlight emerging behaviour in complex prediction absent from monomer predictions. (d) High-confidence complex 3D predicted structures are added to the AlphaFold Database for interactive visualisation and interrogation. (e) Number of high-confidence complexes (blue) and experimentally determined multimers per model/WHO organism (left) in the PDB (grey) per kingdom and viruses (right).

Introduction

Cellular processes are orchestrated by interacting proteins whose structures encode specificity, regulation, and function⁷. Although interaction networks have been mapped extensively through genetic, biochemical, and computational approaches², the lack of structural information for most protein-protein interactions remains a critical bottleneck to advancing biological understanding, since experimentally determined complex structures in the Protein Data Bank (PDB⁸) cover only a small fraction of known protein-protein interactions.

The advent of AlphaFold2⁹ enabled rapid and accurate prediction of protein tertiary (3D) structure. The AlphaFold Protein Structure Database (AFDB)¹⁰⁻¹² provides access to monomeric protein 3D structures across many proteomes. AFDB transformed structural biology and enabled novel basic science discoveries, for instance, the characterisation of novel protein folds¹³, as well as advancing artificial intelligence (AI) in biology, by improving protein 3D prediction methods¹⁴, or for the development of generative artificial intelligence methods¹⁵. However, biological function rarely resides in isolated monomers, and for many proteins, biologically relevant conformations, binding sites, and regulatory features only emerge upon complex formation^{16,17}. Protein-protein interaction data are available from a range of experimental datasets and are analysed using increasingly sophisticated computational methods. These datasets from diverse biological systems are curated and disseminated through multiple public resources, including the Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING²) and IntAct¹⁸. Translating this interaction knowledge into structural models has the potential to substantially increase the information content available to support mechanistic and functional research.

While comprehensive experimental structural characterisation of all protein interactions remains infeasible, methods predicting structures of protein complexes, such as RoseTTAFold¹⁹ and AlphaFold-Multimer⁴, have demonstrated that high-confidence quaternary structure prediction is possible. Several recent studies characterised proteome-scale protein complex structure predictions in order to generate novel insights into systems biology²⁰⁻²⁶. Yet, these efforts are typically limited to

specific organisms, lack consistent confidence calibration, or are not integrated into a stable public infrastructure. As a result, multimer predictions remain difficult to discover, compare, or reuse, preventing their adoption as a routine component of biological analysis, or as training and inference data for AI. Although complex prediction has not yet reached experimental accuracy and still exhibits biases^{27,28}, these models can generate strong initial hypotheses; however, large-scale dry-lab structural characterisation of protein interactions remains intractable due to experimental cost and time.

Here, we introduce a comprehensive study predicting 23,441,822 homomeric and 7,620,644 heteromeric protein complexes belonging to 4,777 proteomes, including 16 model organisms and 30 proteomes prioritised by the World Health Organisation (WHO global health proteomes; **Fig. 1a**) and Swiss-Prot. We extended the computation acceleration strategies previously applied to monomeric 3D structure prediction²⁹ to complex structure prediction, building on AlphaFold-Multimer⁴ to enable predictions at this scale. We further release 1,754,242 *high-confidence* complexes by extending the AlphaFold Database^{4,10} enabling structural interrogation of interaction networks, variant effects at interfaces, and mechanistic hypotheses that were previously inaccessible at scale, with the full set of ~31M predictions to follow, openly accessible for bulk download. Our work provides a path to improved structure prediction throughput, offers insights into complexes at scale, and provides easy access to novel data through AFDB for systems biology.

Results

We predicted the 3D structures for ~23M homodimers derived from 4,777 proteomes in UniProt^{1,4,10} and Swiss-Prot³⁰, including 16 model organisms and 30 WHO global health proteomes. Additionally, ~8M heterodimer candidate pairs were extracted from the “physical protein–protein interaction” set of the STRING Database² by filtering for proteins belonging to the 16 model organisms and 30 WHO global health proteomes (**Fig. 1a**). In contrast to recent studies^{20–22}, we decided not to add additional filtering, such as STRING score thresholds, to increase coverage for these critical proteomes. We used MMseqs2-GPU³ to generate Multiple Sequence Alignments (MSAs) for homodimers against UniRef100³¹ without the use of metagenomic databases. By restricting alignments to the best hit per taxon, we established a pragmatic orthology filter that prevents paralogous sequences from diluting the evolutionary signals essential for predicting protein complexes. For heterodimers, we concatenated previously generated homodimer MSAs, without pairing. We then used AlphaFold-multimer for complex prediction with inference through either an accelerated implementation of ColabFold⁵ or OpenFold⁶ to obtain 3D structures (**Fig. 1b**; Methods).

Once structures were predicted, in addition to model-supplied predicted Local Distance Difference Test (pLDDT) and interface predicted TM-score[s] (ipTM), we computed interaction prediction Score[s] from Aligned Errors (ipSAE)³², pDockQ2³³, and Local Interaction Score[s] (LIS), as well as the number of backbone ($\text{clash}_{\text{backbone}}$) and heavy atom ($\text{clash}_{\text{heavy-atom}}$) clashes (**Fig. 1c**). As metrics like ipSAE are directional, we computed scores for chain A relative to chain B and for chain B relative to chain A, and defined $\text{ipSAE}_{\text{min}}$ as the smaller of these two values. This reduced bidirectional metrics to a single value per complex.

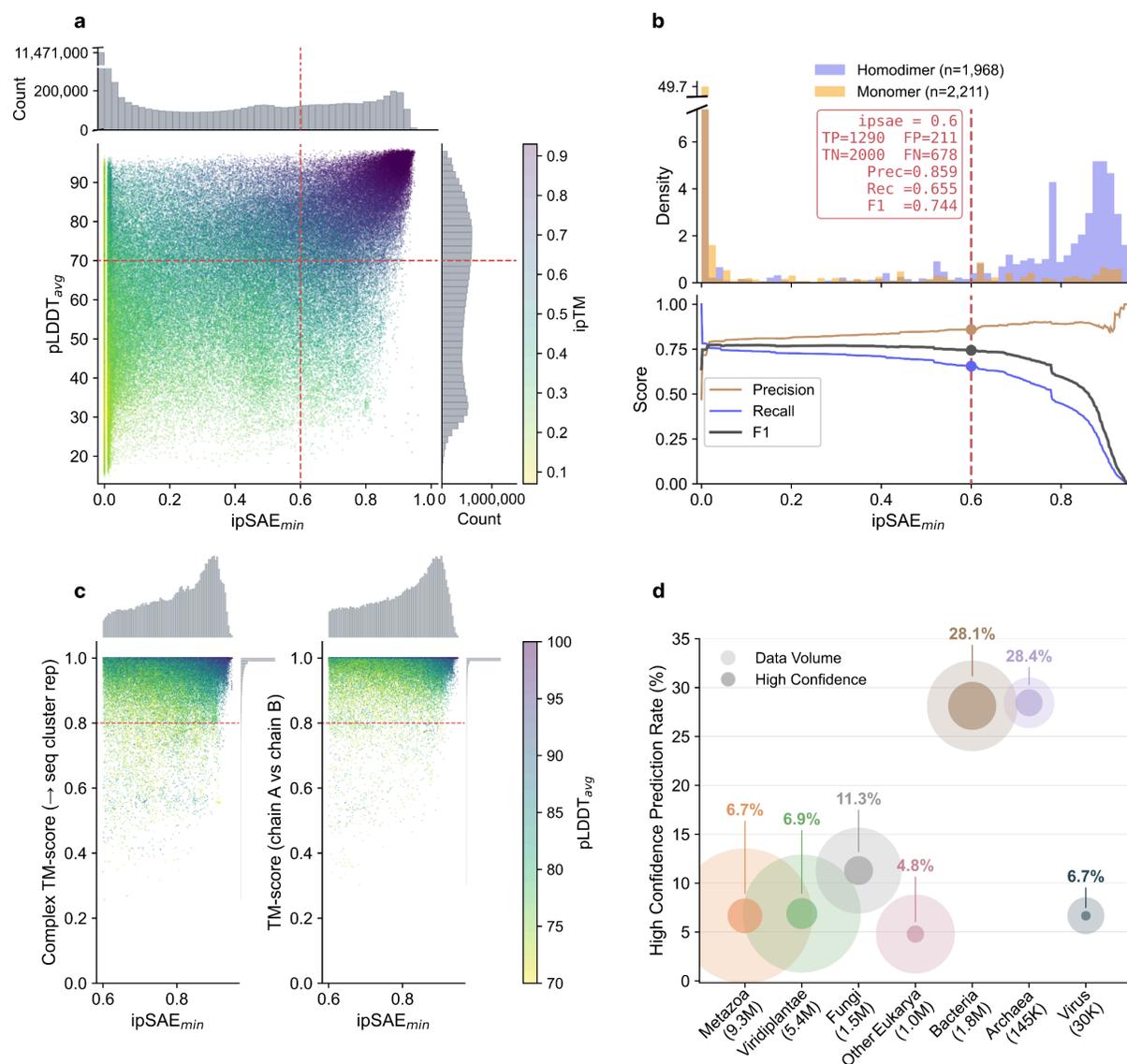


Figure 2: Homodimer analysis defines a high-confidence prediction threshold and assesses structural consistency. (a) Homodimeric predictions are plotted with ipSAE_{min} (x axis), pLDDT_{avg} (y axis) and ipTM (colour) generally displaying agreement. Red dotted lines indicate thresholds for *high-confidence* (ipSAE_{min}=0.6;pLDDT_{avg}=70). (b) ipSAE_{min} appears to be a strong discriminator for monomeric vs homodimeric complexes. (c) The homodimer prediction resulted in consistency within the group of dimers that share near identical sequences, as well as between the two chains. (d) Prediction success rate (inner circle) per major taxonomic clade among all predictions (outer circle).

Through comparison of the homodimeric structure from the PDB and corresponding predictions, the combination of ipSAE_{min} above or equal to 0.6, pLDDT_{avg} above or equal to 70, and less than or equal to 10 backbone clashes emerged as a good estimate for *high-confidence* predictions, resulting in 1.8M homodimeric complexes available through the AlphaFold Database with standardised structural formats, confidence metrics, and metadata. This enables search, visualisation, and bulk access (**Fig. 1d**).

Overall, this resource provides a massive expansion of the known structural interactome (**Fig. 1e**). The *high-confidence* predictions consistently exceed the number of experimental multimer structures in the PDB by one to three orders of magnitude across nearly every proteome, with the exception of highly studied organisms like *Homo sapiens*, *Escherichia coli* and *Saccharomyces cerevisiae*. Across broader taxonomic clades, particularly *Metazoa* and *Viridiplantae*, these *high-confidence* models bridge a vast structural gap, providing a foundation for structural biology at scale.

Filtering *high-confidence* homomeric 3D complexes across thousands of proteomes

Structures for homodimeric complexes were predicted systematically for all annotated sequences across 4,777 UniProt proteomes, including 16 model organisms and 30 WHO global health proteomes. Across the full prediction set, ipSAE_{min}, pLDDT_{avg}, and ipTM showed broad agreement, with high scoring predictions clustering in the upper right of the joint distribution (**Fig. 2a**).

To validate the efficacy of different scoring metrics, we first established a ground truth dataset by filtering experimentally determined PDB structures that had exact-sequence matches to sequences included in our prediction campaign. To reduce overlap with the AlphaFold-Multimer training set, we selected entries released after September 30th, 2021, resulting in 1,968 PDB homodimers expected to be confidently predicted as homodimers. As a negative control, we included 2,211 PDB monomers expected to score poorly in homodimer prediction (**Fig. 2b**). The distributional separation is immediately apparent: the vast majority of monomers accumulate at ipSAE_{min} ≈ 0, while homodimers extend broadly to high scores.

To determine the optimal combination of metrics and cutoffs, we assessed four common scoring metrics: ipTM, ipSAE_{min}, LIS_{min}, and pDockQ2. We analysed the score distributions for both the ground truth and negative sets, and computed precision, recall, and F1 curves across a range of cutoffs. Among the four metrics, ipSAE_{min} showed the clearest distributional separation and the most stable F1 plateau across cutoffs (**Fig. 2b**; Supplementary Fig. 1a). Based on this analysis, we selected ipSAE_{min} with a cutoff of ≥0.6, a literature-supported threshold^{32,34,35}. Our benchmark supports this choice of cutoff: the F1 curve exhibits a broad, stable plateau up to ipSAE_{min} = 0.6, after which it declines sharply, indicating that this cutoff does not meaningfully compromise discriminative performance. This cutoff yields a precision of 0.859, a recall of 0.655, and F1 of 0.744 (**Fig. 2b**). We additionally computed clash_{heavy-atom} and clash_{backbone} scores to identify structures with steric clashes, and observed that clash_{backbone} decreased with increasing ipSAE_{min} (**Supplementary Fig. 2**), supporting the decision of selecting this metric. Considering that pLDDT_{avg} of 70 and above for monomeric structures is defined as high-confidence in AFDB¹⁰, we decided to filter structures for pLDDT_{avg} ≥ 70 for improved overall complex confidence (**Fig. 2a**), which excluded about 15% of the data. Lastly, to reduce the number of structures with excessive clashes, we filtered clash_{backbone} ≤ 10, finally defining high-confidence complexes for the purposes of further analysis as those with ipSAE_{min} ≥ 0.6, pLDDT_{avg} ≥ 70, and clash_{backbone} ≤ 10, which results in 1.8M high-confidence homodimers out of ~23M (~7% retention).

To validate prediction consistency for the *high-confidence* homodimers, we compared the structures of highly homologous proteins. First, we simplified the *high-confidence* homodimer set to 1,719,470 monomers for the purpose of sequence clustering, and clustered these sequences using MMseqs2³⁶ at 0.98 identity and 0.95 coverage. This yielded 1,429,305 clusters of which 148,148 were non-singletons with an average cluster size of 2.96 members. Then, using Foldseek Multimer³⁷, we aligned each predicted dimer to its cluster representative in sequence space. This showed that 95.9% of the complexes maintained either a query- or target-normalised complex TM-score greater than 0.8 (**Fig. 2c**, left), supporting the internal coherence of our dataset. Furthermore, structural

alignment of the two chains within each predicted homodimer using Foldseek³⁸, showed that 98.81% of predictions achieved either a query-normalised or target-normalised TM-score greater than 0.8, confirming prediction consistency between the two chains.

The *high-confidence* homodimer prediction rate varied across taxonomic groups (**Fig. 2d**), with archaea and bacteria exceeding eukaryotes (fungi, plants, animals) by more than 3-fold. This likely reflects the shorter, more compact architecture of prokaryotic proteins, and the higher prevalence of homo-oligomeric assemblies in prokaryotic proteomes. Eukaryotic proteins, by contrast, tend to be longer, more multi-domain, and richer in disordered regions, and more often participate in heteromeric complexes not captured by homodimer modelling.

Finally, to assist users of the AFDB website, particularly those less familiar with predictive methods and their confidence metrics, in interpreting the predicted models, the website classifies entries into three categorical labels based on ipSAE_{\min} : “*very high-confidence*” (≥ 0.8 , high accuracy interface; 972,625 entries), “*confident*” (0.7 to < 0.8 , correct, well-resolved interaction likely; 438,879), and “*low-confidence*” (0.6 to < 0.7 , expected interaction signal, but interpret with caution; 342,738).

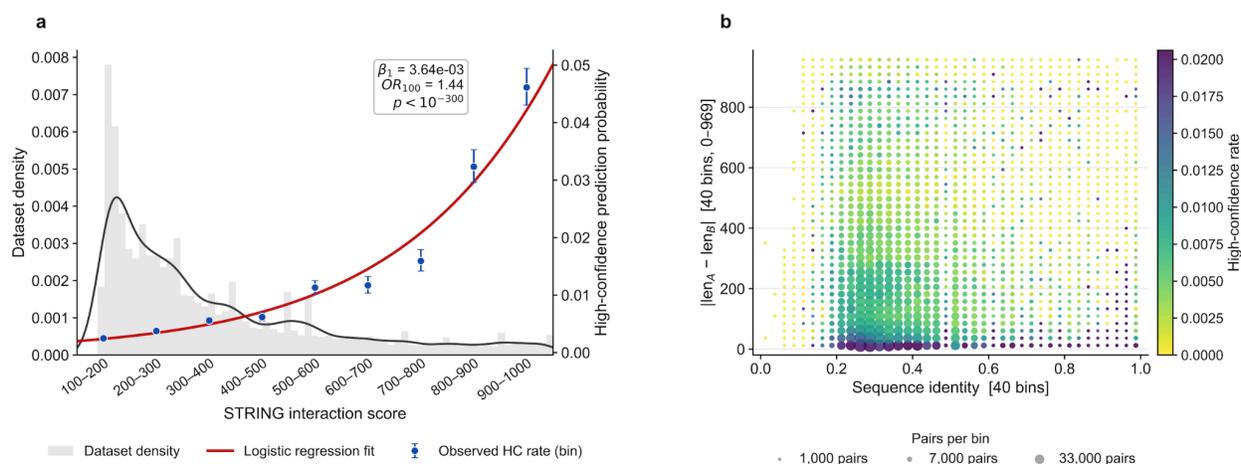


Figure 3: Heterodimeric structures correlate with STRING scores, relative length difference, and chain similarity. (a) The STRING score (x axis) increases the likelihood of *high-confidence* predictions (right-side y axis). Conversely, the number of candidates decreases with higher STRING score. (b) The *high-confidence* prediction rate as a function of inter-chain sequence identity and chain-length asymmetry in heterodimers for the pairs from the ~8M heterodimer set; sparse bins (< 50 pairs) are excluded from colour normalisation. Each circle represents a bin in a 40×40 grid partitioning sequence identity between the chains against absolute chain-length difference (range 0–966 amino acids). Colour indicates the fraction of protein pairs within the bin of which heterodimer prediction meets the high-confidence criteria ($\text{ipSAE}_{\min} \geq 0.6$, $\text{pLDDT}_{\text{avg}} \geq 70$, $\text{clash}_{\text{backbone}} \leq 10$); dot area scales with bin occupancy. On the x axis: chains with higher similarity yield higher ratio of *high-confidence* prediction. On the y axis: *high-confidence* predictions appear more frequently when the absolute length difference between chains is smaller.

Evidence-driven heterodimeric complex prediction

Heterodimeric interactions were derived from STRING enriched for biologically supported interactions (*physical interactions*). We prioritised full coverage within a defined set of high-interest proteomes (listed in *Supplementary Materials*), enabling unbiased exploration of heterodimeric structures and yielding approximately 8M candidate complexes.

We applied the same filtering criteria we derived for homodimers ($ipSAE_{\min} \geq 0.6$, $pLDDT_{\text{avg}} \geq 70$, $\text{clash}_{\text{backbone}} \leq 10$) to the 7,620,644 heterodimer set, which reduced heterodimers to 56,956 *tentatively high-confidence* predictions. For these, we observed a significant relationship ($p < 10^{-300}$) between STRING score and the likelihood of *high-confidence* (Fig. 3a). However, *high-confidence* heterodimer predictions were more frequently observed when the absolute length difference between chains is smaller (lower absolute length difference; x axis of Fig. 3b), and when the sequence identity between the two chains of heterodimer is higher (y axis in Fig. 3b). Notably, even within the lower sequence identity range (for example, 0.2 to 0.4 in Fig. 3b), the decrease in *high-confidence* prediction rate with increasing absolute length difference remained clear. Taken together, these patterns suggest that homodimer-like complexes, or complexes with fewer changes in length or identity of amino acids between chains have a higher likelihood to pass our filtering criteria, compared to heterodimers with larger differences between chains.

Given these potential caveats for our filtering criteria in the heteromeric context, we believe further confidence calibration is still required to identify *highly-confident* heterodimers across relevant axes of biological interest, rather than towards homodimer-like properties. We therefore designate these structures as *tentatively high-confidence*, and plan to further calibrate cut-offs to include a more representative set in a subsequent AlphaFold Database release.

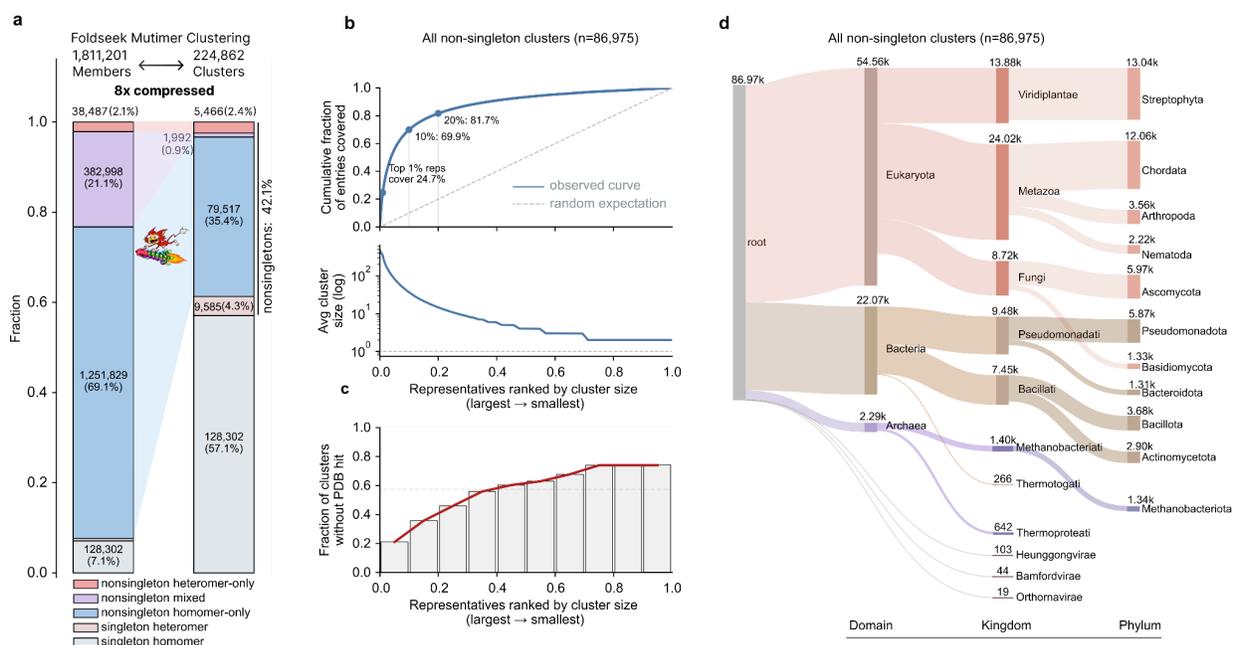


Figure 4: Compressibility of predicted complex space. (a) Clustering of 1,811,201 structures (1,754,242 *high-confidence* homodimers, and 56,959 heterodimers using the same filtering criteria as for homodimers) results in an approximately 8-fold compression of the dataset. Most entries were assigned to homodimer-only clusters, but a small fraction of non-singleton clusters were mixed, containing both homo- and heterodimeric members (b) Non-singleton cluster representatives ranked by cluster size. A relatively small number of large clusters account for most structures: the top 1% of representatives cover approximately 25% of entries, and the top 20% cover approximately 82%. The lower panel shows the cluster-type composition across representative rank bins. (c) Fraction of non-singleton clusters lacking any PDB multimer match at a Foldseek multimer TM-score threshold of 0.65, plotted across representative rank bins. Clusters without PDB support are enriched among smaller clusters. (d) Sankey³⁹ diagram summarizing the taxonomic lowest common ancestors of the 86,975 non-singleton clusters.

Clustering of the predicted complexes

High-confidence homomeric and *tentative high-confidence* heterodimeric structures (1,754,242 homodimers and 56,959 heterodimers) were clustered using Foldseek Multimercluster (manuscript in preparation) with parameter set according to structural similarity of interacting chains and their interfaces, reducing 1,811,201 structures 8-fold into 224,862 clusters, of which 86,975 contained at least one other member (non-singleton clusters; **Fig. 4a**). Homomeric complexes constitute the majority both before and after clustering (~69% and ~82% of non-singleton entries, respectively). mixed clusters containing both homo- and heterodimeric members account for only ~0.9% of non-singleton clusters and mostly occur in the largest clusters (top 10%) (**Fig. 4a**).

The top 1% of non-singleton cluster representatives cover ~25% of all entries, and 20% account for ~82% (**Fig. 4b**), indicating that predicted high-confidence complex space is concentrated around a small number of recurrent folds. Non-singleton clusters without any detectable PDB multimer match were more frequent among smaller clusters than among the largest clusters (**Fig. 4c**), suggesting that structurally recurrent solutions are more likely to overlap with known multimeric space, whereas structurally rarer clusters are more often unsupported by current PDB coverage (**Fig. 3**).

To assess evolutionary conservation, we computed the lowest common ancestor (**Fig. 4d**) for each cluster. Notably, ~9% of non-singleton clusters contain members from at least two different superkingdoms, indicating that these complexes likely originated in a common ancestor and have been maintained as universal building blocks of cellular life.

Representative exemplars and use cases

To illustrate the range of ways in which oligomeric context can affect structural interpretation, we selected representative examples drawn from eukaryotic microbes, fungal and bacterial pathogens, a neglected tropical parasite, and a crop species. These examples were chosen to capture different outcomes, including recovery of domain-swapped folds, improved membrane-protein organization, refinement of inter-domain architecture, and agreement with models derived from curated sequence input.

The transcription elongation factor Eaf (UniProt accession Q55DI5) N-terminal domain-containing protein from *Dictyostelium discoideum*, provides a clear example of a fold that emerges only in an oligomeric context. The monomeric model (AF-Q55DI5-F1) has low confidence (pLDDT_{avg}=50.56) and fragmented β -sheet elements. In contrast, modelling the protein as a homodimer (AF-0000000066503175-v1) yields a well-defined structure (pLDDT_{avg}=86.06) formed by domain swapping between the two chains (**Fig. 5a**). Each chain contributes structural elements that complete the fold of its partner, such that the architecture is assembled across chains rather than within a single polypeptide. A Foldseek search against the Protein Data Bank identified a related architecture in 7okx, in which the fold is likewise formed across chains O and M, supporting the domain-swapped assembly observed in the prediction. This case shows that some folds emerge only in an oligomeric context and may be missed or misrepresented by monomeric prediction alone.

The Autophagy-related protein 33 from *Fonsecaea pedrosoi* (UniProt accession A0A0D2GLV4) provides a membrane-protein example in which oligomeric context improves structural definition. The monomeric model already contains a four-helix bundle, but with relatively low confidence (pLDDT_{avg}=58.91). In contrast, the dimeric model brings two such bundles together into a much more coherent high-confidence assembly (pLDDT_{avg}=76.91, ipSAE_{min}=0.74) and more clearly defines the likely membrane boundaries, with the lower-confidence regions (light blue) lying mainly outside the membrane-spanning layer (**Fig. 5b**).

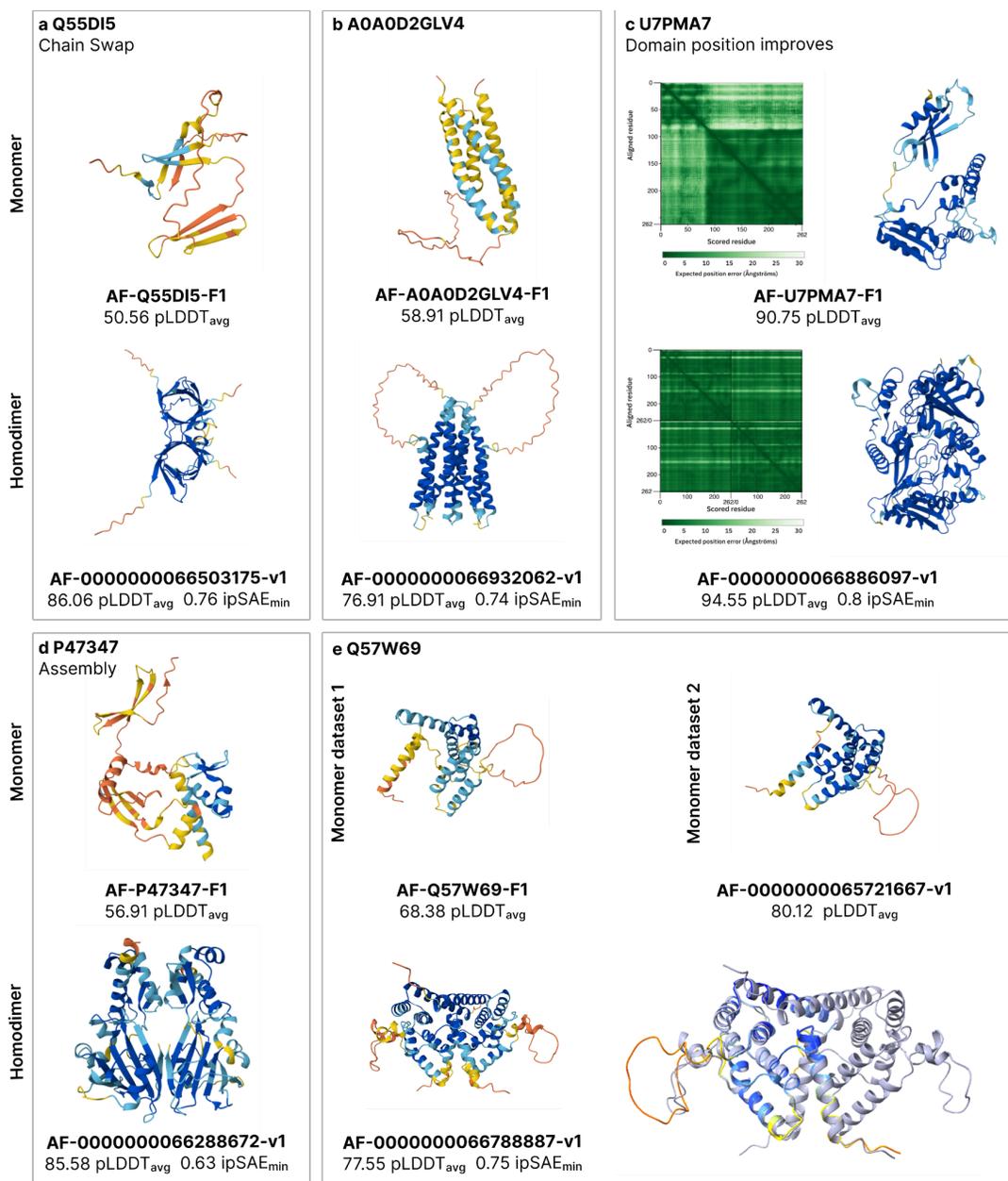


Figure 5: Representative case studies illustrate how the oligomeric context influences the structural interpretation of proteins. (a) For proteins like Q55DI5 (Transcription elongation factor Eaf N-terminal domain-containing protein), the high-confidence fold emerges only in the homodimer prediction through domain swapping, a structure entirely missed by the low-confidence monomeric model. (b) Similarly, for the membrane protein A0A0D2GLV4 (Autophagy-related protein 33), the dimeric model yields a more coherent, high-confidence assembly that better defines membrane boundaries. (c-d) Multimer prediction can refine the global inter-domain architecture even for an already confident monomeric model, such as U7PMA7 (an AB hydrolase-1 domain-containing protein) or substantially rescue low-confidence monomeric predictions, as seen with P47347 (Uncharacterised HTH-type regulator MG101). (d) P47347 (*Mycoplasma genitalium* HTH-type transcriptional regulator MG101). The dimeric model shows higher confidence than the monomer prediction, consistent with stabilisation of the fold in an oligomeric context, although the predicted interface remains less certain. (e) For *Trypanosoma brucei brucei* Q57W69, both MSA quality and oligomeric context can improve the resulting model confidence in the same manner.

Measuring the high-confidence stretches of the transmembrane helices gives distances of 34-39 Å, which is similar to the thickness achieved by the lipids packing inside a lipid bilayer (~30-40 Å), i.e. the main component of cell membranes. This case suggests that, for some membrane proteins, monomeric prediction can recover the core topology, whereas oligomeric modelling is needed to better resolve the full assembly.

The monomeric AlphaFold model of *Sporothrix schenckii* AB hydrolase-1 domain-containing protein (AF-U7PMA7-F1) is already highly confident overall (**Fig. 5c**), with $pLDDT_{avg}=94.7$, indicating that the local fold is well specified. However, modelling the protein as a dimer improves the relative positioning of the domains, as seen in the reduced uncertainty in the Predicted Aligned Error (PAE) plot, suggesting that oligomeric context helps constrain the inter-domain arrangement rather than the local secondary structure itself. This case illustrates that even when a monomeric model is confident, multimeric prediction can still refine the global architecture and identify interacting interfaces.

A related pattern is seen for the *Mycoplasma genitalium* uncharacterised HTH-type transcriptional regulator MG101 (UniProt accession P47347; **Fig. 5d**). In the monomeric prediction, overall confidence is poor ($pLDDT_{avg}=56$), whereas the dimeric model increases confidence substantially ($pLDDT_{avg}=85$), consistent with stabilization of the fold in an oligomeric context. Because HTH transcription factors commonly function as dimers, this improvement is consistent with a biologically relevant assembly in which partner interactions help define the final structure. Although it must be noted that the $ipSAE_{min}$ is 0.63, indicating lower confidence in the interface region which can likely be attributed to the fact that for a transcriptional regulator the assembly will only be fully complete after also including double-stranded DNA. This example highlights how multimer modelling can rescue proteins whose monomeric predictions are not confident, even when the specifics of the interface are predicted with modest confidence.

For the uncharacterised *Trypanosoma brucei brucei* protein (UniProt accession Q57W69), different modelling inputs produced materially different structural hypotheses (**Fig. 5e**). The original model showed only moderate confidence ($pLDDT_{avg}=68.38$), whereas use of a curated multiple sequence alignment, following the trypanosomatid-focused approach of Wheeler lab⁴⁰, improved the prediction to $pLDDT_{avg}=80.12$. The dimeric model agrees closely with this improved dataset-derived structure: superposition of AF-0000000066788887-model_v1 chain A with AF-0000000065721667-model_v1 chain A gives a Root Mean Square Deviation (RMSD) of 0.67 Å over 166 pruned atom pairs. This example shows that, for proteins from underrepresented lineages, both MSA composition and oligomeric context can influence the final structural hypothesis.

Conclusion

The transition from monomeric to proteome-scale quaternary structural coverage represents a paradigm shift for the AlphaFold Database. Across these cases, we demonstrate that oligomeric context does not only increase model confidence; it can alter how the structure is interpreted at the level of fold, assembly and domain arrangement. These results underscore: for some proteins, particularly poorly characterised or taxonomically underrepresented examples, monomeric prediction alone may provide an incomplete or even misleading structural picture. Predicting complexes at scale can instead reveal the potential interface landscape, suggest oligomeric states or identify the interfaces likely to drive complex formation. Our comprehensive structural resource spanning reference proteomes, Swiss-Prot and proteins prioritised by the WHO, including neglected diseases, will facilitate large-scale analyses to identify general principles of protein-protein interface

formation and enable new research and applications across health, biotechnology, and fundamental biological research.

Availability

1,754,242 high-confidence homodimer predictions, selected using the criteria $\text{ipSAE}_{\min} \geq 0.6$, $\text{pLDDT}_{\text{avg}} \geq 70$, and backbone clash score ≤ 10 , are available as individual entry pages through AlphaFold Database at alphafold.ebi.ac.uk. To support AFDB users in interpreting the structure models, surfaced entries are further categorised in “*very high confidence*” ($\text{ipSAE}_{\min} \geq 0.8$), “*confident*” ($0.7 \leq \text{ipSAE}_{\min} < 0.8$), and “*low confidence*” ($0.6 \leq \text{ipSAE}_{\min} < 0.7$). Homodimers not passing the previously defined threshold, together with their interface scores, are provided on the FTP page ftp.ebi.ac.uk/pub/databases/alphafold. These models and MSAs are also available through bulk download to support large-scale computational analysis, benchmarking, and method development, while reducing the risk of over-interpretation in routine biological use.

ColabFold with improved throughput is available at github.com/sokrypton/ColabFold in release 1.6.0. Acceleration libraries cuEquivariance (docs.nvidia.com/cuda/cuequivariance) and TensorRT (docs.nvidia.com/deeplearning/tensorrt), that were used to improve OpenFold throughput, are freely available with Apache 2.0 licensing. OpenFold with TensorRT and cuEquivariance integration is available at github.com/aqlaboratory/openfold.

Acknowledgements

We would like to thank NVIDIA colleagues, in particular Kyle Gion, Christian Hundt, Tobias Lasser, Isabel Wilkinson, Anthony Costa, Xin Yu, Hari Sadasivan, Youhan Lee, Yuxing Peng for support. EMBL-EBI also acknowledges their colleagues Oana Stroe, Gemma Wood and Victoria Hatch for their support. Martin Steinegger acknowledges support by the National Research Foundation of Korea grants (2020M3A9G7103933, RS-2021-NR061659 and RS-2021-NR056571, RS-2024-00396026), Novo Nordisk Foundation (NNF24SA0092560), and Creative-Pioneering Researchers Program through Seoul National University. Milot Mirdita acknowledges support from the National Research Foundation of Korea (grant RS-2023-00250470). This work was supported by the BBSRC, UK Research and Innovation [20-BBSRC/NSF-BIO], Google DeepMind and the European Molecular Biology Laboratory. The AlphaFold Protein Structure Database is supported by Google DeepMind and the European Molecular Biology Laboratory.

Competing interests

Niccolò A. E. Venanzi, Darren Hsu, Nilkanth Patel, Alejandro Chacón, Duc Tran, Quan Vu, Boris Fomitchev, Brianda Santini Lopez, Kyle Tretina, Kieran Didi, Micha Livne, Prashant Sohani, Christian Dallago are employed at NVIDIA. Nick Dietrich, Oleg Kovalevskiy, Dariusz Lasecki, Agata Laydon, Risha Patel, Augustin Židek are employed at Google DeepMind. Martin Steinegger declares an outside interest in Stylus Medicine as a scientific advisor.

Contributions

Damian Bertoni (Software, Validation), Jennifer Fleming (Conceptualization, Validation, Formal analysis, Supervision, Project administration, Writing - Original Draft, Writing - Review & Editing), Yonathan Goldtzvik (Validation), Sreenath Nair (Methodology, Software, Investigation, Formal analysis, Data Curation, Supervision), Paulyna Magaña (Validation), Urmila Paramval (Validation, Software), Ivanna Pidruchna (Validation, Software), Ahsan Tanweer (Investigation, Methodology, Data Curation, Validation, Software), Maxim I. Tsenkov (Investigation, Software, Validation, Formal

analysis, Data Curation, Methodology), Joseph Ellaway (Validation, Software, Data Curation), Jeannie Austin (Project administration), Melanie Vollmar (Validation, Formal analysis), Niccolò A. E. Venanzi (Software, Validation, Formal analysis, Investigation, Methodology, Data Curation), Darren Hsu (Software, Validation, Data Curation, Methodology), Nilkanth Patel (Software, Validation, Data Curation, Methodology), Alejandro Chacón (Software, Validation), Duc Tran (Software, Validation), Quan Vu (Software), Boris Fomitchev (Software), Brianda Santini Lopez (Software, Validation, Data Curation), Kyle Tretina (Writing - Original Draft, Writing - Review & Editing), Maciej Majewski (Software, Validation, Methodology, Investigation, Formal analysis), Kieran Didi (Formal analysis), Micha Livne (Data Curation), Prashant Sohani (Data Curation), Yewon Han (Methodology, Software, Validation, Investigation, Formal analysis, Data Curation, Visualisation, Writing - Results & Methods), Sooyoung Cha (Validation, Investigation, Formal analysis, Visualisation, Writing - Methods), Milot Mirdita (Conceptualization, Methodology, Software, Investigation, Formal analysis, Data Curation, Supervision, Writing - Review & Editing), Christian Dallago (Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition), Martin Steinegger (Conceptualization, Methodology, Software, Validation, Investigation, Formal analysis, Resources, Data Curation, Supervision, Project administration, Funding acquisition, Writing - Original Draft, Writing - Review & Editing), Sameer Velankar (Conceptualization, Methodology, Software, Data Curation, Validation, Investigation, Resources, Supervision, Project administration, Funding acquisition, Writing - Review & Editing), Nick Dietrich (Writing - Review & Editing), Oleg Kovalevskiy (Writing - Review & Editing), Dariusz Lasecki (Writing - Review & Editing), Agata Laydon (Writing - Review & Editing), Risha Patel (Writing - Review & Editing), Augustin Žídek (Writing - Review & Editing)

References

1. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**, D609–D617 (2025).
2. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646 (2023).
3. Kallenborn, F. *et al.* GPU-accelerated homology search with MMseqs2. *Nat Methods* **22**, 2024–2027 (2025).
4. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021) doi:10.1101/2021.10.04.463034.
5. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
6. Ahdritz, G. *et al.* OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat Methods* **21**, 1514–1524 (2024).
7. Greenblatt, J. F., Alberts, B. M. & Krogan, N. J. Discovery and significance of protein-protein interactions in health and disease. *Cell* **187**, 6501–6517 (2024).
8. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* **47**, D520–D528 (2019).
9. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
10. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* **52**, D368–D375 (2024).
11. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human

- proteome. *Nature* **596**, 590–596 (2021).
12. Bertoni, D. *et al.* AlphaFold Protein Structure Database 2025: a redesigned interface and updated structural coverage. *Nucleic Acids Res* **54**, D358–D362 (2026).
 13. Bordin, N. *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun Biol* **6**, 160 (2023).
 14. Passaro, S. *et al.* Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv* (2025) doi:10.1101/2025.06.14.659707.
 15. Geffner, T. *et al.* Proteina: Scaling flow-based protein structure generative models. *arXiv [cs.LG]* (2025) doi:10.48550/arXiv.2503.00710.
 16. Matthews, J. M. & Sunde, M. Dimers, oligomers, everywhere. *Adv Exp Med Biol* **747**, 1–18 (2012).
 17. Burré, J., Sharma, M. & Südhof, T. C. α -Synuclein assembles into higher-order multimers upon membrane binding to promote SNARE complex formation. *Proc Natl Acad Sci U S A* **111**, E4274–83 (2014).
 18. Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358–63 (2014).
 19. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
 20. Schmid, E. W. *et al.* Proteome-wide screening for human protein-protein interactions. *bioRxiv* (2025) doi:10.1101/2025.11.10.687652.
 21. Zhang, J. *et al.* Predicting protein-protein interactions in the human proteome. *Science* **390**, eadt1630 (2025).
 22. Catoiu, E. A. *et al.* QSProteome: a community-driven interactive platform for large-scale exploration and evaluation of predicted protein complex structures. *Nucleic Acids Res* **54**, D661–D672 (2026).
 23. Schmid, E. W. & Walter, J. C. Predictomes, a classifier-curated database of AlphaFold-modeled protein-protein interactions. *Mol Cell* **85**, 1216–1232.e5 (2025).
 24. Schweke, H. *et al.* An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999–1010.e15 (2024).
 25. Humphreys, I. R. *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
 26. Moriwaki, Y. *et al.* Predicting protein complexes in biosynthetic gene clusters. *bioRxiv* (2025) doi:10.1101/2025.10.26.684697.
 27. Danielsson, S. N. & Elofsson, A. Reliable identification of homodimers using AlphaFold. *bioRxiv* (2025) doi:10.1101/2025.11.27.691011.
 28. Lee, C. Y. *et al.* Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation. *Mol Syst Biol* **20**, 75–97 (2024).
 29. Didi, K. *et al.* Efficient protein structure prediction from compact computers to datacenters with OpenFold-TRT. *bioRxiv* (2026) doi:10.64898/2026.03.11.711233.
 30. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol Biol* **406**, 89–112 (2007).
 31. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
 32. Dunbrack, R. L., Jr. Rēs ipSAE loquuntur: What’s wrong with AlphaFold’s score and how to fix it. *bioRxiv* (2025) doi:10.1101/2025.02.10.637595.
 33. Zhu, W., Shenoy, A., Kundrotas, P. & Elofsson, A. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics* **39**, (2023).
 34. Overath, M. D. *et al.* Predicting experimental success in DE Novo binder design: A meta-analysis

- of 3,766 experimentally characterised binders. *bioRxiv* 2025.08.14.670059 (2025)
doi:10.1101/2025.08.14.670059.
35. What happened in the Nipah Protein Design Competition so far? + Prediction markets for protein design. <https://www.adaptyvbio.com/blog/nipah-submissions/>.
 36. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
 37. Kim, W. *et al.* Rapid and sensitive protein complex alignment with Foldseek-Multimer. *Nat Methods* **22**, 469–472 (2025).
 38. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246 (2024).
 39. Kennedy, A. B. W. & Sankey, H. R. The thermal efficiency of steam engines. Report of the committee appointed to the council upon the subject of the definition of a standard or standards of thermal efficiency for steam engines: With an introductory note. (including appendixes and plate at back of volume). *Minutes Proc. Inst. Civ. Eng.* **134**, 278–312 (1898).
 40. Wheeler, R. J. A resource for improved predictions of Trypanosoma and Leishmania protein three-dimensional structure. *PLoS One* **16**, e0259871 (2021).
 41. Wojdyr, M. GEMMI: A library for structural biology. *J. Open Source Softw.* **7**, 4200 (2022).
 42. Didi, K. *et al.* Scaling Atomistic Protein Binder Design with Generative Pretraining and Test-Time Compute. in *The Fourteenth International Conference on Learning Representations* (2026).
 43. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).
 44. Lee, S., Kim, J., Mirdita, M., Gilchrist, C. L. M. & Steinegger, M. Easy and interactive taxonomic profiling with Metabuli App. *Bioinformatics* **41**, (2025).

Methods

Sequence and interaction dataset construction

Sequence selection and interaction definition

We selected 23,441,822 UniProt 2025_04 sequences by filtering UniProt by a set of proteomes and for protein (monomers) with length between 15 and 1,500 amino acids. The set of 4,777 proteomes were selected from Swiss-Prot, proteomes within the World Health Organisation (WHO) global health proteomes, and the top downloaded reference proteomes in UniProt. From this set, homodimers were simply derived by duplicating each monomer into a complex, thus resulting in 23,441,822 homodimers, with the longest heterodimers in the set containing a maximum of 3,000 amino acids.

Heterodimers were instead derived by extracting physical interaction evidence provided by the STRING database. In particular, the file [protein.physical.links.v12.0.txt.gz](https://string-db.org/cgi/download) (11.1 GB) CC-BY-4.0, obtained from <https://string-db.org/cgi/download> in January of 2026, was downloaded. This file contains annotations of physically interacting partners with their corresponding STRING score. From these interactions, we filtered for sequences within a prioritised set of proteomes from 16 model organisms and 30 global health-relevant proteomes, resulting in 7,620,644 candidate complexes with two distinct chains. Further, these interactions were filtered to produce heterodimers of maximally 3,000 amino acids. No further filtering, such as STRING score thresholds, were applied, to obtain the highest coverage.

String ID to Uniprot accession mapping

To map STRING (v12.0) physical interaction proteins to our UniProt proteomes, we applied a three-step strategy with decreasing priority. (1) UniProt ID mapping: STRING provides a direct STRING-to-UniProt identifier mapping via the file [protein.aliases.v12.0.txt.gz](https://string-db.org/cgi/download) (CC-BY-4.0), downloaded from <https://string-db.org/cgi/download> in January 2026. As this file contains obsolete UniProt accessions, we re-mapped all entries against the UniProt 2025_04 release, obtained from https://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/knowledgebase/complete/ in January 2026. (2) CRC64 hash matching: For proteins not resolved by step 1, we matched CRC64 sequence hashes between STRING and UniProt entries. (3) MMseqs2 sequence search: For the remaining unmapped proteins, we ran per-taxon MMseqs2 searches (STRING sequences as queries against our reference set) requiring 100% sequence identity and $\geq 95\%$ bidirectional coverage (--min-seq-id 1 -c 0.95 --cov-mode 0 --alignment-mode 3 -s 4), retaining only the best hit per query based on E-value and bit score.

For both the ID mapping and CRC64 steps, we applied a taxon priority filter: when a STRING protein could be mapped to multiple UniProt accessions, same-species mappings (where the taxon ID extracted from the STRING identifier matched the UniProt taxon ID via the UniProt taxonomy dump) were preferred; cross-taxon mappings were retained only when no same-species match existed. When multiple methods successfully mapped the same STRING protein, the highest-priority result was kept (ID mapping > CRC64 > MMseqs2).

Finally, STRING physical interaction pairs were deduplicated such that each unordered pair (A-B / B-A) appeared only once, and the mapped UniProt accessions were converted to internal identifiers (AF-series IDs) using a reference table, producing per-taxon interaction files. This pipeline achieved a mapping rate of 73.1% for the model organism set (5,503,385 out of 7,528,643 STRING entries) and 82.4% for the WHO global health proteome set (2,145,870 out of 2,604,057). Redundancy removal resulted in 7,620,644 candidates for heterodimeric prediction.

Inference tools and approaches

To improve computational efficiency, we separated the two compute-demanding workloads of MSA generation and structure prediction.

MSA generation

We generated homodimer MSAs by leveraging a pre-release version of ColabFold 1.6.0 using the MMseqs2-GPU backend (version 18-8cc5c). In particular, we leveraged the *colabfold_search* tool with a new MSA filtering strategy for homomeric predictions. This strategy keeps only the best hit per taxon based on alignment score, essentially returning a monomer MSA in a3m format filtered to contain only the highest scoring hit per taxon. The MSA is then duplicated during AlphaFold-multimer's featurization for homodimeric prediction. *colabfold_search* was run using the following parameters: `--use-env 0 --pairing_strategy 1 --pair-mode paired --filter 2 --db-load-mode 2`. The underlying sequence database was the pre-built ColabFold GPU-accelerated search database uniref30_2302, which provides clustered coverage of all UniRef100 sequences.

To generate heterodimer MSAs, we simply concatenated the previously obtained homodimer/monomer MSAs for two distinct sequences, without additional pairing. This design was chosen on the one hand, to decrease computational complexity, and on the other, as these input MSAs were already restricted to the best hit per taxon, the need for pairing was reduced. This rationale was supported by our investigation into different MSA pairing strategies; comparing taxonomy-based pairing against simple concatenation revealed that additional pairing did not clearly yield better downstream predictions, especially at higher ipTM threshold (ipTM > 0.8). Thus, we opted for a simple concatenation strategy, which allows the direct re-use of homodimer MSAs, reducing computational demands (**Supplementary Fig. 3**).

Structure prediction

Protein complex 3D structure prediction from MSAs was executed either through ColabFold's *colabfold_batch* command (primarily, for homodimers), or by using an accelerated implementation of OpenFold that leveraged NVIDIA TensorRT and cuEquivariance libraries. Both tools utilised the same set of parameters, namely: one set of weights from AlphaFold Multimer (model_1_multimer_v3), four recycles/iterations with early stopping, and no relaxation. The outputs from the two inference tools were verified to yield equivalent results, as previously reported²⁹. Given the scale of the prediction campaign, the choice of inference parameters was aimed at the reduction of computational demand, while maintaining prediction accuracy. To further reduce overhead, we implemented a parameter in ColabFold (`--skip-output msa,plots,paes_json`) to only output essential files such as the structure (.pdb) and quality (.json) metrics, and optimised MSA read-in. Failures like out of memory errors or errors due to missing atoms were observed during structure prediction in the presence of large MSAs/long sequences, and non-standard amino acids (e.g. "X"), respectively. These failure modes occurred in a small fraction of the data (<5%), which were not further investigated.

High-performance-compute scaling

In order to increase throughput at scale (i.e., in a multi-GPU, multi-node setting), we wrapped MSA execution and structure predictor in Slurm pipelines which spawned multiple inference processes per GPU on each node, maximising GPU utilisation. GPUs independently process chunks of data reading from a queueing system, making the inference pipeline naively parallelisable to the number

of available GPUs and nodes. The pipelines were executed on a DGX H100 superpod instance, using units of nodes for scale (i.e., at minimum, 8 H100 GPUs).

For ColabFold-based homodimer inference, higher throughput was obtained by packing homodimers of equal length into a single batch to process, sorted by their MSA-depth in descending order. This reduced the amount of Jax (0.7.2) model recompilations, thus increasing prediction throughput. This trick however does not work when processing heterodimers, given that the length of individual chains differs.

For OpenFold, whether for homodimers or heterodimers, this packing strategy is not needed, as the method does not require model re-compilation. However, given a dependency between sequence length and execution time, reserving longer sequences for individual jobs may be beneficial if operating with specific Slurm runtimes. To further optimise the process, input featurizations (CPU-bound) were performed for the next input query alongside the inference step for the current query (GPU-bound).

Post-processing toolkit and data products

ColabFold and OpenFold outputs (i.e., .pdb and .json files) were converted into AFDB-compliant data products using an extension of the AFDB-Integration-Kit (manuscript in preparation), a multi-stage post-processing toolkit that computes interface and structural quality metrics, generates ModelCIF-compliant mmCIF and Binary CIF files, annotates secondary structure, validates outputs against AFDB schemas, and packages metadata into batched JSON files for database ingestion.

Interface quality metrics such as ipSAE, ipTM, pDockQ2, and LIS were derived from PAE matrices and 3D coordinates, with ipSAE computed through a C++ implementation that processes shards of multiple data in batch mode, increasing throughput. As ipSAE is directional, we computed scores for both chain orderings and defined $ipSAE_{min} = \min(ipSAE_{A \rightarrow B}, ipSAE_{B \rightarrow A})$ to reduce each complex to a single conservative estimate of interface quality. Analogous minimum operations were applied to LIS (LIS_{min}). Backbone and heavy-atom clash counts and interface residue assignments were identified via GPU-accelerated scripts that pack batches of up to 512 structures into GPU-resident tensors and dispatch atomic distance kernels with `torch_cluster.radius_graph`. Secondary structure annotation was inferred with PyDSSP (github.com/ShintaroMinami/PyDSSP; commit e251a43), a vectorised NumPy/PyTorch implementation that eliminates subprocess overhead and processes entire batches in memory. JSON parsing in the toolkit was replaced with *orjson* (3.11.4), reducing deserialisation of large PAE matrices by roughly an order of magnitude. Similarly, PDB structure parsing was replaced with *Gemmi*⁴¹.

Each stage additionally parallelises its own work across `ProcessPoolExecutor`` process pools, bypassing the Python GIL for CPU-bound operations, while metadata lookups were batched into a single DuckDB query per data shard, and cached in module-level globals shared across workers, eliminating per-model database roundtrips. DuckDB memory usage is capped to prevent out-of-memory events on shared nodes, and downstream shard aggregation streams data through PyArrow batches to avoid loading full datasets into memory.

To distribute this workload across nodes in our HPC setting, we implemented a Slurm array job framework based entirely on file-level coordination: a driver script partitions a flat list of model IDs into contiguous shards of at most 5,000 structures per array task, with each task operating independently and deterministically from its array index alone. Each node writes outputs to local scratch memory, uploads completed batches directly to object storage via `s5cmd``, and records only small marker files on a shared Lustre filesystem, avoiding I/O contention at scale. Fault tolerance was achieved through two-level automatic resume, skipping completed pipeline stages within a

shard and exiting immediately for already-processed shards on resubmission, making runs across hundreds of nodes robust to node failures or Slurm job preemption.

Validation and analyses

Structural coverage analysis

To quantify the extent to which predicted complexes expand structural coverage beyond experimentally determined multimers, we compared the number of high-confidence predicted complexes (1,754,242 homodimers and 56,959 heterodimers) with the number of experimentally determined multimeric structures deposited in the Protein Data Bank that are mapped to UniProt (downloaded in January 2026), per organism. PDB multimer counts were tallied by NCBI Taxonomy identifier, and taxonomic lineage information was resolved using NCBI taxonomy files (names.dmp, nodes.dmp). Organisms with fewer than 10 high-confidence predicted structures were excluded to avoid unreliable estimates. For organism-level comparison, each species was represented by both its PDB multimer count and its predicted high-confidence complex count. For kingdom-level comparison (**Fig. 1e**, right), taxons were aggregated into seven broad categories: Metazoa, Viridiplantae, Fungi, Other Eukaryotes, Bacteria, Archaea, and Viruses. Organisms with no corresponding PDB multimer entry were retained and displayed with predicted counts only.

Cutoff validation dataset

To establish ground truth and negative control sets for confidence threshold calibration, we used PDB chain-level annotations downloaded in March 2026. These annotations map each PDB entry to its stoichiometry, release date, and chain-level sequence alignment to UniProt accessions. We retained only entries for which the PDB chain sequence exactly matched the aligned UniProt sequence, ensuring unambiguous correspondence between predicted and experimental structures. To minimise overlap with the AlphaFold-Multimer training set, we further restricted the dataset to entries released after 30th September 2021. From this filtered subset, we extracted homodimers (1,968 entries) as positive controls expected to be confidently predicted, and monomers (2,211 entries) as negative controls expected to score poorly in homodimer prediction.

Sequence cluster comparison

Sequence clusters were generated using MMseqs2³⁶ (release 18) from the 23,441,822 input sequence set. Cluster representatives were obtained by clustering sequences at 98% sequence identity and 95% coverage using the parameters `--min-seq-id 0.98 -c 0.95 --cov-mode 0`. Structural similarity within each cluster was evaluated using Foldseek³⁸ (commit d6204679). A Foldseek database was constructed from the 1,754,242 high-confidence structures, and representative structures were aligned against cluster members using foldseek *structurealign* with `--alignment-type 2`. Structural similarity between representative and member structures was quantified using TM-score.

Within-homodimer chain comparison

To assess structural consistency between chains in predicted homodimers, structural alignments were performed between chain A and chain B of each complex using Foldseek (commit d6204679). A Foldseek database was constructed from the 1,754,242 high-confidence structures, and chain pairs were aligned using monomer structural alignment. A prefilter mapping was used to directly match

chain A to chain B for each predicted complex, and structural similarity was quantified using Foldseek TM-score metrics.

Clustering analyses

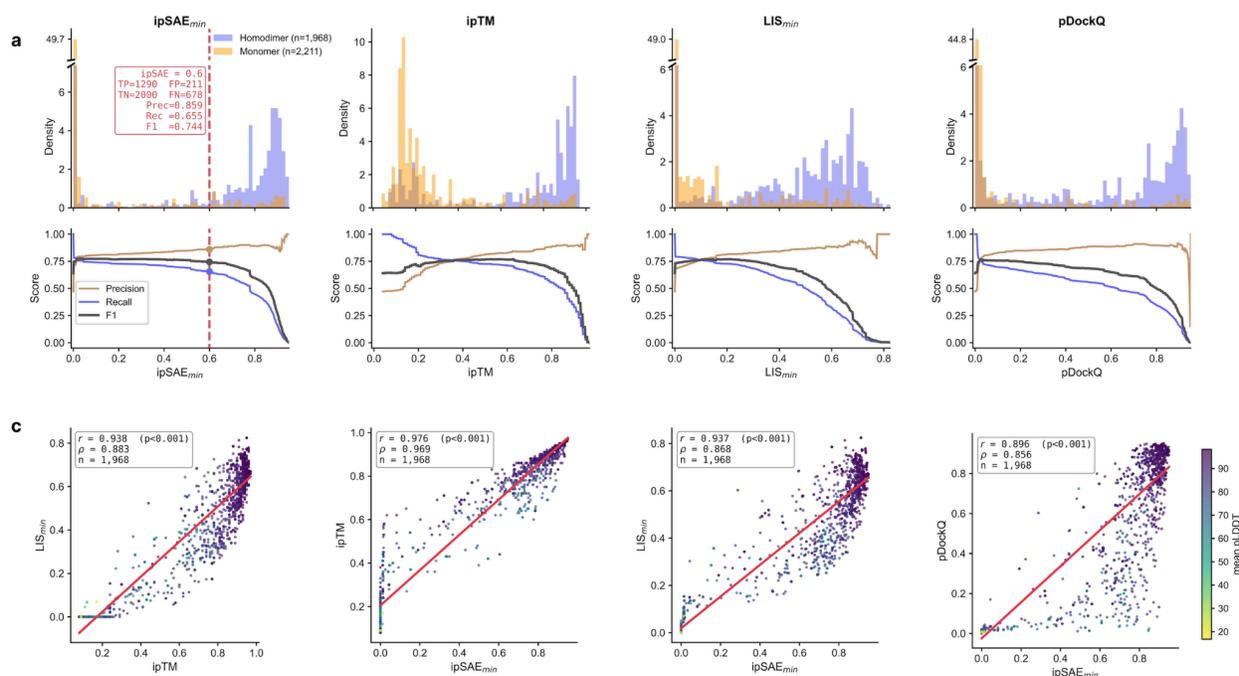
1,811,201 structures from 1,754,242 *high-confidence* homomeric and 56,959 *tentative high-confidence* heterodimeric structures were clustered using Foldseek Multimercluster (commit d6204679). Cluster representatives were obtained by clustering structures at 60% 3Di sequence coverage, 30% interface similarity and 70% chain similarity using the parameters `-c 0.6 --interface-lddt-threshold 0.3 --chain-tm-threshold 0.7 --cov-mode 0`. These parameters follow those used for the Teddymer dataset⁴².

Taxonomic analysis

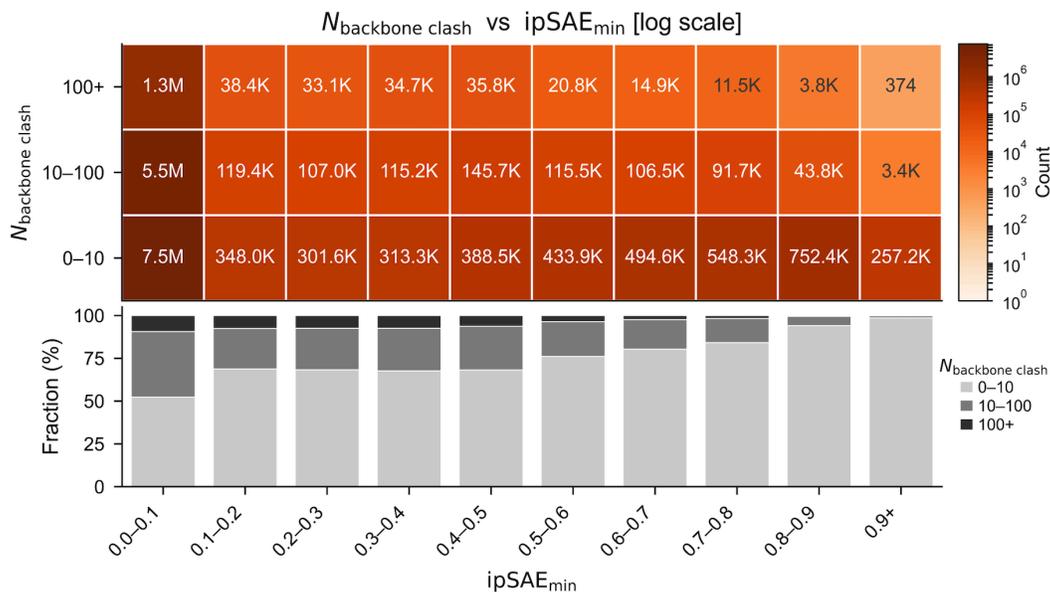
Taxonomic lineages and rank assignments were resolved using the NCBI Taxonomy database (taxdump, downloaded in February 2026); specifically, nodes.dmp and names.dmp were used to construct parent-child relationships and to assign each taxon to standardised ranks (e.g., superkingdom, phylum). These rank assignments were used both for computing cluster-level lowest common ancestors (LCAs) and for stratifying prediction success rates across taxonomic clades (**Fig. 2d**). Based on UniProt taxonomic assignments and using the MMseqs2 taxonomy's lca module⁴³, we computed for each non-singleton cluster, the LCA from its member taxa, and visualized the resulting taxonomic distribution as a Sankey diagram using Metabuli-App⁴⁴.

Structural search of the high confidence subset against PDB

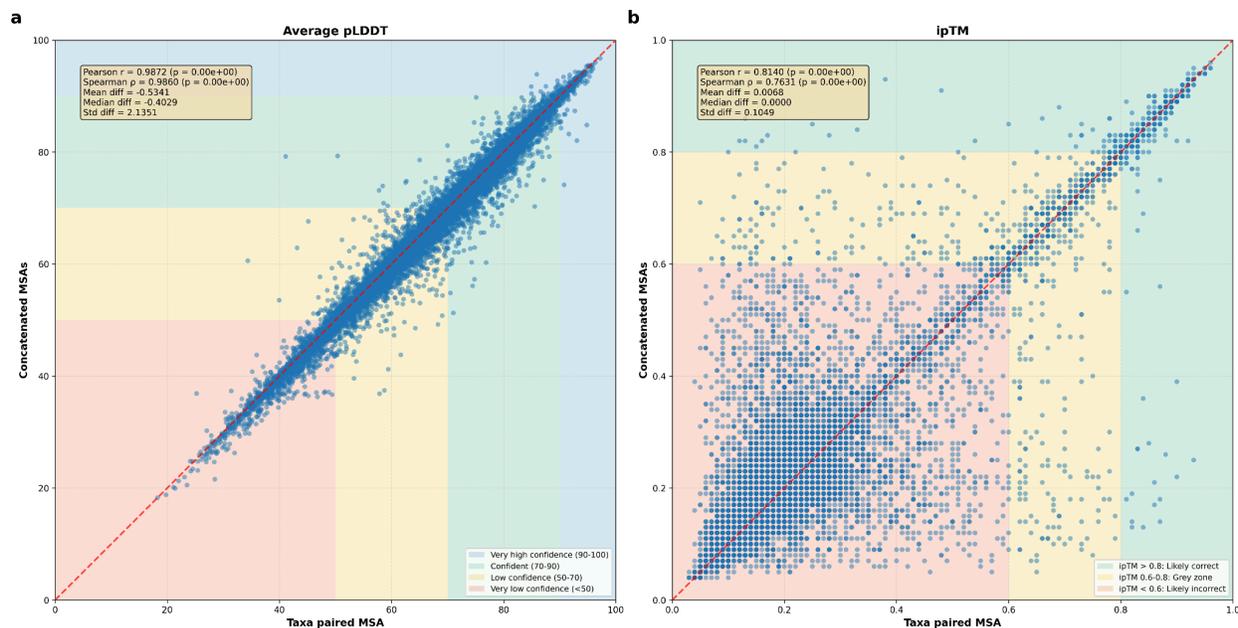
To assess structural coverage against experimentally determined structures, we searched the 1,754,242 high-confidence homomeric and 56,959 tentative high-confidence heterodimeric predicted structures against the PDB100, a clustered version of the PDB downloaded in January 2026, using Foldseek Multimer (commit d6204679). We used the following parameters: `--cov-mode 2 --tmscore-threshold 0.65 --cluster-search 1`. Coverage mode 2 (query coverage) was chosen because the query set consists of predicted dimeric structures searched against PDB entries that may contain larger multimeric assemblies; this mode ensures that individual chains within higher-order complexes can be matched. The TM-score threshold of 0.65 was adopted following the benchmarks reported in the Foldseek Multimer publication. The flag `--cluster-search 1` enables search against the full PDB rather than cluster representatives only.



Supplementary Figure 1: additional homodimer validation metrics. (a) Different metrics (x axis) against the density of either a homodimer set expected to be predicted as homodimer (blue), or a monomer set not expected to be predicted as a homodimer (yellow). (b) Agreement between ipSAE_{min} or ipTM (x axis) and other quality metrics (x and y axes).



Supplementary Figure 2: Backbone clashes appear less frequently at high ipSAE_{min} values. (Top) Heatmap showing the number of heterodimer structural models (from a random sample of 500,000) distributed across bins of ipSAE_{min} (x-axis, binned in 0.1 intervals) and backbone clash count (y-axis; 0-10, 10-100, and 100+). Cell values indicate raw counts. (Bottom) Stacked bar chart showing the fractional composition of backbone clash categories within each ipSAE_{min} bin. Models with high ipSAE_{min} (≥ 0.6) are predominantly clash-free (number of backbone clashes < 10), whereas low ipSAE_{min} bins are enriched for models with extensive steric clashes.



Supplementary Figure 3: Comparison of pLDDT (a) and ipTM (b) scores for predicted structures using ColabFold (via colabfold_batch) using two distinct MSA generation strategies as inputs. On the x axis: taxa paired MSAs following the standard colabfold_search pipeline with MMseqs2-GPU as a backend. On the y axis: concatenated homodimer (taxa filtered monomer) MSAs computed with colabfold_search using MMseqs2-GPU as a backend. For average pLDDT, there is generally great agreement between the two input MSA strategies (plot coloured by AFDB defined confidence bands). For ipTM, there is also general agreement, with more noise in the “likely incorrect” regions, and narrower limits of agreement with greater correlation in the “likely correct” region, despite outliers in both directions.

Supplementary Material, AlphaFold Database expands to proteome-scale quaternary structures.

Han, Tsenkov, Venanzi et al.

00242543,UP000242554,UP000242636,UP000242638,UP000242757,UP000242798,UP000242823,UP000242942,UP000242950,UP000242976,UP000242991,UP000243217,UP000243283,UP000243411,UP000243459,UP000243492,UP000243495,UP000243498,UP000243523,UP000243650,UP000243666,UP000243699,UP000243778,UP000243845,UP000243884,UP000243975,UP000244005,UP000244309,UP000244905,UP000244962,UP000244978,UP000245119,UP000245133,UP000245207,UP000245320,UP000245340,UP000245341,UP000245380,UP000245488,UP000245695,UP000245711,UP000245910,UP000246171,UP000246263,UP000246397,UP000246436,UP000246464,UP000246679,UP000246914,UP000246998,UP000247120,UP000247192,UP000247426,UP000247437,UP000247485,UP000247498,UP000247586,UP000247647,UP000247790,UP000248340,UP000248405,UP000248423,UP000248481,UP000248482,UP000248483,UP000248484,UP000248925,UP000248929,UP000249118,UP000249287,UP000249293,UP000249343,UP000249363,UP000249390,UP000249673,UP000249829,UP000250241,UP000251241,UP000251314,UP000251714,UP000252167,UP000252519,UP000253472,UP000254069,UP000254082,UP000254429,UP000254519,UP000254807,UP000254920,UP000255036,UP000255233,UP000255367,UP000255382,UP000255518,UP000256485,UP000256499,UP000256970,UP000257109,UP000258159,UP000258290,UP000258798,UP000259373,UP000259509,UP000259687,UP000260425,UP000260530,UP000261340,UP000261360,UP000261420,UP000261480,UP000261580,UP000261600,UP000261640,UP000261680,UP000261681,UP000262004,UP000262320,UP000263690,UP000264006,UP000264120,UP000264800,UP000264820,UP000264840,UP000264883,UP000265000,UP000265140,UP000265300,UP000265325,UP000265520,UP000265618,UP000265692,UP000266841,UP000267029,UP000267096,UP000267145,UP000267218,UP000267524,UP000267821,UP000268673,UP000269641,UP000269669,UP000269793,UP000269945,UP000270661,UP000270924,UP000271272,UP000271868,UP000271974,UP000272105,UP000272662,UP000274082,UP000274429,UP000274504,UP000274756,UP000275408,UP000275480,UP000276254,UP000276295,UP000277198,UP000278548,UP000278807,UP000279389,UP000279551,UP000279909,UP000280188,UP000280281,UP000281986,UP000282613,UP000283360,UP000283509,UP000283530,UP000283616,UP000283805,UP000283841,UP000283868,UP000284095,UP000284621,UP000284702,UP000285060,UP000285120,UP000285710,UP000286134,UP000286640,UP000286641,UP000287701,UP000287853,UP000287872,UP000288216,UP000288725,UP000289347,UP000289348,UP000289664,UP000289726,UP000289738,UP000289886,UP000289930,UP000290289,UP000290636,UP000290640,UP000290900,UP000291000,UP000291020,UP000291022,UP000291343,UP000291422,UP000291933,UP000292052,UP000292209,UP000293596,UP000293823,UP000294530,UP000294599,UP000294673,UP000294855,UP000295192,UP000295252,UP000295530,UP000296049,UP000296307,UP000296678,UP000296700,UP000297031,UP000297245,UP000297703,UP000297855,UP000298030,UP000298196,UP000298264,UP000298551,UP000298616,UP000298636,UP000298652,UP000299084,UP000301870,UP0003034864,UP000305881,UP000306102,UP000307000,UP000307430,UP000307440,UP000309816,UP000310200,UP000311919,UP000314980,UP000314981,UP000314982,UP000314983,UP000314984,UP000314987,UP000315363,UP000315496,UP000315995,UP000316012,UP000316079,UP000316621,UP000316759,UP000316968,UP000317371,UP000317650,UP000318125,UP000318571,UP000318821,UP000320239,UP000320707,UP000320735,UP000320896,UP000321157,UP000321393,UP000321440,UP000321555,UP000321612,UP000321868,UP000321893,UP000322000,UP000322631,UP000322667,UP000322983,UP000323046,UP000323067,UP000323257,UP000324091,UP000324222,UP000324639,UP000324748,UP000324870,UP000324871,UP000326062,UP000326328,UP000326396,UP000326532,UP000326565,UP000326695,UP000326950,UP000326961,UP000327000,UP000327013,UP000327085,UP000327118,UP000327236,UP000327439,UP000327468,UP000328092,UP000335636,UP000336361,UP000336676,UP000338437,UP000338646,UP000406184,UP000409147,UP000419144,UP000422232,UP000424527,UP000428325,UP000433483,UP000434101,UP000434172,UP000436088,UP000437017,UP000439965,UP000440125,UP000444721,UP000447434,UP000448177,UP000449547,UP000452235,UP000463857,UP000464024,UP000464105,UP000467105,UP000467132,UP000467840,UP000471190,UP000471633,UP000472240,UP000472241,UP000472264,UP000472265,UP000472268,UP000472272,UP000472273,UP000472274,UP000472275,UP000472276,UP000472277,UP000473826,UP000475037,UP000475862,UP000480763,UP000481043,UP000483018,UP000485058,UP000492820,UP000492821,UP000494040,UP000494165,UP000494206,UP000494272,UP000498640,UP000499800,UP000500822,UP000500930,UP000501105,UP000501169,UP000501914,UP000501937,UP000502003,UP000502298,UP000502572,UP000502823,UP000503318,UP000503349,UP000503405,UP000503440,UP000503557,UP000504601,UP000504602,UP000504603,UP000504604,UP000504605,UP000504606,UP000504607,UP000504608,UP000504609,UP000504610,UP000504612,UP000504614,UP000504616,UP000504623,UP000504624,UP000504626,UP000504629,UP000504630,UP000504632,UP000504633,UP000504634,UP000504635,UP000504639,UP000504640,UP000509322,UP000509510,UP000509626,UP000515123,UP000515124,UP000515125,UP000515126,UP000515129,UP000515131,UP000515132,UP000515135,UP000515140,UP000515145,UP000515146,UP000515148,UP000515150,UP000515150,UP000515151,UP000515152,UP000515153,UP000515154,UP000515156,UP000515161,UP000515162,UP000515163,UP000515164,UP000515165,UP000515200,UP000515202,UP000515203,UP000515204,UP000515208,UP000515211,UP000515847,UP000516359,UP000516361,UP000516369,UP000517252,UP000523087,UP000524187,UP000525078,UP000526125,UP000527355,UP000537131,UP000540568,UP000543174,UP000543804,UP000549775,UP000550136,UP000551758,UP000555649,UP000557509,UP000557566,UP000558488,UP000562238,UP000564677,UP000565468,UP000575874,UP000579812,UP000583929,UP000585474,UP000591844,UP000593563,UP000593564,UP000593571,UP000593576,UP000593579,UP000594034,UP000594220,UP000594260,UP000594261,UP000594262,UP000594342,UP000594402,UP000594454,UP000594638,UP000594865,UP000595437,UP000595783,UP000595790,UP000595892,UP000595894,UP000596276,UP000596660,UP000596661,UP000596742,UP000597762,UP000602510,UP000610960,UP000612585,UP000613401,UP000614350,UP000615613,UP000618051,UP000619457,UP000622580,UP000622797,UP000623687,UP000624244,UP000625711,UP000626092,UP000627573,UP000631114,UP000636800,UP000637239,UP000639338,UP000639772,UP000646548,UP000646548,UP000646187,UP000652761,UP000654075,UP000654913,UP000659654,UP000662840,UP000663068,UP000663205,UP000663292,UP000664032,UP000664048,UP000664991,UP000672009,UP000673383,UP000674179,UP000675881,UP000675900,UP000677054,UP000678393,UP000678499,UP000680417,UP000681035,UP000681084,UP000682530,UP000683360,UP000693970,UP000694018,UP000694240,UP000694380,UP000694381,UP000694384,UP000694385,UP000694386,UP000694387,UP000694388,UP000694389,UP000694390,UP000694391,UP000694392,UP000694393,UP000694394,UP000694395,UP000694398,UP000694399,UP000694400,UP000694402,UP000694403,UP000694404,UP000694405,UP000694406,UP000694407,UP000694409,UP000694411,UP000694412,UP000694414,UP000694415,UP000694417,UP000694421,UP000694422,UP000694423,UP000694425,UP000694427,UP000694428,UP000694520,UP000694521,UP000694523,UP000694543,UP000694544,UP000694545,UP000694546,UP000694548,UP000694554,UP000694556,UP000694557,UP000694558,UP000694559,UP000694561,UP000694564,UP000694565,UP000694566,UP000694580,UP000694660,UP000694840,UP000694843,UP000694844,UP000694845,UP000694853,UP000694856,UP000694857,UP000694861,UP000694863,UP000694865,UP000694871,UP000694886,UP000694904,UP000694905,UP000694906,UP000694910,UP000694915,UP000694918,UP000694923,UP000694924,UP000694930,UP000694941,UP000694949,UP000694950,UP000695000,UP000695022,UP000695023,UP000695026,UP000700334,UP000701341,UP000711996,UP000716291,UP000719412,UP000734854,UP000736672,UP000738359,UP000744769,UP000747542,UP000756132,UP000767238,UP000770661,UP000777482,UP000782854,UP000788993,UP000789617,UP000790347,UP000791440,UP000792457,UP000793478,UP000794436,UP000797356,UP000799539,UP000800092,UP000803844,UP000805418,UP000811609,UP000812440,UP000816034,UP000817156,UP000821853,UP000821866,UP000823588,UP000823872,UP000824633,UP000826271,UP000827986,UP000828062,UP000828234,UP000828321,UP000828519,UP000828573,UP000828581,UP000828719,UP000828777,UP000828861,UP000829196,UP000829401,UP000829907,UP000829999,UP000831086,UP000831120,UP000831195,UP000831534,UP000831813,UP000831833,UP000833502,UP000833506,UP000835761,UP000836335,UP000836404,UP000836841,UP000837675,UP000838412,UP000839563,UP000886885,UP000887013,UP000887159,UP000887520,UP000887540,UP000887561,UP000887562,UP000887563,UP000887567,UP000887568,UP000887569,UP000887572,UP000887572,UP001054821,UP001055185,UP001055553,UP001055712,UP001056012,UP001056425,UP001058974,UP001063698,UP001085076,UP001105220,UP001108240,UP001108280,UP001139955,UP001140949,UP001147733,UP001152024,UP001152320,UP001152484,UP001152562,UP001152747,UP001152759,UP001152918,UP001153076,UP001153269,UP001153365,UP001153737,UP001154078,UP001154282,UP001154329,UP001154400,UP001154402,UP001154860,UP001155700,UP001157006,UP001159641,UP001161247,UP001162076,UP001162480,UP001162541,UP001164049,UP001164776,UP001164929,UP001165190,UP001165780,UP001166674,UP001172457,UP001174136,UP001177003,UP001178507,UP001179946,UP001181741,UP001186944,UP001187343,UP001187531,UP001190640,UP001195914,UP001200034,UP001205105,UP001206925,UP001215939,UP001222027,UP001223588,UP001224775,UP001227230,UP001229421,UP001230385,UP001231189,UP001231518,UP001237071,UP001237278,UP001237829,UP001243989,UP001245389,UP001249851,UP001253637,UP001253851,UP001256966,UP001273209,UP001283361,UP001286313,UP001292079,UP001292094,UP001295423,UP001295469,UP001314229,UP001317516,UP001318040,UP001321473,UP001331515,UP001331761,UP001334084,UP001356427,UP001359559,UP001362899,UP001364617,UP001366552,UP001374535,UP001374584,UP001375240,UP001378592,UP001395587,UP001408789,UP001412239,UP001434419,UP001443914,UP001445076,UP001456562,UP001457282,UP001460270,UP001472978,UP001474421,UP001479290,UP001487740,UP001497480,UP001497497,UP001501274,UP001501940,UP001507899,UP001508024,UP001508042,UP001508043,UP001508044,UP001508046,UP001508047,UP001515400,UP001515401,UP001515402,UP001515403,UP001515404,UP001515405