RESEARCH ARTICLE SUMMARY

STRUCTURE PREDICTION

Computed structures of core eukaryotic protein complexes

Ian R. Humphreys†, Jimin Pei†, Minkyung Baek†, Aditya Krishnakumar†, Ivan Anishchenko, Sergey Ovchinnikov, Jing Zhang, Travis J. Ness, Sudeep Banjade, Saket R. Bagde, Viktoriya G. Stancheva, Xiao-Han Li, Kaixian Liu, Zhi Zheng, Daniel J. Barrero, Upasana Roy, Jochen Kuper, Israel S. Fernández, Barnabas Szakal, Dana Branzei, Josep Rizo, Caroline Kisker, Eric C. Greene, Sue Biggins, Scott Keeney, Elizabeth A. Miller, J. Christopher Fromme, Tamara L. Hendrickson, Qian Cong*‡, David Baker*‡

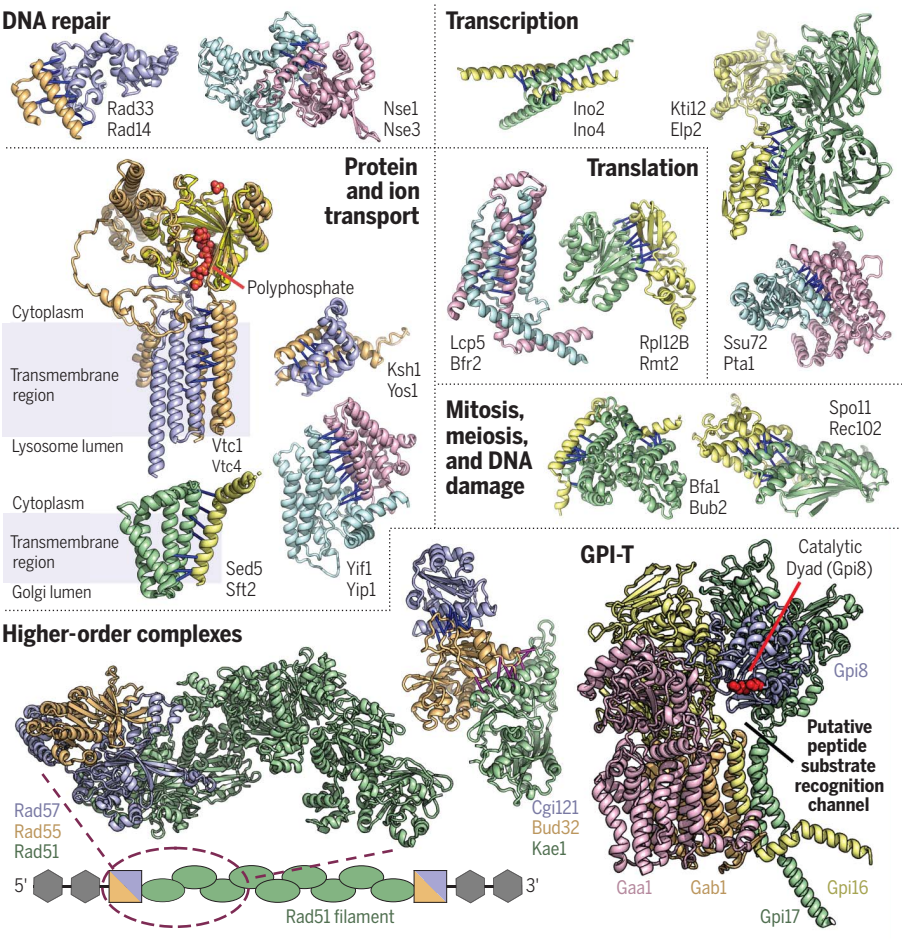
INTRODUCTION: Protein-protein interactions play critical roles in biology, but the structures of many eukaryotic protein complexes are unknown, and there are likely many interactions not yet identified. High-throughput experimental methods such as yeast two-hybrid and affinity-purification mass spectrometry have been used to identify interactions in multiple organisms, but there are inconsistencies between different datasets, and the methods do not provide high-resolution structural information. Here, we use deep learning methods to systematically identify and build structures for the protein complexes that mediate key processes in eukaryotes.

RATIONALE: Interacting proteins often coevolve, and in prokaryotes, evolutionary information can be used to identify interactions on the proteome scale at an accuracy higher than that of experimental screens. Extending this method to eukaryotes is complicated because there are fewer genome sequences available, resulting in weaker coevolutionary signals. The deep learning methods RoseTTAFold and AlphaFold, have a rich understanding of protein sequence-structure relationships, and so could help overcome this limitation.

RESULTS: We developed a coevolution-guided protein interaction identification pipeline that incorporates a rapidly computable version of RoseTTAFold with the slower but more accurate AlphaFold to systematically evaluate interactions between 8.3 million pairs of yeast proteins. RoseTTAFold alone has comparable performance in identifying protein-protein interactions to that of large-scale experimental methods; combination with AlphaFold increases identification accuracy. In total, we constructed models for 106 previously unidentified assemblies and 806 that were structurally uncharacterized.

These complexes provide rich insights into a range of biological processes from transcription, translation, and DNA repair to protein transport and modification. For example, Rad51 plays a pivotal role in DNA repair through homologous recombination, and mutations are associated with Fanconi anemia and cancer in humans. Rad55 and Rad57 are positive regulators of Rad51 assembly on single-stranded DNA. Our Rad55–Rad57–Rad51 complex model suggests that Rad55–Rad57 can bind at the 5′ end of the Rad51 single-stranded DNA filament and may stabilize the filament conformation of Rad51. Glycosylphosphatidylinositol transamidase (GPI-T) is a pentameric enzyme complex that catalyzes the attachment of GPI anchors to the C terminus of proteins. GPI-T is structurally uncharacterized, and mutations in subunits of the complex have been implicated in neurodevelopmental disorders and cancer in humans. Our model of the five-protein assembly shows that the previously identified catalytic dyad is positioned adjacent to a channel formed by three other subunits that could function in C-terminal GPI-T signal peptide recognition.

CONCLUSION: Our approach extends the range of large-scale deep learning–based structure modeling from monomeric proteins to protein assemblies. Following up on the many new interactions and complex structures should advance the understanding of a wide range of eukaryotic cellular processes and provide new targets for therapeutic intervention. Our results herald a new era of structural biology in which computation plays a fundamental role in both interaction discovery and structure determination. ■



Examples of predicted complexes.

The list of authors and their affiliations is available in the full article online.
*Corresponding author. Email: qian.cong@utsouthwestern.edu (Q.C.); dabaker@uw.edu (D.B.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as I. R. Humphreys *et al.*, *Science* **374**, eabm4805 (2021). DOI: 10.1126/science.abm4805

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abm4805

RESEARCH ARTICLE

STRUCTURE PREDICTION

Computed structures of core eukaryotic protein complexes

Ian R. Humphreys[1,2]†, Jimin Pei[3,4]†, Minkyung Baek[1,2]†, Aditya Krishnakumar[1,2]†, Ivan Anishchenko[1,2], Sergey Ovchinnikov[5,6], Jing Zhang[3,4], Travis J. Ness[7]‡, Sudeep Banjade[8], Saket R. Bagde[8], Viktoriya G. Stancheva[9], Xiao-Han Li[9], Kaixian Liu[10], Zhi Zheng[10,11], Daniel J. Barrero[12], Upasana Roy[13], Jochen Kuper[14], Israel S. Fernández[15], Barnabas Szakal[16], Dana Branzei[16,17], Josep Rizo[4,18,19], Caroline Kisker[14], Eric C. Greene[13], Sue Biggins[12], Scott Keeney[10,11,20], Elizabeth A. Miller[9], J. Christopher Fromme[8], Tamara L. Hendrickson[7], Qian Cong[3,4]*§, David Baker[1,2,21]*§

Protein-protein interactions play critical roles in biology, but the structures of many eukaryotic protein complexes are unknown, and there are likely many interactions not yet identified. We take advantage of advances in proteome-wide amino acid coevolution analysis and deep-learning–based structure modeling to systematically identify and build accurate models of core eukaryotic protein complexes within the *Saccharomyces cerevisiae* proteome. We use a combination of RoseTTAFold and AlphaFold to screen through paired multiple sequence alignments for 8.3 million pairs of yeast proteins, identify 1505 likely to interact, and build structure models for 106 previously unidentified assemblies and 806 that have not been structurally characterized. These complexes, which have as many as five subunits, play roles in almost all key processes in eukaryotic cells and provide broad insights into biological function.

Yeast two-hybrid (Y2H), affinity-purification mass spectrometry (APMS), and other high-throughput experimental approaches have identified many pairs of interacting proteins in yeast and other organisms (*1*–*5*), but there are discrepancies between sets generated using the different methods and considerable false-positive and false-negative rates (*6*–*8*). Because residues at protein-protein interfaces are expected to coevolve, the likelihood that any two proteins interact can be assessed by identifying and aligning the ortholog sequences of the two proteins in many different species, joining them to create paired multiple sequence alignments (pMSAs), and then determining the extent to which changes in the sequences of orthologs for the first protein covary with ortholog sequence changes for the second (*9*, *10*). Such amino acid coevolution has been used to guide modeling of complexes for cases in which the structures of the partners are known (*11*, *12*) and to systematically identify pairs of interacting proteins in prokaryotes with an accuracy higher than that of experimental screens (*9*). Recent deep-learning–based advances in protein structure prediction (*13*, *14*) have the potential to increase the power of such approaches as they now enable accurate modeling not only of protein monomer structures but also protein complexes (*13*).

We set out to combine proteome wide coevolution-guided protein interaction identification with deep-learning–based protein structure modeling to systematically identify and determine the structures of eukaryotic protein assemblies (Fig. 1A). We faced several challenges in directly applying to eukaryotes the statistical methods we had found effective in identifying coevolving pairs in prokaryotes (*8*). First, far fewer genome sequences are available for eukaryotes than prokaryotes: The average number of orthologous sequences (excluding nearly identical copies with >95% sequence identity) is on the order of 10,000 for bacterial proteins but 1000 for eukaryotic proteins. Thus, multiple sequence alignments for pairs of eukaryotic proteins contain fewer diverse sequences, making it more difficult for statistical methods to distinguish true coevolutionary signal from the noise. Second, eukaryotes in general have a larger number of genes, making comprehensive pairwise analysis more computationally intensive and increasing the background noise. Third, mRNA splicing in eukaryotes further increases the number of protein species, resulting in errors in gene predictions and complicating sequence alignments. Fourth, eukaryotes underwent several rounds of genome duplications in multiple lineages (*15*), and it can be difficult to distinguish orthologs from paralogs, which is important for detecting coevolutionary signal because the protein interactions of interest are likely to be conserved in orthologs in other species but less so in paralogs.

To mitigate the first three challenges, we chose to predict protein complexes for the yeast *Saccharomyces cerevisiae* as the starting point because there are a large number of fungal genomes (*16*), the genome is relatively small (6000 genes in total), and there is relatively little mRNA splicing (*17*). Furthermore, because the interactome of yeast has been extensively studied, there is a "gold standard" set (see materials and methods) of known interactions to evaluate the accuracy of predicted interactions and structures.

To distinguish orthologs from paralogs, we started from OrthoDB (*18*), a hierarchical catalog of orthologs across 1271 eukaryote genomes, and supplemented each orthologous group with sequences from 4325 eukaryote proteomes that we assembled from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/genome) and the Joint Genome Institute (*19*). Among these, 2026 are fungal proteomes spanning 14 phyla (47 classes). We compared the sequences for each protein in each of the additional 4325 proteomes against those of the most closely related species in the OrthoDB database and used the reciprocal best hit criterion (*20*) to identify orthologs (fig. S1); these were then added to the corresponding orthologous group. A complication is that each species frequently contains multiple proteins belonging to the same orthologous group, leading to ambiguity in determining which protein should be included. These multiple copies may represent alternatively spliced forms of the same gene, parts of the same gene that were split

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA. [4]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [5]Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA, USA. [6]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA, USA. [7]Department of Chemistry, Wayne State University, Detroit, MI, USA. [8]Department of Molecular Biology and Genetics, Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. [9]MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK. [10]Molecular Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [11]Gerstner Sloan Kettering Graduate School of Biomedical Sciences, New York, NY, USA. [12]Howard Hughes Medical Institute, Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [13]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [14]Rudolf Virchow Center for Integrative and Translational Bioimaging, University of Würzburg, Würzburg, Germany. [15]Department of Structural Biology, St Jude Children's Research Hospital, Memphis, TN, USA. [16]IFOM, the FIRC Institute of Molecular Oncology, Via Adamello 16, 20139, Milan, Italy. [17]Istituto di Genetica Molecolare, Consiglio Nazionale delle Ricerche (IGM-CNR), Via Abbiategrasso 207, 27100, Pavia, Italy. [18]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX, USA. [19]Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [20]Howard Hughes Medical Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [21]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.
***Corresponding author. Email: qian.cong@utsouthwestern.edu (Q.C.); dabaker@uw.edu (D.B.)**
†These authors contributed equally to this work. ‡Present address: Sanofi, Cambridge, MA, USA. §These authors contributed equally to this work.

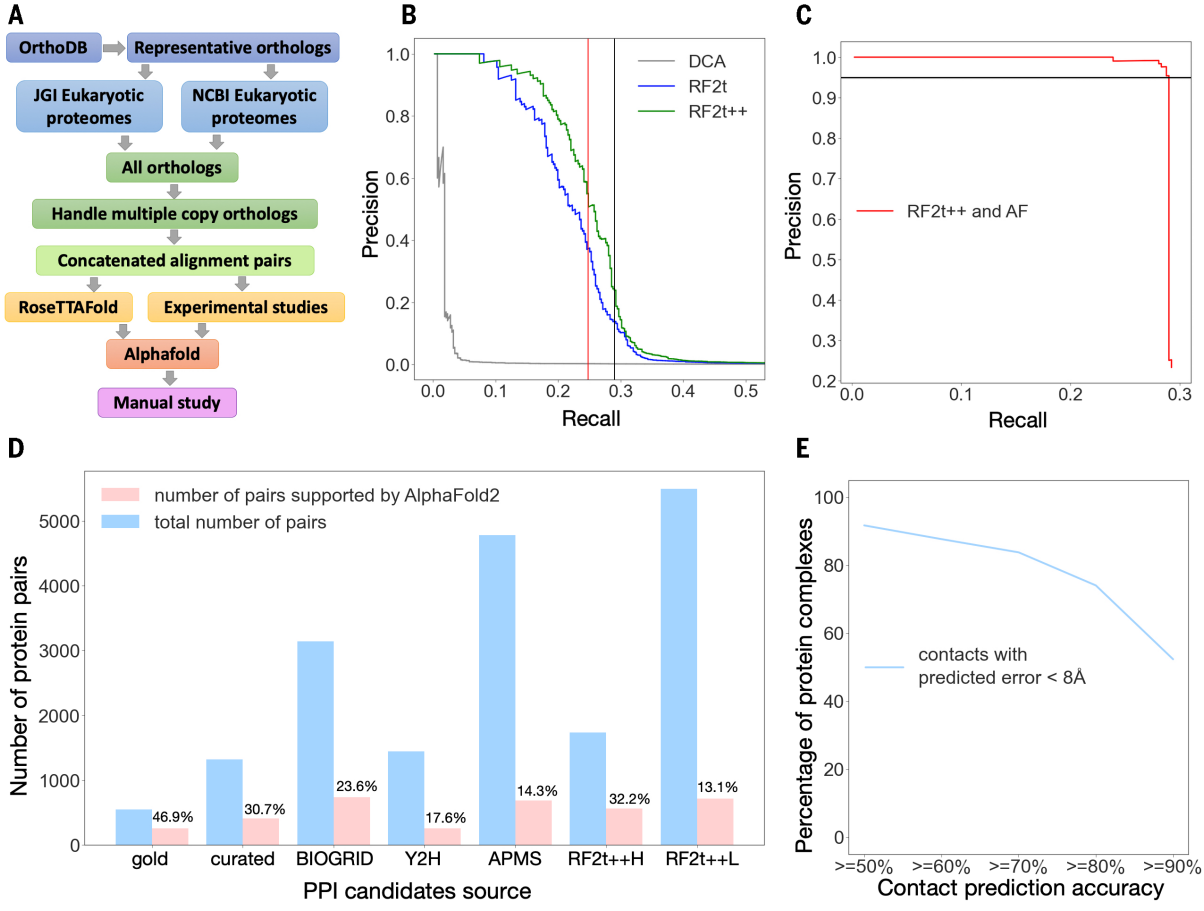


Fig. 1. Evaluation of protein interaction and structure prediction accuracy. (**A**) The PPI screen pipeline. (**B**) Performance (precision at different levels of recall) of different methods in picking out gold standard PPIs from the set of 4.3 million pMSAs [precision: number of true positives above a cutoff divided by the total number of pairs above this cutoff; recall: number of true positives above cutoff divided by the total number of true positives (gold standard PPIs)]. Pairs were ranked by the top coevolution score or contact probability between residue pairs. DCA: direct coupling analysis. RF2t: top contact probability between residues of two proteins by RF two-track model. RF2t++, optimized RF2t (see materials and methods). RF2t++ predictions better than the cutoff shown in vertical black line (RF2t++L in Fig. 1C) were processed with AF; recall of gold standard PPIs at this cutoff is 29%, and precision is 23%. RF2t++ results with a more stringent cutoff (red vertical line) are also shown in Fig. 1C (RF2t++H). (**C**) AF contact probability ranking of complexes selected by RF2t++ in (B); complexes with scores above the horizontal black line were selected for further analysis. (**D**) Number of high-scoring (top contact probability >0.67) AF predictions in PPI sets from different sources. (**E**) Distribution of percent of AF predicted interprotein contacts with predicted error <8 Å found in contact (<8 Å) in closely related experimental structures.

into multiple pieces because of errors in gene prediction, or recent gene expansions specific to certain lineages. We dealt with these possibilities by keeping only the longest isoform of each gene, merging pieces of the same gene, and selecting the copy with the highest sequence identity to single-copy orthologs in other species. For 4090 out of ~6000 yeast proteins, we were able to assign a single-copy yeast protein to orthologs in other species, and we generated pMSAs for all 4090 × 4089/2 = 8,362,005 pairwise combinations of these proteins (fig. S2). We focused on 4,286,433 pairs with alignments containing over 200 sequences to increase prediction accuracy and less than 1300 amino acids to accelerate computation (fig. S3).

In a first set of calculations, we found that even with the advantages of *S. cerevisiae* and improved ortholog identification, the statistical method (direct coupling analysis, DCA) we had used in our previous coevolution-guided protein-protein interaction (PPI) screen in prokaryotes (*9*) [the more accurate GREMLIN (*11*) method is too slow for this] could not effectively distinguish a gold standard set of 768 yeast protein pairs known to interact (*5*) (http://interactome.dfci.harvard.edu/S_cerevisiae/) from the much larger set (768,000 pairs) of primarily noninteracting pairs (Fig. 1B, gray curve, area under the curve: 0.016). Progress required a more accurate and sensitive, but still rapidly computable, method to evaluate protein interactions based on pMSAs.

We explored the application of the deep-learning–based structure prediction methods, RoseTTAFold (RF) and AlphaFold (AF), to this problem. Even though RF was originally trained on monomeric protein sequences and structures, it can accurately predict the structures of protein complexes given pMSAs with a sufficient number of sequences (*13*). We found that a lighter-weight (10.7 million parameters) RF two-track model (figs. S4 and S5) provided a good trade-off between compute time and accuracy: The model requires 11 s (about 100 times faster than AF) to process a pMSA of 1000 amino acids on a NVIDIA TITAN RTX graphic processing unit, and it can effectively distinguish gold standard PPIs among much larger sets of randomly paired proteins. The very short time required to analyze an individual pMSA made it possible to process all 4.3 million pMSAs. This method considerably outperformed DCA in distinguishing gold standard interactions from random pairs (Fig. 1B, blue curve, area under the curve:

0.219), using the highest predicted contact probability over all pairs of residues in the two proteins as a measure of the propensity for two proteins to interact (fig. S6). Performance was further improved (Fig. 1B, green curve, area under the curve: 0.248) by correcting overestimations of predicted contact probabilities between the C-terminal residues of the first protein and the N-terminal residues of the second protein, and of predicted interactions for a subset of proteins showing hub-like interactions with many other proteins (see materials and methods and figs. S7 and S8). The much better performance of RF than DCA likely stems from the extensive information on protein sequence-structure relationships embedded in the RF deep neural network; DCA, by contrast, operates solely on protein sequences with no underlying protein structure model.

We next explored whether AF residue-residue contact predictions could further distinguish interacting from noninteracting protein pairs. Like RF, AF was trained on monomeric protein structures, but given the good results with two-track RF on protein complexes and the higher accuracy of AF [also a two-track network followed by a three-dimensional (3D) structure module] on monomers, we reasoned that it might similarly have higher accuracy than RF on complexes; to enable modeling of protein complexes using AF, we modified the positional encoding in the AF script (see materials and methods). AF was too slow to be applied to the entire set of 4.3 million pMSAs [this would require 0.1 to 1 million graphics processing unit (GPU) hours]; instead we applied AF to the 5495 protein pairs with the highest RF support (indicated by the black vertical line in Fig. 1B). Using the highest AF contact probability over all residue pairs as a measure of interaction strength, we found that the combination of RF followed by AF provided excellent performance (Fig. 1C and figs. S9 and S11). Almost all the gold standard pairs were ranked higher than the negative controls, allowing selection of a set of 715 candidate PPIs with an expected precision of 95% at an AF contact probability cutoff of 0.67 (black horizontal line in Fig. 1C); we refer to this RF plus AF procedure as the de novo PPI screen, and the resulting set of predicted interactions, the de novo PPI set, below.

Owing to the trade-off between compute time and accuracy, and the necessity of setting a stringent threshold to avoid large numbers of false positives given the very large number of total pairs, we were concerned that some interacting proteins might not coevolve sufficiently to be identified robustly in our all-versus-all RF screen. Given the excellent performance of AF in distinguishing gold standard interactions among the RF filtered pairs, we also applied AF to pMSAs for PPIs reported in the literature, including those identified in high-throughput experimental screens. Similarly to our de novo PPI screen procedure, we considered protein pairs with AF contact probability larger than 0.67 to be confident interacting partners. We found that 47% of the gold standard PPIs were confidently predicted, with lower ratios (31 and 24%) for candidate PPIs from the literature (http://interactome.dfci.harvard.edu/S_cerevisiae/download/LC_multiple.txt) (*3*) or supported by low-throughput experiments according to BIOGRID (*21*) (Fig. 1D). The ratio of confidently predicted PPIs is even lower for protein pairs identified by Y2H (18%) or APMS (14%) screens (table S1), consistent with the known larger fraction of false positives in large-scale experimental screens (*8*, *22*). The fast RF two-track model used in the de novo screen has an accuracy comparable to or better than that of the large-scale experimental screens when assessed in this way: With a high-stringency RF cutoff (indicated by the red vertical line in Fig. 1B), the fraction of confidently predicted pairs among PPIs identified by RF is 32%, similar to the accuracy of low-throughput experiments; with a lower stringency cutoff (indicated by the black vertical line in Fig. 1B), this fraction becomes closer to that of the large-scale experimental screens, but somewhat fewer true PPIs are missed than with the higher cutoff (Fig. 1D).

In total, we identified 715 likely interacting pairs from the "de novo RF → AF" screen, and 1251 from the "pooled experimental sets → AF" screen, of which 461 overlap, resulting in a total of 1505 PPIs (see figs. S11 to S13 for interface size and secondary-structure distributions for the predicted complex structures). Out of these, 699 have been structurally characterized, 700 have some supporting experimental data from literature and databases, and 106 have not, to our knowledge, been previously described. To evaluate the accuracy of the predicted 3D structure of protein complexes, we used as a benchmark the 699 pairs with experimental structures in the Protein Data Bank (PDB). For 92% of these pairs, at least 50% of confident (predicted aligned error <8 Å) AF-predicted contacts are present in the experimental structures (Fig. 1E and fig. S14). The models do miss many contacts observed in the experimental structures, however, likely owing to lower residue-residue coevolution (fig. S15).

With these benchmark results providing confidence in the accuracy of the new complex interaction predictions and 3D models of the predicted complexes, we analyzed the structure models for the 806 complexes for which high-resolution structural information was not available. We classified these models into groups on the basis of their biological functions and provide examples of complexes in each functional class in Figs. 2 to 4. A first set of complexes are involved in maintenance and processing of genetic information: DNA repair, mitosis and meiosis checkpoints, transcription, and translation (Fig. 2). A second set of complexes play roles in protein translocation, transport through the secretory pathway, the cytoskeleton, and cell organelles (Fig. 3). A third set of complexes are involved in metabolism (Fig. 4). Examples of protein complexes in which proteins of unknown function are predicted to interact with well-characterized ones are shown in Fig. 4: These interactions provide hints about the function of the uncharacterized proteins and could help identify new components of previously characterized assemblies. In cases where three or more proteins were predicted to mutually interact, we generated models of the full assemblies by using as input a sequence alignment for the entire complex (see materials and methods). Examples of these larger assemblies are shown in Fig. 5; in most cases, the pairwise interactions are quite similar to those for the independently built binary complexes, but simultaneous modeling of the full complex has the advantage of allowing conformational changes that could accompany full assembly.

It is not possible to analyze the functional implications of all of the new complexes in a single paper. Instead, as an illustration of the insights that can be gained from these, we focus on a few selected examples in the following sections. To enable broader study of the functional implications of the full set of models, we have made them available at https://modelarchive.org/doi/10.5452/ma-bak-cepc and additional information is provided in the supplementary Excel file.

Complexes involved in DNA homologous recombination and repair

The homologous recombination required for accurate chromosome segregation during meiosis is initiated by DNA double-strand breaks made by Spo11 (*23*). Spo11 is essential for sexual reproduction in most eukaryotes (*24*, *25*), but mechanistic insight has been limited by a deficit of high-resolution structural information. We predict the structures of complexes of Spo11 with its essential partners Ski8 and Rec102 (Fig. 2 and figs. S16 and S17). The predicted Spo11–Ski8 structure is supported by cross-linking and mutagenesis data (*26*, *27*). Our model resembles a previous model based on the Ski3–Ski8 complex, with Ski8 contacting a sequence in Ski3 that is similar to the sequence $QREIF_{380}$ in Spo11 (*27*, *28*) (fig. S17A), but suggests a more extensive interaction surface than previously appreciated (*29*, *30*) (fig. S17, B and C). Rec102 was proposed to be a remote homolog of the transducer domain of the Top6B subunit of

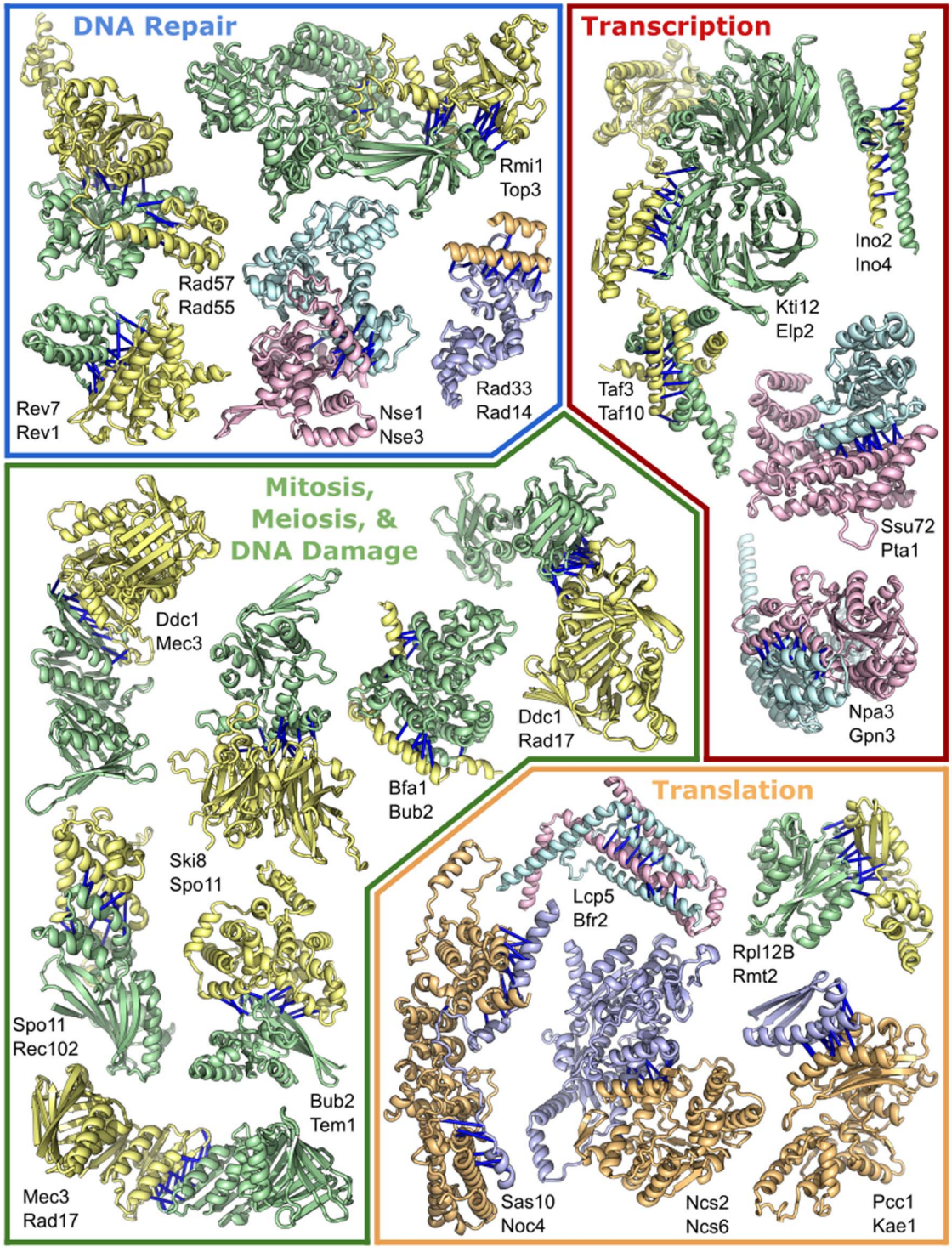


Fig. 2. Protein complexes involved in transcription, translation, and DNA repair. Top predicted residue-residue contacts are indicated with bars. Pair color indicates the method of identification: pairs from the "pooled experimental sets → AF" screen are in yellow and green, pairs from the "de novo RF → AF" screen are in blue and light orange; and pairs present in both datasets are teal and pink. Full names of these proteins are in table S2.

archaeal topoisomerase VI (*31*), which couples adenosine 5′-triphosphate–dependent dimerization of Top6B subunits to DNA cleavage by Top6A subunits (*32*). Our predicted Rec102–Spo11 complex resembles the Top6A–Top6B interface: a four-helix bundle consisting of two C-terminal helices from Rec102 and two helices from Spo11 (the first helix of the winged helix domain (WHD) plus a more N-terminally located helix) (fig. S17D). Alanine substitutions in this portion of Rec102 disrupt interaction with Spo11 and block meiotic recombination in vivo (*27*). The model clarifies the Spo11 portion of this interface, which was not well structured in previous

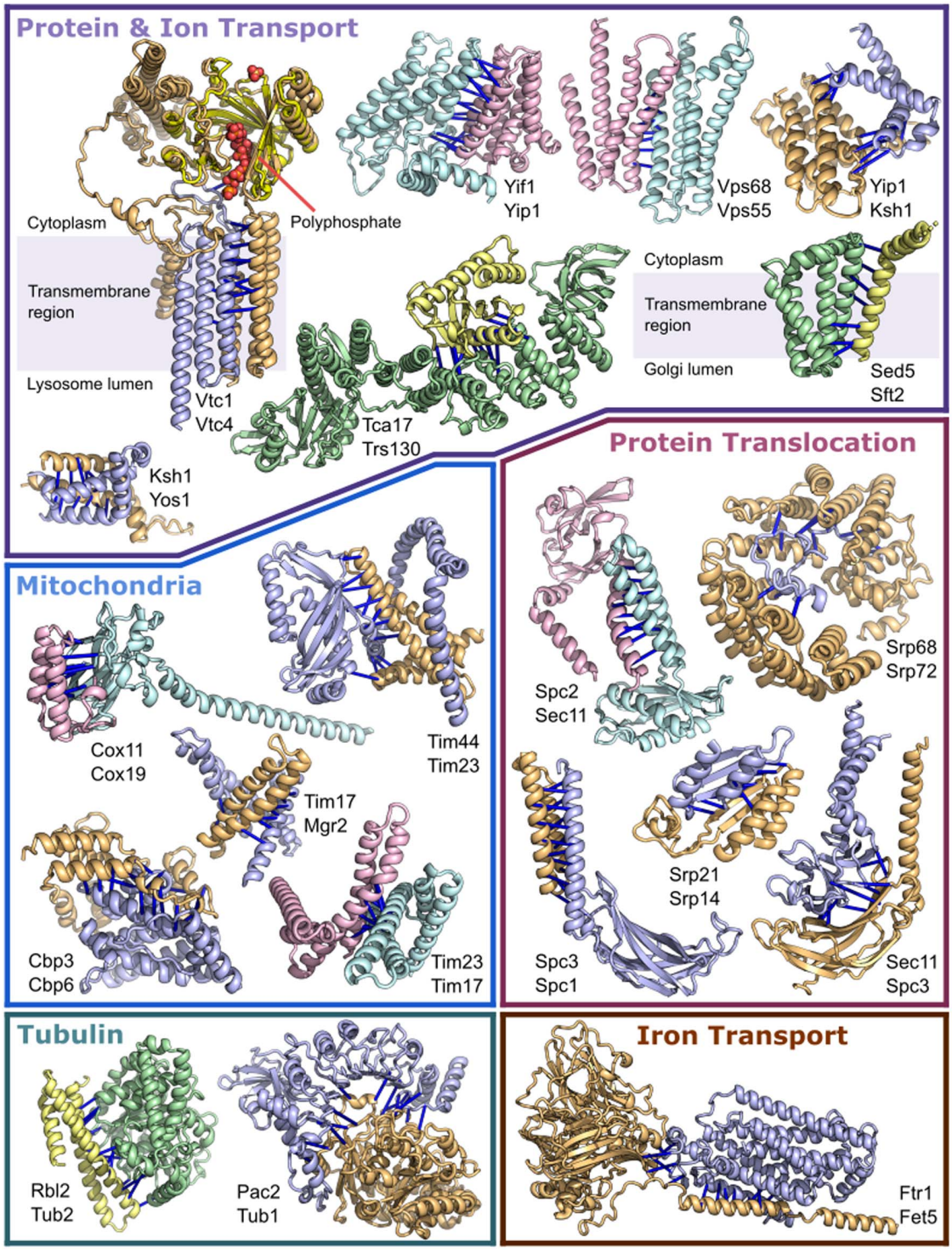


Fig. 3. Protein complexes involved in molecule transport, membrane translocation, and mitochondria. Bars and coloring as in Fig. 2. Full names for proteins are in table S3. Membrane-spanning regions are annotated on Vtc1-Vtc4 and Sed5-Sft2. Top left: model of Vtc1-Vtc4 complex, with superimposed crystal structure (PDB: 3G3Q, chain A) of the VTC4 (bright yellow) with phosphate bound (red balls).

homology models (*27*, *31*). Both Rec102 and Top6B have long, helical arms that feed into the Spo11 interface; our model predicts a different angle for this arm and contains a kink that corresponds to a conserved sequence motif $EYPMVF_{192}$ in *Saccharomyces* that is missing in both archaeal TopoVI and mammals (fig. S17, D and E). Mutations in this region can suppress *rec104* conditional alleles (*33*), suggesting that this part of Rec102 is important for integrating Rec104 function into the Spo11 core complex.

The highly conserved Rad51 protein, which is central to DNA repair, carries out key reactions during homologous recombination, and mutations in human paralogs are associated with Fanconi anemia and multiple types of cancer (*34*). Rad51 paralogs can be positive regulators

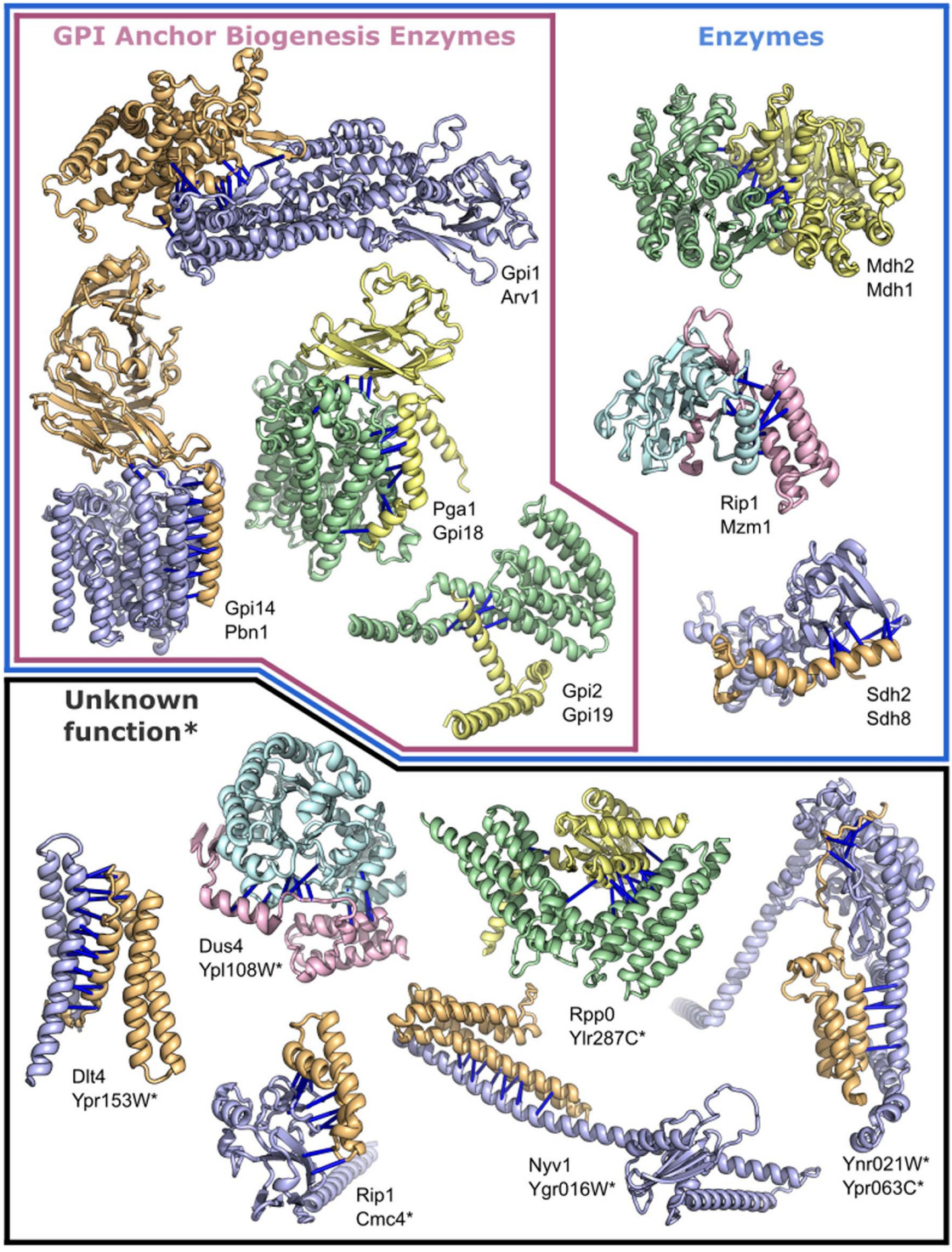


Fig. 4. Protein complexes involved in metabolism, GPI anchor biosynthesis, or including a protein of unknown function. Coloring is as in Figs. 2 and 3. Proteins annotated in the Uniprot database as uncharacterized proteins are denoted with an asterisk. Full names for these proteins are in table S4.

of Rad51 activity (*35*); in yeast, the Rad51 paralogs Rad55 and Rad57 form a stable homodimer that accelerates assembly of Rad51 filaments on single-stranded DNA (ssDNA) during homologous recombination through a transient interaction with Rad51 (*36*). The lack of structural data for the Rad55–Rad57 complex and its interface with Rad51 has limited mechanistic understanding of this process. We generated a model of the trimeric Rad55–Rad57–Rad51 complex, which in combination with the known Rad51 filament structure (*37*), suggests that Rad55–Rad57 binds at the 5′ end of the Rad51 filament where it could promote growth of the Rad51 filament in a directional manner (Fig. 5B and fig. S18).

Nucleotide excision repair (NER) requires a search for lesions in DNA that is mediated

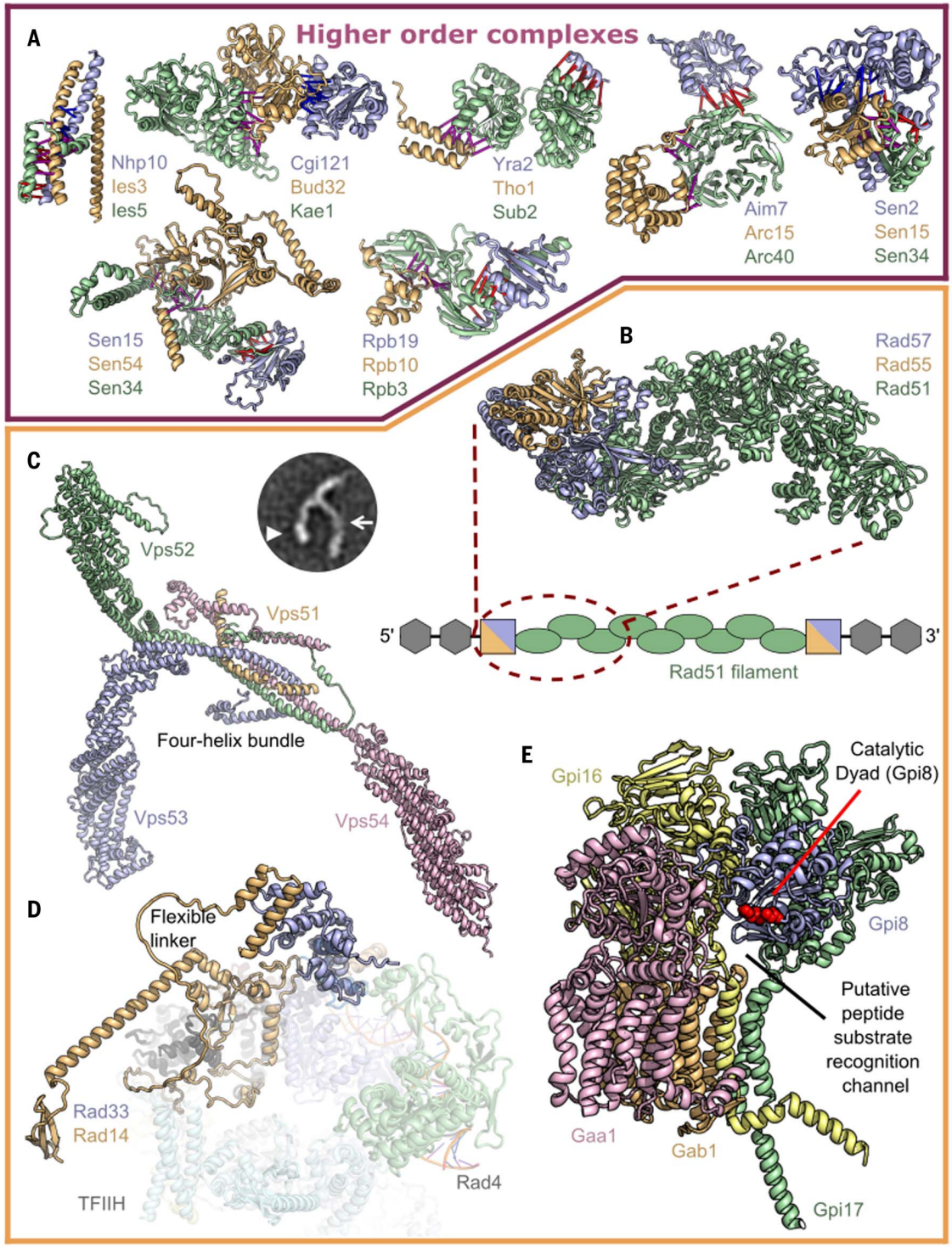


Fig. 5. Higher-order protein complexes. (**A**) Top predicted residue-residue contacts for trimers are indicated with bars. Bar color corresponds to the interacting protein pair; protein 1:2 are blue, 1:3 are red, 2:3 are purple. Full names of each protein within the complex are in table S5. (**B**) Model of Rad55–Rad57–Rad51 and cartoon depiction of placement of this complex in the larger Rad51 filament. Additional information is in fig. S18. (**C**) GARP complex model constructed by predicting structure of central hetero-oligomeric helical bundle, and superimposing models of individual components onto this. 2D class average of GARP complex with minor adaptation (*77*); reprinted by permission from Springer Nature Customer Service Center GmbH. Alternative GARP models are in fig. S24. (**D**) Rad33–Rad14 complex model superimposed onto previously determined TFIIH/Rad4–Rad23–Rad33 complex structure (7k04). See fig. S19 for additional details. (**E**) GPI-T pentamer model highlighting a possible peptide substrate recognition channel adjacent to the catalytic dyad. See fig. S27 for additional details.

by a conserved complex containing Rad4 (XPC—Xeriderma pigmentosum group C protein in humans), Rad23 (HR23B), and Rad33 (Centrin2) in yeast. The Rad4–Rad23–Rad33 complex is essential for global genome NER and is the major player in initial damage recognition (*38*). Rad14 (XPA) is recruited at a later stage and activates the helicase Rad3 (XPD) subunit of the general transcription and DNA repair factor IIH complex (TFIIH, consisting of Rad3, Ssl2, Ssl1, Tfb1, Tfb2, Tfb4, and Tfb5) through the release of the TFIIK (CAK) complex following interactions with the TFIIH subunits Tfb5 (p8) and Ssl2 (XPB), and double-stranded DNA (*39*). The structures of Rad14 that are currently available only comprise the extended DNA binding domain and lack the N and C terminus, where the latter interacts with Tfb5. We generated a model of the complex between full-length Rad14 and Rad33 that resolves much of the current structural ambiguity in this system (Fig. 2 and fig S19B), shedding light on how Rad14 may be recruited to the Rad4–Rad23–Rad33 complex. Placing this model into a cryo–electron microscopy (EM) map comprising XPA (Rad14) and TFIIH bound to DNA

(39) suggests how the Rad14 C terminus, which fits into previously unmodeled density, interacts with TFIIH. The long central helix observed in the Centrin2 (Rad33) structure (40) is kinked about 90° in our Rad33–Rad14 complex model (fig. S19B); both conformations are feasible and are compatible for the interaction with Rad14. In a recent cryo-EM structure of the TFIIH/Rad4–Rad23–Rad33 initial recognition complex (41), only the C- terminal part of Rad33 was determined. Superposition of Rad33 in the Rad33–Rad14 complex model onto this structure (Fig. 5D) shows how Rad14 can interact with the Rad4–Rad23–Rad33 recognition complex (38, 42) while maintaining the TFIIH interaction, bridging the steps of initial damage recognition and damage verification. Our model suggests that Rad14 and Rad4 can be present at the same time in the repair cascade; cross-talk between these important proteins could modulate downstream events.

Complexes involved in translation and ribosome regulation

Throughout evolution, the eukaryotic machinery for protein production has expanded in size and complexity (43), which facilitated the development of sophisticated mechanisms for the regulation of gene expression at the post-transcriptional level (44) and increased integration with the cellular environment (45). The expanded complexity of the eukaryotic translational machinery came at the cost of a highly complex process for ribosome maturation (46). We generate models of complexes that had not been structurally characterized previously that involve components of the translation apparatus (Fig. 2 and fig. S20). Two complexes, Rpl12B–Rmt2 and Rpl7A–Fpr4, involving enzymes that introduce protein modifications such as arginine methylations or proline isomerization (47), provide insight into mechanisms that expand the chemical diversity of ribosomal proteins at functional sites (48) and possibly regulate translation (49). A complex between components of the U3 ribosome-maturation factor and a protein involved in the regulation of glycerol, Lcp5–Sgd1 (50), could play a role in coupling translation with metabolism. A complex between eukaryotic initiation factor 2B (eIF2B), an auxiliary factor for eIF2 recycling after guanosine 5′-triphosphate hydrolysis, and transcriptional factor regulator Dig2 could help couple translation and transcription: The delivery of the first aminoacyl-tRNA (Met-tRNA$_i^{Met}$) is a key event in eukaryotic translation regulation by the GTPase eIF2 (51), and targeting eIF2 through its nucleotide exchanger eIF2B is a basal mechanism of translation regulation. This possible cross-talk between ribosome-maturation pathways and metabolic sensors, and translation initiation regulators such as eIF2, with transcription factors suggests exciting new avenues to further map the highly integrated nature of translation within eukaryotic cells.

Complexes involving ubiquitin and small ubiquitin-like modifier (SUMO) ligases

Reversible covalent modifications of proteins with ubiquitin and SUMO modulate protein-protein interactions, cellular localization, and stability (52). SUMO E3 ligases facilitate SUMO transfer, and Siz1, Siz2, Mms21, and Zip3 are the known SUMO ligases in budding yeast (52). Our model of the Siz2 and Mms21 SUMO ligase complex (fig. S21A) suggests that both E3s could act jointly to modify DNA-associated substrates, perhaps through the DNA binding SAP domain of Siz2 (53) or involving the Mms21 (Nse2)–containing Smc5–6 complex, which modulates DNA recombination, replication, and repair (54, 55). The Smc5–6 complex contains another RING-finger E3 ligase–like subunit, Nse1 (56), that interacts with Nse3 and Nse4. Our model of the yeast Nse1–Nse3–Nse4 complex (fig. S21B) is similar to a structure determined for the *Xenopus laevis* complex, despite the sequences of the yeast and *Xenopus* proteins being too distant for similarity to be detectable by BLAST.

SUMO-targeted ubiquitin ligases (STUbLs) are ubiquitin ligases that recognize SUMO-modified proteins. A STUbL consisting of the Slx8 ubiquitin ligase and the associated protein Slx5 functions in proteasome-mediated turnover of several proteins associated with DNA replication, repair, and chromosome structure (57–59). Our model of the Slx5-Slx8 complex (fig. S21C) provides insight into how these two proteins may collectively recognize their substrates. In addition, we generated a lower-confidence but intriguing model of a previously undescribed complex between Slx8 and Cue3 [coupling of ubiquitin conjugation to endoplasmic reticulum (ER) degradation protein 3] (fig. S21D), possibly linking ubiquitination of substrates to protein degradation in ER.

Complexes involved in chromosome segregation

The heterodecameric complex DASH/Dam1 (Dam1c) is composed of 10 proteins—Ask1, Dad1, Dad2, Dad3, Dad4, Dam1, Duo1, Hsk3, Spc19, and Spc34—which come together to form a "T" shape and can further oligomerize into rings (60, 61). During mitosis, these heterodecamers strengthen the attachment between kinetochores and microtubules (62) by oligomerizing to form either partial or complete rings around microtubules and further contacting kinetochore components (63–65). Microtubules are required for in vivo ring formation, but a structure of the Dam1c ring complex from *Chaetomium thermophilum* was determined in the absence of microtubules using monovalent salts (66). We generated structure models of nine binary complexes (Dad2–Ask1, Dad2–Hsk3, Dad2–Spc1, Dad4–Hsk3, Dam1–Duo1, Duo1–Dad1, Spc19–Dad1, Spc34–Duo1, and Spc34–Spc19) that encompass several members of Dam1c (fig. S22). These complexes are largely consistent with the Dam1c structure, suggesting that the findings from the thermophile structure can likely be extended to *S. cerevisiae*. We went beyond previous structural data by predicting the structure of a potential interdecamer interaction between a loop on Spc19 and the N terminus of Dad1, which could be important for ring formation in vivo (66).

Complexes involved in molecule transport and membrane trafficking

The small-membrane protein Ksh1 is essential for growth and conserved across eukaryotes, and plays an unknown role in protein secretion (67). We predicted structures of complexes between Ksh1 and two membrane proteins reported to form a complex: Yos1 and Yip1. This complex also includes Yif1 and interacts with Rab GTPases (68) (Fig. 3). These structures suggest that Ksh1 is a fourth member of this enigmatic complex that is essential to the secretory pathway and explain how Ksh1 can play a role in secretion despite its small size of 72 amino acids.

The vacuolar transporter chaperone (VTC) is a five-subunit complex that synthesizes polyphosphate to regulate cellular phosphate concentrations (69). Structures are only known for some soluble portions of this complex, including the catalytic domain of the Vtc4 subunit (70). Our model of the previously not structurally characterized Vtc1–Vtc4 subcomplex suggests that the cytosolic active site is positioned by the complex to feed the polyphosphate product through a membrane pore into the lumen of the lysosome (Fig. 3).

The ESCRT-III complex is involved in a number of cellular membrane remodeling pathways, including receptor down-regulation, membrane repair, and cell division (71, 72). Our predicted interface between the Vps2 and Vps24 subunits of the ESCRT-III complex resembles the polymerization interface of a different ESCRT-III subunit, Snf7 (73), providing insight into the roles of these previously uncharacterized ESCRT-III subunits and highlighting the generality of this mode of interaction in ESCRT-III complexes. Notably, previously unpublished mutations (fig. S23) in Vps24 that prevent ESCRT function in multivesicular body sorting are located on the predicted interface between Vps2 and Vps24, supporting our model and the functional importance of the Vps2–Vps24 interaction. Vps55 and Vps68 are conserved membrane proteins that are important for endosomal cargo sorting; our

predicted structure (Fig. 2) of their interaction provides clues about the mechanism of their function (*74*).

The GARP complex is a multisubunit tethering complex (MTC) that mediates docking and fusion of vesicles with the Golgi apparatus (*75*). Our approach generated models for binary complexes involving the four GARP subunits, and we further modeled the entire complex (fig. S24A). In this model, the four subunits assemble through a four-helix bundle. In each of the three larger subunits, Vps52, Vps53, and Vps54, C-terminal domains comprising "CATCHR" folds emanate from the bundle. This architecture resembles portions of the cryo-EM structure of the Exocyst complex, a distinct MTC that mediates fusion of vesicles at the plasma membrane (*76*), which possesses two separate four-helix bundles organizing its eight subunits. In our prediction, the "CATCHR" domains appear to be somewhat flexibly linked to the central four-helix bundle, and hence we overlaid the structure predictions for Vps52, Vps53, and Vps54, respectively, onto the central four-helix bundle (Fig. 5C and fig. S24B). The resulting model has a marked resemblance to previously published 2D classes (fig. S24C) from a negative-stain EM analysis of the GARP complex (*77*). These predictions will facilitate structure-guided experiments to elucidate the mechanism of MTC function.

Golgi-resident protein, Grh1, forms a tethering complex with Uso1 and Bug1 that interacts with the COPII coat protein complex, Sec23–Sec24. The tether is thought to participate in COPII vesicle capture (*78*, *79*), but the mechanism remains unclear. The C terminus of Grh1 contains a predicted intrinsically disordered region (IDR) with a net positively charged cluster and a triple-proline motif (fig. S25, A and B). Our model of the Sec23–Grh1 complex contains an interface between the Sec23 gelsolin domain and the PPP motif of Grh1 (*80*), and an interface between the Grh1 IDR and Sec23 involving a disorder-to-helical transition (fig. S25C). A similar multivalent interface also drives interaction between Sec23 and the COPII coat scaffolding protein, Sec31 (*81*). Our model suggests that the combinatorial multivalent interaction between Grh1 and Sec23 may compete with the interaction between Sec31 and Sec23 to promote vesicle uncoating; consistent with this model, Grh1 is recruited to glutathione *S*-transferase (GST)–Sec23, dependent on the IDR, and competes for Sec31 binding (fig. S25D).

SNARE [soluble *N*-ethylmaleimide-sensitive factor (NSF) attachment protein (SNAP) receptor] proteins drive intracellular membrane fusion between transport vesicles and organelles (*82*). Our predicted complex structure between the SNARE Sed5 and the uncharacterized transmembrane protein Sft2 unexpectedly predicted an interaction between transmembrane domains of the two proteins (Fig. 3). SNARE localization is thought to occur through interactions of cytoplasmic domains with cytoplasmic sorting factors, but this prediction, together with genetic evidence (*83*), suggests that SNARE localization or function may be subject to additional mechanisms through interactions with transmembrane protein regulators. Membrane fusion requires the formation of a four-helix bundle (called the SNARE complex) between the vesicle SNARE and the target membrane SNAREs (*84*, *85*). The bundle is formed by the SNARE motifs, which are 60 to 70 amino acids with heptad repeats and the ability to form coiled-coil structures. Models of binary complexes of SNARE-motif–containing proteins frequently differ from their classic conformation in the SNARE four-helical bundle (fig. S26A), probably because all four chains are required to form the stable complex (*86*). Indeed, modeling the four SNARE proteins (Ufe1, Use1, Sec20, and Sec22) that are known to mediate the fusion between Golgi-derived retrograde transport vesicles with ER (*87*) together resulted in a complex that resembles a typical SNARE complex (*84*) (fig. S26, B and C). This example highlights the potential pitfalls of modeling only binary complexes when the functional assembly involves more than two chains.

GPI transamidase complex

Glycosylphosphatidylinositol transamidase (GPI-T) is a pentameric enzyme complex of unknown structure (*88*–*90*) that catalyzes the attachment of GPI anchors to the C terminus of specific substrate proteins, based on recognition of a C-terminal signal peptide (*91*). GPI-T catalyzes the removal of this signal sequence, replacing it with a new amide bond to an ethanolamine phosphate in the GPI anchor. The five subunits of *S. cerevisiae* GPI-T are Gpi8 (which contains the catalytic active site), Gpi16, Gaa1, Gpi17, and Gab1 (*88*, *92*, *93*). Our large-scale modeling approach generated models for the following binary complexes: Gpi8–Gpi17, Gab1–Gaa1, Gab1–Gpi17, and Gaa1–Gpi16. We subsequently modeled the full-length, pentameric GPI-T in one shot, starting from the sequences of all components (Fig. 5E). Several features of this model are consistent with previous characterization of this enzyme. *S. cerevisiae* GPI-T can be purified as a core heterotrimer, containing only Gpi8, Gpi16, and Gaa1 (*92*); our GPI-T model confirms extensive interactions between the soluble domains of these three subunits. This model also recapitulates the disulfide bond between Gpi8 (Cys^{85}) and Gpi16 (Cys^{202}), previously characterized for human GPI-T (*94*) [the existence of this disulfide bond in yeast GPI-T has been called into question (*90*)]. Gaa1 is essential for binding of the GPI anchor to GPI-T (*95*), and the hydrophobic Gab1 is also predicted to participate in anchor recognition (*88*). Our model positions the transmembrane regions of Gaa1 and Gab1 against each other. The catalytic dyad in Gpi8 (Cys^{199} and His^{157}) faces these transmembrane domains, and abuts a highly conserved face of Gaa1, proposed to recognize the GPI anchor glycans (*96*, *97*). In our model, the positions of these subunits are consistent with binding of the GPI anchor to position its modifying amine in the Gpi8 active site for catalysis. Gpi16 is immediately adjacent to these interactions and is likely involved in anchor recognition. The functional role of Gpi17 has been elusive, but our model now suggests that Gpi17, together with Gpi8 and Gpi16, forms a recognition channel for the C-terminal GPI-T signal peptide (fig. S27) adjacent to the catalytic dyad (Fig 5E). In vivo, GPI-T is expected to be a dimer of pentamers, with dimerization occurring on one face of the caspase-like Gpi8 subunit (*92*, *97*, *98*). This decameric complex was too large for us to model computationally; however, the pentameric complex that we present here leaves open the dimerization face of Gpi8 consistent with probable dimerization. It also suggests that Gaa1 and Gpi17 would participate in dimerization of this enzyme. In humans, mutations in GPI-T subunits are associated with neurodevelopmental disorders (*99*). Each subunit contributes to different cancer mechanisms, in some cases by perturbing GPI anchoring of specific receptors and in others by separating from GPI-T to alter disparate signal transduction pathways (*89*). Now, with a structural model in hand, these mechanisms can be examined at a molecular level.

Limitations of the current method

As with any new method, it is important when interpreting the results (our large set of predicted complex structures) to keep in mind the limitations of the approach. First, our study is not comprehensive, so conclusions should not be drawn about absences; in particular, we eliminated proteins that arose from recent duplication owing to difficulty in identifying orthologs in other organisms, and thus only surveyed two-thirds of the entire yeast proteome. Second, the approach likely misses interactions restricted to a small set of organisms, or that vary rapidly during evolution, owing to weaker coevolutionary signals. Third, the approach likely works less well for transient interactions that generally involve smaller and weaker interfaces, which may be under lower selective pressure, in particular those involving intrinsically disordered regions, which are poorly represented in the PDB. The majority of known interactions identified by our approach are likely obligate assemblies and involve ordered structural elements. Fourth, interactions between single hydrophobic or amphipathic

helices, such as single transmembrane helices or coiled coils, may be overpredicted (in initial studies of human complexes, interactions solely between single-pass transmembrane regions appear to be over-represented). Fifth, and perhaps most important, for proteins that form high-order obligate protein complexes, binary complex models may be quite inaccurate, as illustrated by the SNARE example.

Conclusion

Our approach extends the range of large-scale deep-learning–based structure modeling from monomeric proteins to protein assemblies. As highlighted by the above examples, following up on the many new complexes presented here should advance understanding of a wide range of eukaryotic cellular processes and provide new targets for therapeutic intervention. The methods can be extended directly to large-scale mapping of interactions in the human proteome, but considerably more compute time will be required given the much larger total number of protein pairs, and models may be somewhat less accurate owing to weaker coevolutionary signal for the subset of human proteins specific to higher eukaryotes and for the many closely related paralogs arising from gene duplication. Investigating interactions of individual proteins or subsets of proteins— for example, deorphanization of orphan receptors—should be immediately accessible using our approach provided there are sufficient sequence homologs. Training RF and AF on protein complexes should further improve performance of both methods (*100*), particularly for protein pairs with fewer homologs and/or weaker and more transient interactions, and reduce the dependence on ortholog identification. Together with the advances in monomeric structure prediction, our results herald a new era of structural biology in which computation plays a fundamental role in both interaction discovery and structure determination.

Methods

As described in detail in the supplementary materials and methods, we developed a multistep bioinformatics and deep learning pipeline for identifying pairs of proteins likely to interact and modeling the 3D structures of the corresponding protein complexes. The steps of this pipeline are illustrated schematically in Fig. 1A. First, comprehensive orthologous groups of genes were generated and yeast genes were mapped to these groups; second, multiple sequence alignments of orthologous sequences were generated for each pair of yeast proteins; third, contact probability was computed for each protein pair using RoseTTAFold; and fourth, interaction probability was reevaluated, and complex structures were modeled using AlphaFold. The experimental data-guided PPI screening pipeline is very similar except that in the third stage, instead of using RoseTTAFold, we used experimental data primarily derived from large-scale screens to identify PPI candidates.

REFERENCES AND NOTES

1. T. Ito *et al*., A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4569–4574 (2001). doi: 10.1073/pnas.061034498; pmid: 11283351
2. S. R. Collins *et al*., Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics* **6**, 439–450 (2007). doi: 10.1074/mcp.M600381-MCP200; pmid: 17200106
3. T. Reguly *et al*., Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J. Biol.* **5**, 11 (2006). doi: 10.1186/jbiol36; pmid: 16762047
4. P. Uetz *et al*., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623–627 (2000). doi: 10.1038/35001009; pmid: 10688190
5. H. Yu *et al*., High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008). doi: 10.1126/science.1158684; pmid: 18719252
6. O. Kuchaiev, M. Rasajski, D. J. Higham, N. Przulj, Geometric de-noising of protein-protein interaction networks. *PLOS Comput. Biol.* **5**, e1000454 (2009). doi: 10.1371/journal.pcbi.1000454; pmid: 19662157
7. A. M. Edwards *et al*., Bridging structural biology and genomics: Assessing protein interaction data with known complexes. *Trends Genet.* **18**, 529–536 (2002). doi: 10.1016/S0168-9525(02)02763-4; pmid: 12350343
8. J. P. Mackay, M. Sunde, J. A. Lowry, M. Crossley, J. M. Matthews, Protein interactions: Is seeing believing? *Trends Biochem. Sci.* **32**, 530–531 (2007). doi: 10.1016/j.tibs.2007.09.006; pmid: 17980603
9. Q. Cong, I. Anishchenko, S. Ovchinnikov, D. Baker, Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019). pmid: 31296772
10. A. G. Green *et al*., Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* **12**, 1396 (2021). doi: 10.1038/s41467-021-21636-z; pmid: 33654096
11. S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014). doi: 10.7554/eLife.02030; pmid: 24842992
12. T. A. Hopf *et al*., Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014). doi: 10.7554/eLife.03430; pmid: 25255213
13. M. Baek *et al*., Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021). doi: 10.1126/science.abj8754; pmid: 34282049
14. J. Jumper *et al*., Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi: 10.1038/s41586-021-03819-2; pmid: 34265844
15. A. Meyer, M. Schartl, Gene and genome duplications in vertebrates: The one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* **11**, 699–704 (1999). doi: 10.1016/S0955-0674(99)00039-3; pmid: 10600714
16. I. V. Grigoriev *et al*., MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42** (D1), D699–D704 (2014). doi: 10.1093/nar/gkt1183; pmid: 24297253
17. M. Spingola, L. Grate, D. Haussler, M. Ares Jr., Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *RNA* **5**, 221–234 (1999). doi: 10.1017/S1355838299981682; pmid: 10024174
18. E. M. Zdobnov *et al*., OrthoDB in 2020: Evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **49** (D1), D389–D393 (2021). doi: 10.1093/nar/gkaa1009; pmid: 33196836
19. A. Clum *et al*., DOE JGI Metagenome Workflow. *mSystems* **6**, e00804-20 (2021). doi: 10.1128/mSystems.00804-20; pmid: 34006627
20. D. P. Wall, H. B. Fraser, A. E. Hirsh, Detecting putative orthologs. *Bioinformatics* **19**, 1710–1711 (2003). doi: 10.1093/bioinformatics/btg213; pmid: 15593400
21. R. Oughtred *et al*., The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021). doi: 10.1002/pro.3978; pmid: 33070389
22. H. Huang, B. M. Jedynak, J. S. Bader, Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLOS Comput. Biol.* **3**, e214 (2007). doi: 10.1371/journal.pcbi.0030214; pmid: 18039026
23. S. Keeney, C. N. Giroux, N. Kleckner, Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88**, 375–384 (1997). doi: 10.1016/S0092-8674(00)81876-0; pmid: 9039264
24. B. de Massy, Initiation of meiotic recombination: How and where? Conservation and specificities among eukaryotes. *Annu. Rev. Genet.* **47**, 563–599 (2013). doi: 10.1146/annurev-genet-110711-155423; pmid: 24050176
25. H. Murakami, S. Keeney, Regulating the formation of DNA double-strand breaks in meiosis. *Genes Dev.* **22**, 286–292 (2008). doi: 10.1101/gad.1642308; pmid: 18245442
26. C. Arora, K. Kee, S. Maleki, S. Keeney, Antiviral protein Ski8 is a direct partner of Spo11 in meiotic DNA break formation, independent of its cytoplasmic role in RNA metabolism. *Mol. Cell* **13**, 549–559 (2004). doi: 10.1016/S1097-2765(04)00063-2; pmid: 14992724
27. C. Claeys Bouuaert *et al*., Structural and functional characterization of the Spo11 core complex. *Nat. Struct. Mol. Biol.* **28**, 92–102 (2021). doi: 10.1038/s41594-020-00534-w; pmid: 33398171
28. F. Halbach, P. Reichelt, M. Rode, E. Conti, The yeast ski complex: Crystal structure and RNA channeling to the exosome complex. *Cell* **154**, 814–826 (2013). doi: 10.1016/j.cell.2013.07.017; pmid: 23953113
29. S. Steiner, J. Kohli, K. Ludin, Functional interactions among members of the meiotic initiation complex in fission yeast. *Curr. Genet.* **56**, 237–249 (2010). doi: 10.1007/s00294-010-0296-0; pmid: 20364342
30. S. Tessé, A. Storlazzi, N. Kleckner, S. Gargano, D. Zickler, Localization and roles of Ski8p protein in Sordaria meiosis and delineation of three mechanistically distinct steps of meiotic homolog juxtaposition. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12865–12870 (2003). doi: 10.1073/pnas.2034282100; pmid: 14563920
31. T. Robert *et al*., The TopoVIB-Like protein family is required for meiotic DNA double-strand break formation. *Science* **351**, 943–949 (2016). doi: 10.1126/science.aad5309; pmid: 26917764
32. K. D. Corbett, P. Benedetti, J. M. Berger, Holoenzyme assembly and ATP-mediated conformational dynamics of topoisomerase VI. *Nat. Struct. Mol. Biol.* **14**, 611–619 (2007). doi: 10.1038/nsmb1264; pmid: 17603498
33. L. Salem, N. Walter, R. Malone, Suppressor analysis of the Saccharomyces cerevisiae gene REC104 reveals a genetic interaction with REC102. *Genetics* **151**, 1261–1272 (1999). doi: 10.1093/genetics/151.4.1261; pmid: 10101155
34. M. R. Sullivan, K. A. Bernstein, RAD-ical New Insights into RAD51 Regulation. *Genes* **9**, 629 (2018). doi: 10.3390/genes9120629; pmid: 30551670
35. J. San Filippo, P. Sung, H. Klein, Mechanism of eukaryotic homologous recombination. *Annu. Rev. Biochem.* **77**, 229–257 (2008). doi: 10.1146/annurev.biochem.77.061306.125255; pmid: 18275380
36. U. Roy *et al*., The Rad51 paralog complex Rad55-Rad57 acts as a molecular chaperone during homologous recombination. *Mol. Cell* **81**, 1043–1057.e8 (2021). doi: 10.1016/j.molcel.2020.12.019; pmid: 33421364
37. A. B. Conway *et al*., Crystal structure of a Rad51 filament. *Nat. Struct. Mol. Biol.* **11**, 791–796 (2004). doi: 10.1038/nsmb795; pmid: 15235592
38. K. Sugasawa, J. Akagi, R. Nishi, S. Iwai, F. Hanaoka, Two-step recognition of DNA damage for mammalian nucleotide excision repair: Directional binding of the XPC complex and DNA strand scanning. *Mol. Cell* **36**, 642–653 (2009). doi: 10.1016/j.molcel.2009.09.035; pmid: 19941824
39. G. Kokic *et al*., Structural basis of TFIIH activation for nucleotide excision repair. *Nat. Commun.* **10**, 2885 (2019). doi: 10.1038/s41467-019-10745-5; pmid: 31253769
40. J. R. Thompson, Z. C. Ryan, J. L. Salisbury, R. Kumar, The structure of the human centrin 2-xeroderma pigmentosum group C protein complex. *J. Biol. Chem.* **281**, 18746–18752 (2006). doi: 10.1074/jbc.M513667200; pmid: 16627479
41. T. van Eeuwen *et al*., Cryo-EM structure of TFIIH/Rad4-Rad23-Rad33 in damaged DNA opening in nucleotide excision

repair. *Nat. Commun.* **12**, 3338 (2021). doi: 10.1038/s41467-021-23684-x; pmid: 34099686
42. T. Riedl, F. Hanaoka, J.-M. Egly, The comings and goings of nucleotide excision repair factors on damaged DNA. *EMBO J.* **22**, 5293–5303 (2003). doi: 10.1093/emboj/cdg489; pmid: 14517266
43. S. Klinge, F. Voigts-Hoffmann, M. Leibundgut, N. Ban, Atomic structures of the eukaryotic ribosome. *Trends Biochem. Sci.* **37**, 189–198 (2012). doi: 10.1016/j.tibs.2012.02.007; pmid: 22436288
44. A. G. Hinnebusch, I. P. Ivanov, N. Sonenberg, Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016). doi: 10.1126/science.aad9868; pmid: 27313038
45. J. A. Saba, K. Liakath-Ali, R. Green, F. M. Watt, Translational control of stem cell function. *Nat. Rev. Mol. Cell Biol.* **22**, 671–690 (2021). doi: 10.1038/s41580-021-00386-2; pmid: 34272502
46. S. Klinge, J. L. Woolford Jr.., Ribosome assembly coming into focus. *Nat. Rev. Mol. Cell Biol.* **20**, 116–131 (2019). doi: 10.1038/s41580-018-0078-y; pmid: 30467428
47. K. M. Mulvaney *et al*., Molecular basis for substrate recruitment to the PRMT5 methylosome. *Mol. Cell* **81**, 3481–3495.e7 (2021). doi: 10.1016/j.molcel.2021.07.019; pmid: 34358446
48. Z. L. Watson *et al*., Structure of the bacterial ribosome at 2 Å resolution. *eLife* **9**, e60482 (2020). doi: 10.7554/eLife.60482; pmid: 32924932
49. J. M. Małecki *et al*., Human METTL18 is a histidine-specific methyltransferase that targets RPL3 and affects ribosome biogenesis and function. *Nucleic Acids Res.* **49**, 3185–3203 (2021). doi: 10.1093/nar/gkab088; pmid: 33693809
50. F. Dragon *et al*., A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* **417**, 967–970 (2002). doi: 10.1038/nature00769; pmid: 12068309
51. L. R. Kenner *et al*., eIF2B-catalyzed nucleotide exchange and phosphoregulation by the integrated stress response. *Science* **364**, 491–495 (2019). doi: 10.1126/science.aaw2922; pmid: 31048491
52. S. Jentsch, I. Psakhye, Control of nuclear activities by substrate-selective and protein-group SUMOylation. *Annu. Rev. Genet.* **47**, 167–186 (2013). doi: 10.1146/annurev-genet-111212-133453; pmid: 24016193
53. I. Psakhye, S. Jentsch, Protein group modification and synergy in the SUMO pathway as exemplified in DNA repair. *Cell* **151**, 807–820 (2012). doi: 10.1016/j.cell.2012.10.021; pmid: 23122649
54. D. Menolfi, A. Delamarre, A. Lengronne, P. Pasero, D. Branzei, Essential Roles of the Smc5/6 Complex in Replication through Natural Pausing Sites and Endogenous DNA Damage Tolerance. *Mol. Cell* **60**, 835–846 (2015). doi: 10.1016/j.molcel.2015.10.023; pmid: 26698660
55. S. Agashe *et al*., Smc5/6 functions with Sgs1-Top3-Rmi1 to complete chromosome replication at natural pause sites. *Nat. Commun.* **12**, 2111 (2021). doi: 10.1038/s41467-021-22217-w; pmid: 33833229
56. G. De Piccoli, J. Torres-Rosell, L. Aragón, The unnamed complex: What do we know about Smc5-Smc6? *Chromosome Res.* **17**, 251–263 (2009). doi: 10.1007/s10577-008-9016-8; pmid: 19308705
57. I. Psakhye, F. Castellucci, D. Branzei, SUMO-Chain-Regulated Proteasomal Degradation Timing Exemplified in DNA Replication Initiation. *Mol. Cell* **76**, 632–645.e6 (2019). doi: 10.1016/j.molcel.2019.08.003; pmid: 31519521
58. A. Waizenegger *et al*., Mus81-Mms4 endonuclease is an Esc2-STUbL-Cullin8 mitotic substrate impacting on genome integrity. *Nat. Commun.* **11**, 5746 (2020). doi: 10.1038/s41467-020-19503-4; pmid: 33184279
59. I. Psakhye, D. Branzei, SMC complexes are guarded by the SUMO protease Ulp2 against SUMO-chain-mediated turnover. *Cell Rep.* **36**, 109485 (2021). doi: 10.1016/j.celrep.2021.109485; pmid: 34348159
60. J. J. L. Miranda, P. De Wulf, P. K. Sorger, S. C. Harrison, The yeast DASH complex forms closed rings on microtubules. *Nat. Struct. Mol. Biol.* **12**, 138–143 (2005). doi: 10.1038/nsmb896; pmid: 15640796
61. S. Westermann *et al*., Formation of a dynamic kinetochore-microtubule interface through assembly of the Dam1 ring complex. *Mol. Cell* **17**, 277–290 (2005). doi: 10.1016/j.molcel.2004.12.019; pmid: 15664196
62. C. L. Asbury, D. R. Gestaut, A. F. Powers, A. D. Franck, T. N. Davis, The Dam1 kinetochore complex harnesses microtubule dynamics to produce force and movement. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9873–9878 (2006). doi: 10.1073/pnas.0602249103; pmid: 16777964
63. V. H. Ramey *et al*., Subunit organization in the Dam1 kinetochore complex and its ring around microtubules. *Mol. Biol. Cell* **22**, 4335–4342 (2011). doi: 10.1091/mbc.e11-07-0659; pmid: 21965284
64. J. O. Kim *et al*., The Ndc80 complex bridges two Dam1 complex rings. *eLife* **6**, e21069 (2017). doi: 10.7554/eLife.21069; pmid: 28191870
65. C. T. Ng *et al*., Electron cryotomography analysis of Dam1C/DASH at the kinetochore-spindle interface in situ. *J. Cell Biol.* **218**, 455–473 (2019). doi: 10.1083/jcb.201809088; pmid: 30504246
66. S. Jenni, S. C. Harrison, Structure of the DASH/Dam1 complex shows its role at the yeast kinetochore-microtubule interface. *Science* **360**, 552–558 (2018). doi: 10.1126/science.aar6436; pmid: 29724956
67. F. Wendler *et al*., A genome-wide RNA interference screen identifies two novel components of the metazoan secretory pathway. *EMBO J.* **29**, 304–314 (2010). doi: 10.1038/emboj.2009.350; pmid: 19942856
68. M. Heidtman, C. Z. Chen, R. N. Collins, C. Barlowe, Yos1p is a novel subunit of the Yip1p-Yif1p complex and is required for transport between the endoplasmic reticulum and the Golgi complex. *Mol. Biol. Cell* **16**, 1673–1683 (2005). doi: 10.1091/mbc.e04-10-0873; pmid: 15659647
69. Y. Desfougères, R. U. Gerasimaitė, H. J. Jessen, A. Mayer, Vtc5, a Novel Subunit of the Vacuolar Transporter Chaperone Complex, Regulates Polyphosphate Synthesis and Phosphate Homeostasis in Yeast. *J. Biol. Chem.* **291**, 22262–22275 (2016). doi: 10.1074/jbc.M116.746784; pmid: 27587415
70. M. Hothorn *et al*., Catalytic core of a membrane-associated eukaryotic polyphosphate polymerase. *Science* **324**, 513–516 (2009). doi: 10.1126/science.1168120; pmid: 19390046
71. M. Vietri, M. Radulovic, H. Stenmark, The many functions of ESCRTs. *Nat. Rev. Mol. Cell Biol.* **21**, 25–42 (2020). doi: 10.1038/s41580-019-0177-4; pmid: 31705132
72. J. H. Hurley, ESCRTs are everywhere. *EMBO J.* **34**, 2398–2407 (2015). doi: 10.15252/embj.201592484; pmid: 26311197
73. S. Tang *et al*., Structural basis for activation, assembly and membrane binding of ESCRT-III Snf7 filaments. *eLife* **4**, e12548 (2015). doi: 10.7554/eLife.12548; pmid: 26670543
74. C. Schluter *et al*., Global analysis of yeast endosomal transport identifies the vps55/68 sorting complex. *Mol. Biol. Cell* **19**, 1282–1294 (2008). doi: 10.1091/mbc.e07-07-0659; pmid: 18216282
75. S. Siniossoglou, H. R. Pelham, An effector of Ypt6p binds the SNARE Tlg1p and mediates selective fusion of vesicles with late Golgi membranes. *EMBO J.* **20**, 5991–5998 (2001). doi: 10.1093/emboj/20.21.5991; pmid: 11689439
76. K. Mei *et al*., Cryo-EM structure of the exocyst complex. *Nat. Struct. Mol. Biol.* **25**, 139–146 (2018). doi: 10.1038/s41594-017-0016-2; pmid: 29335562
77. H.-T. Chou, D. Dukovski, M. G. Chambers, K. M. Reinisch, T. Walz, CATCHR, HOPS and CORVET tethering complexes share a similar architecture. *Nat. Struct. Mol. Biol.* **23**, 761–763 (2016). doi: 10.1038/nsmb.3264; pmid: 27428774
78. R. Behnia, F. A. Barr, J. J. Flanagan, C. Barlowe, S. Munro, The yeast orthologue of GRASP65 forms a complex with a coiled-coil protein that contributes to ER to Golgi traffic. *J. Cell Biol.* **176**, 255–261 (2007). doi: 10.1083/jcb.200607151; pmid: 17261844
79. M. Schuldiner *et al*., Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507–519 (2005). doi: 10.1016/j.cell.2005.08.031; pmid: 16269340
80. W. Ma, J. Goldberg, TANGO1/cTAGE5 receptor as a polyvalent template for assembly of large COPII coats. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10061–10066 (2016). doi: 10.1073/pnas.1605916113; pmid: 27551091
81. V. G. Stancheva *et al*., Combinatorial multivalent interactions drive cooperative assembly of the COPII coat. *J. Cell Biol.* **219**, e202007135 (2020). doi: 10.1083/jcb.202007135; pmid: 32997735
82. T. C. Südhof, J. E. Rothman, Membrane fusion: Grappling with SNARE and SM proteins. *Science* **323**, 474–477 (2009). doi: 10.1126/science.1161748; pmid: 19164740
83. S. Conchon, X. Cao, C. Barlowe, H. R. Pelham, Got1p and Sft2p: Membrane proteins involved in traffic to the Golgi complex. *EMBO J.* **18**, 3934–3946 (1999). doi: 10.1093/emboj/18.14.3934; pmid: 10406798
84. R. B. Sutton, D. Fasshauer, R. Jahn, A. T. Brunger, Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. *Nature* **395**, 347–353 (1998). doi: 10.1038/26412; pmid: 9759724
85. R. Jahn, R. H. Scheller, SNAREs—Engines for membrane fusion. *Nat. Rev. Mol. Cell Biol.* **7**, 631–643 (2006). doi: 10.1038/nrm2002; pmid: 16912714
86. J. Rizo, Mechanism of neurotransmitter release coming into focus. *Protein Sci.* **27**, 1364–1391 (2018). doi: 10.1002/pro.3445; pmid: 29893445
87. L. Burri *et al*., A SNARE required for retrograde transport to the endoplasmic reticulum. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9873–9877 (2003). doi: 10.1073/pnas.1734000100; pmid: 12893879
88. Y. Hong *et al*., Human PIG-U and yeast Cdc91p are the fifth subunit of GPI transamidase that attaches GPI-anchors to proteins. *Mol. Biol. Cell* **14**, 1780–1789 (2003). doi: 10.1091/mbc.e02-12-0794; pmid: 12802054
89. D. G. Gamage, T. L. Hendrickson, GPI transamidase and GPI anchored proteins: Oncogenes and biomarkers for cancer. *Crit. Rev. Biochem. Mol. Biol.* **48**, 446–464 (2013). doi: 10.3109/10409238.2013.831024; pmid: 23978072
90. L. Yi *et al*., Disulfide Bond Formation and N-Glycosylation Modulate Protein-Protein Interactions in GPI-Transamidase (GPIT). *Sci. Rep.* **7**, 45912 (2017). doi: 10.1038/srep45912; pmid: 28374821
91. P. Moran, I. W. Caras, A nonfunctional sequence converted to a signal for glycophosphatidylinositol membrane anchor attachment. *J. Cell Biol.* **115**, 329–336 (1991). doi: 10.1083/jcb.115.2.329; pmid: 1717483
92. P. Fraering *et al*., The GPI transamidase complex of Saccharomyces cerevisiae contains Gaa1p, Gpi8p, and Gpi16p. *Mol. Biol. Cell* **12**, 3295–3306 (2001). doi: 10.1091/mbc.12.10.3295; pmid: 11598210
93. K. Ohishi, N. Inoue, T. Kinoshita, PIG-S and PIG-T, essential for GPI anchor attachment to proteins, form a complex with GAA1 and GPI8. *EMBO J.* **20**, 4088–4098 (2001). doi: 10.1093/emboj/20.15.4088; pmid: 11483512
94. K. Ohishi, K. Nagamune, Y. Maeda, T. Kinoshita, Two subunits of glycosylphosphatidylinositol transamidase, GPI8 and PIG-T, form a functionally important intermolecular disulfide bridge. *J. Biol. Chem.* **278**, 13959–13967 (2003). doi: 10.1074/jbc.M300586200; pmid: 12582175
95. S. Vainauskas, A. K. Menon, A conserved proline in the last transmembrane segment of Gaa1 is required for glycosylphosphatidylinositol (GPI) recognition by GPI transamidase. *J. Biol. Chem.* **279**, 6540–6545 (2004). doi: 10.1074/jbc.M312191200; pmid: 14660601
96. U. Meyer, M. Benghezal, I. Imhof, A. Conzelmann, Active site determination of Gpi8p, a caspase-related enzyme required for glycosylphosphatidylinositol anchor addition to proteins. *Biochemistry* **39**, 3461–3471 (2000). doi: 10.1021/bi992186o; pmid: 10727241
97. D. G. Gamage *et al*., The soluble domains of Gpi8 and Gaa1, two subunits of glycosylphosphatidylinositol transamidase (GPI-T), assemble into a complex. *Arch. Biochem. Biophys.* **633**, 58–67 (2017). doi: 10.1016/j.abb.2017.09.006; pmid: 28893510
98. J. L. Meitzler, J. J. Gray, T. L. Hendrickson, Truncation of the caspase-related subunit (Gpi8p) of Saccharomyces cerevisiae GPI transamidase: Dimerization revealed. *Arch. Biochem. Biophys.* **462**, 83–93 (2007). doi: 10.1016/j.abb.2007.03.035; pmid: 17475206
99. T. T. M. Nguyen *et al*., Bi-allelic Variants in the GPI Transamidase Subunit PIGK Cause a Neurodevelopmental Syndrome with Hypotonia, Cerebellar Atrophy, and Epilepsy. *Am. J. Hum. Genet.* **106**, 484–495 (2020). doi: 10.1016/j.ajhg.2020.03.001; pmid: 32220290
100. R. Evans *et al*., Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 [Preprint]. 4 October 2021. doi: 10.1101/2021.10.04.463034
101. M. Baek, L. Heo, R. Ndem, neilfleckSCRI, RosettaCommons/RoseTTAFold: RoseTTAFold update: Including the simpler version for PPI screening, Zenodo (2021). doi: 10.5281/zenodo.5639837

ACKNOWLEDGMENTS

We thank E. Horvitz, N. V. Grishin, H. Park, and J. H. Thomas for helpful discussions; L. Goldschmidt and A. Guillory for computing resource management; and L. Stewart for logistical support. Additionally, we are grateful to M. Bard, T. N. Davis, D. G. Drubin,

M. J. Dunham, S. D. Emr, F. Hughson, J. Hurley, K. Murakami, N. Nakamura, E. Nogales, R. Schekman, S. Shan, S. Showman, K. Sugasawa, and S. Suzuki for their correspondence and biological expertise. We thank S. Burley, B. Vallat, and J. Westbrook at the RCSB Protein Data Bank and T. Schwede, G. Tauriello, A. Waterhouse, and S. Bienert at SWISS-MODEL for hosting our model structures at the ModelArchive. **Funding:** This work was supported by Microsoft (M.B., D.B., and Azure compute time and expertise), Amgen (D.B. and I.R.H.), Southwestern Medical Foundation (J.P. and Q.C.), the Washington Research Foundation (M.B. and Q.C.), the Howard Hughes Medical Institute (D.B., S.B., S.K., and generous compute time on Janelia), National Science Foundation (NSF) Cyberinfrastructure for Biological Research (CIBR, Award DBI 1937533 to D.B. and I.A.), CPRIT training grant (RP210041 to J.Z.), UK Medical Research Council (MRC_UP_1201/10 to E.A.M.), HHMI Gilliam Fellowship (D.J.B.), the Deutsche Forschungsgemeinschaft (KI-562/11-1 and KI-562/7-1 to C.K.), NIH/NIGMS (R21AI156595 to S.O., R35GM136258 to J.C.F., R35NS097333 to J.R., R35GM118026 and R01CA221858 to E.C.G.), HHMI fellowship of the Damon Runyon Cancer Research Foundation (DRG2273-16 to S.B. and DRG2389-20 to K.L.), AIRC investigator and the European Research Council Consolidator (IG23710 and 682190 to D.Br.), and the Defense Threat Reduction Agency (HDTRA1-21-1-0007 to D.B.). We also thank The National Energy Research Scientific Computing Center (NERSC) for providing computing time (project m3962 at NERSC). **Author contributions:** Q.C. and D.B. conceived the research; J.P. and Q.C. prepared the sequence alignments used in the screen; M.B. implemented the RoseTTAFold pipeline; M.B. and S.O. repurposed AlphaFold for complex modeling; J.P., J.Z., and Q.C. designed the PPI screening procedure; I.R.H., M.B., I.A., and Q.C. carried out the screen; I.R.H., A.K., and Q.C. analyzed and presented the results; I.R.H., A.K., Q.C., and D.B. coordinated the collaborative efforts; T.J.N., S.B., S.R.B., V.G.S., X.H.L., K.L., Z.Z., D.J.B., U.R., J.K., I.S.F., B.S., D.B., J.R., C.K., E.C.G., S.B., S.K., E.A.M., J.C.F., and T.L.H. provided biological insights on specific examples; Q.C. and D.B. drafted the manuscript and all other authors contributed to the description of specific examples; all authors discussed the results and commented on the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Structures of highly confident pairs with accompanying pMSAs and metadata are available at the ModelArchive: https://modelarchive.org/doi/10.5452/ma-bak-cepc. RoseTTAFold two-track version is available at https://github.com/RosettaCommons/RoseTTAFold or Zenodo (*101*). AlphaFold was fetched from https://github.com/deepmind/alphafold on 16 July 2021 (v2.0.0). Code for a GPU implementation of DCA and the modifications to the AlphaFold predictions script are provided in supplementary materials and methods.

SUPPLEMENTARY MATERIALS

View/request a protocol for this paper from *Bio-protocol*.

20 September 2021; accepted 2 November 2021
Published online 11 November 2021
10.1126/science.abm4805