Science

RESEARCH ARTICLES

Cite as: J. Zhang *et al*., *Science* 10.1126/science.adt1630 (2025).

Predicting protein-protein interactions in the human proteome

Jing Zhang[1,2,3]†, Ian R. Humphreys[4,5]†, Jimin Pei[1,2,3]†, Jinuk Kim[6,7], Chulwon Choi[6], Rongqing Yuan[1,2,3], Jesse Durham[1,2,3], Siqi Liu[3,8,9], Hee-Jung Choi[6], Minkyung Baek[6,10], David Baker[4,5,11], Qian Cong[1,2,3]*

[1]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [3]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA. [4]Department of Biochemistry, University of Washington, Seattle, WA, USA. [5]Institute for Protein Design, University of Washington, Seattle, WA, USA. [6]Department of Biological Sciences, Seoul National University, Seoul, Republic of Korea. [7]Division of Biological Science and Technology, Yonsei University, Wonju, Republic of Korea. [8]Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [9]Children's Medical Center Research Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. [10]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea. [11]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

***Corresponding author. Email: qian.cong@utsouthwestern.edu**

†These authors contributed equally to this work.

Protein-protein interactions (PPI) are essential for biological function. Coevolutionary analysis and deep learning (DL) based protein structure prediction have enabled comprehensive PPI identification in bacteria and yeast, but these approaches have had limited success for the more complex human proteome. We overcame this challenge by enhancing the coevolutionary signals with 7-fold deeper multiple sequence alignments harvested from 30 petabytes of unassembled genomic data and developing a new DL network trained on augmented datasets of domain-domain interactions from 200 million predicted protein structures. We systematically screened 200 million human protein pairs and predicted 17,849 interactions with an expected precision of 90%, of which 3,631 interactions were not identified in previous experimental screens. Three-dimensional models of these predicted interactions provide numerous hypotheses about protein function and mechanisms of human diseases.

Detecting the interacting partners of proteins and determining the three-dimensional (3D) structures of protein complexes are essential to understanding protein function (*1*). Large-scale experimental methods such as yeast two-hybrid (Y2H) and affinity-purification mass spectrometry (APMS) have been used to identify protein-protein interactions (PPI) across the human proteome (*2*). While powerful, these methods detect PPIs under non-physiological conditions and suffer from considerable false-positive and false-negative rates (*2*). This is illustrated by the fact that although the human interactome is estimated to contain 74,000 to 200,000 PPIs (*3*), the experimental data from PPI databases, UniProt (*4*), BioGrid (*5*), and STRING (*6*), suggest that over 1 million (M) human PPIs may exist. Additionally, high-confidence PPIs from these databases show low consistency (Fig. 1A), with only 3988 PPIs regarded as confident by all three.

Computational approaches have been developed to complement experimental studies in resolving the human interactome. They predict PPIs based on homology to known interacting partners, quality of the predicted interfaces, and functional associations between proteins (*7*–*10*). Coevolutionary analysis (residues at PPI interfaces coevolve to preserve their interactions) has been combined with 3D structure prediction methods, such as AlphaFold (AF) (*11*) and RoseTTAFold (RF) (*12*), to identify PPIs on a proteome-wide scale in bacteria and yeast (*13*, *14*). The probability for any protein pair to interact is estimated by 1) constructing multiple sequence alignments (MSAs) of orthologs of the two proteins across many species, 2) concatenating sequences of the same species to build paired MSAs (pMSAs), and 3) inferring the probabilities of residue-residue interactions between the two proteins based on coevolution and complex structure modeling (*15*).

Although similar approaches have been used to detect interactions between select human proteins (*16*, *17*), conducting a de novo proteome-wide PPI screen in humans remains challenging due to the daunting computational scale and limited number of animal genomes, which reduces the statistical power for detecting coevolution between human proteins. Our previously developed lightweight deep learning (DL) networks for rapid PPI screening, which performed well in bacteria (*18*) and yeast (*14*), are considerably less accurate in humans. Although AF2 achieves higher accuracy than our lightweight PPI networks, it requires excessive computational resources, making it impractical for proteome-wide application.

We set out to overcome these challenges and systematically identify human PPIs. To enhance the statistical power of coevolutionary analysis, we harnessed petabytes of untapped genomic sequence data in draft eukaryotic genomes and genomic reads. To efficiently analyze hundreds of millions of human protein pairs, we developed a fast and

accurate method for PPI prediction by 1) optimizing the architecture of state-of-the-art DL networks and 2) augmenting the training set of experimentally determined PPIs in the PDB with a much larger distillation set of domain-domain interactions (DDIs) harvested from accurately predicted monomer structures in the AlphaFold protein structure Database (AFDB) (*19*) based on the assumption that DDI interfaces resemble PPI interfaces in coevolutionary and physicochemical properties.

Harnessing untapped genomic sequence data

Biologists typically rely on databases with protein sequences annotated from genomes or from metagenomic assemblies. However, most available eukaryotic genomes, particularly animal genomes, lack annotated protein sequences. Identifying protein-coding regions in eukaryotic genomes is non-trivial due to the complicated and frequent lineage-specific rules of translation initiation and mRNA splicing (*20*). Hence, it is not a routine practice to annotate and deposit the protein-coding sequences for draft eukaryotic genomes: among the 36,840 eukaryotic genomes available at NCBI as of June, 2024, only 7,355 (20%) were associated with annotated proteins. Moreover, assembling diploid, heterozygous, and repeat-rich eukaryotic genomes is challenging without long-read sequencing (*21*). The NCBI Sequence Read Archive (SRA) database has over 30 petabytes of shotgun sequencing reads, most of which remain unassembled and unanalyzed for protein-coding content.

To enrich the coevolutionary signals in our MSAs, we mined the NCBI Genome and SRA databases to extract whole genome and whole transcriptome datasets for 21,414 diverse Eukaryotes, focusing on Chordates and Arthropods, two large phyla of higher Eukaryotes. We selected one representative genomic dataset per species, including 1) 2,424 species whose proteome sequences were previously annotated from genomes, 2) 6,863 species with draft genomes but no protein annotations, and 3) 12,128 species with whole genome or whole transcriptome shotgun reads. We developed a bioinformatic pipeline (supplementary methods M2) to assemble protein-coding sequences from each type of dataset, utilizing splicing-site-aware sequence aligners (*22*, *23*) and reference protein sets from model organisms (Fig. 1B). All predicted protein sequences were aligned to their human orthologs, if available, and we used reciprocal best-hit criteria (*24*) to distinguish orthologs from paralogs. We name the resulting alignments omicMSAs, because they are directly derived from genomic data. Our omicMSAs include sequences from 21,414 species spanning 9,905 genera, 2,727 families, and 626 orders, drastically expanding the taxonomic diversity of available protein sequences in UniRef100, which only contains 3,082 species (Fig. 1C). The pMSAs built from omicMSAs are thus manyfold deeper than those built using eukaryotic sequences in UniRef100.

We reasoned that the rich evolutionary information in omicMSAs should improve the ability of DL networks to distinguish true PPIs from false ones. We tested this hypothesis using benchmarks derived from PPI databases. There are 75,739, 77,147, and 68,761 confident human PPIs from UniRef, BioGrid (multi-validated), and STRING (physical interactions with combined score > 700), respectively. However, only 3,988 (5.3–5.8%) PPIs are shared by all (Fig. 1A), from which we selected 3,000 positive control pairs (data S1). The negative control set contains 30,000 random pairs of human proteins showing no evidence for their interactions (data S1). We used precision-recall curves to evaluate the performance of different MSAs and DL networks. Because the signal-to-noise ratio for proteome-wide PPI screening is approximately 1:1000 (74k–200k true PPIs out of 200M pairs) (*25*), we down-weighted each positive control pair by a factor of 100 when computing precision to mirror this ratio.

We compared omicMSAs against MSAs built by other commonly used strategies. The first is HHblits (*26*) against UniRef, a widely used approach to prepare MSA inputs for RF and AF. HHblits by default filters the output MSAs at 90% sequence identity, and we named this strategy UniRef90. Filtering MSAs before building pMSAs drastically reduces the number of taxa that can be paired; to increase pMSA depth, we disabled this filter and termed the alternative strategy UniRef100. Metagenomic sequences are frequently used to obtain deeper MSAs and improve structure prediction (*27*). We used the ColabFold MSA pipeline (*28*) to combine UniRef and metagenomic sequences.

We found that omicMSA led to the best performance for all the tested methods (fig. S8), including RF-2t used in our previous yeast PPI screen (*14*), the newly developed RF2-PPI in this study, and AF2, in distinguishing true PPIs from random pairs (red curves in Fig. 1, E to G). The ColabFold MSA (grey curves in Fig. 1, E to G) improves over UniRef100 when used with AF2, but not other tools. Metagenomic data, mostly sequenced from environmental samples, do not contain sequences of multicellular Eukaryotes and are not annotated with the assumption of mRNA splicing (*29*); thus, they do not necessarily improve human PPI prediction. In contrast, we focused on assembling sequences from genomic data of animals with pipelines aware of mRNA splicing. All the tools we tested were trained using MSAs built from UniRef and metagenomic sequences, and they are not adapted to work with omicMSA. Thus, the superior performance of omicMSA demonstrates the power of stronger evolutionary signals gleaned from a much wider range of taxa. The available eukaryotic genomic data are constantly growing (~20% more species per year), providing even greater potential in the future.

RoseTTAFold2-PPI: A DL network for rapid PPI identification

Although intra- and intermolecular contacts are governed by the same physicochemical rules, intermolecular contacts are more frequently transient and contain a higher fraction of electrostatic interactions (*30*). AlphaFold-multimer (AFmm) (*31*) and AlphaFold3 (AF3) (*32*) were trained with PPIs in the PDB to improve their ability to model the 3D structures of protein complexes. However, such a training regime and inference goal may not improve a DL network's ability to distinguish true PPIs from non-interacting proteins. Without exposure to non-interacting protein pairs, the network may fail to learn that most protein pairs do not interact, resulting in a bias toward producing interacting structure models indiscriminately. Trained on PPIs from the PDB and optimized for speed, we developed RF2-lite (*13*) to distinguish true PPIs from random pairs. RF2-lite is 10 times faster than AF2, but shows considerably lower accuracy than AF2. We reason that one factor limiting the accuracy of current methods is the amount of training data. At 30% sequence identity, there are 24,358 clusters of heteromeric PPIs in the PDB (December, 2023), among which many pairs are human proteins or their close homologs (MMseqs e-value < 0.00001, identity > 30%); after removing these complexes to avoid information leakage for human PPI prediction, only 13,231 clusters remain (Fig. 2A, left).

AlphaFold2 enhanced its outstanding performance through self-distillation, a process where it was trained on high-confidence predictions from an earlier version of the model (*11*). This suggests that high-quality predictions of protein complexes should similarly improve the performance of PPI prediction networks. However, despite over 200M predicted structures of protein monomers in AFDB (*33*), predicted complex structures remain scarce. We hypothesized that extracted DDIs from the 200M AF2 models in the AFDB (*33*) could substantially enlarge the training dataset for PPI prediction networks and improve performance. Domains are structural and evolutionary units that are recombined to create proteins throughout evolution (*34*), and the interfaces between domains of the same protein closely resemble those between distinct protein partners (*35*).

We previously developed a method to segment AF2 models into domains based on structural features (inter-residue distances and predicted aligned errors (PAEs)) and homology to previously classified domains from PDB entries (*36*). We repurposed this method to integrate structural features and domain annotations of UniProt entries from InterPro (*37*) (Fig. 2B and supplementary methods M3.2). We found 12.4M multi-domain proteins among the 53.7M AFDB models filtered at 50% sequence identity (*38*). These models contain 22.6M high-quality (mean PAE within domains < 8) domain pairs from the same proteins. We focused on pairs with at least 25 inter-residue contacts (distance < 6Å) and high confidence in their interaction (mean inter-domain PAE < 8), resulting in 1M DDIs. Clustering these DDIs at 30% sequence identity and removing close homologs of human proteins resulted in 223,441 clusters (Fig. 2A, right), 17-fold larger than the PDB PPI training set.

We used both the PDB PPI and AFDB DDI sets to develop RF2-PPI based on the RF2 architecture (*39*), which integrates the MSA, inter-residue, and 3D structure features through attention-based blocks (Fig. 2C). Additionally, we included monomeric 3D structures to ensure that RF2-PPI learns principles of protein folding; we added negative controls (random protein or domain pairs from the same organism with no evidence of being functionally related) to ensure that RF2-PPI learns to distinguish interacting partners from random pairs. The monomers, positive DDIs, negative DDIs, positive PPIs, and negative PPIs were combined at a 2:1:1:1:1 ratio to generate the training dataset.

We explored a number of strategies to optimize the performance of RF2-PPI (supplementary methods M4.2); we found that adding the DDI training set and removing the 3D structure features remarkably improved its performance (figs. S17 and S18). The latter may be because weak human PPIs are frequently mediated by simple motifs (single helices or intrinsically disordered regions (IDRs)), but training data are dominated by interactions with larger interfaces. Direct integration of 3D features may drive the network to emphasize on the geometric and chemical complementarity at the PPI interfaces instead of the coevolutionary signals essential for predicting weak PPIs. The 3D structure module of AF2 is not directly integrated into the Evoformer, its core module for decoding coevolutionary signals, possibly allowing AF2 to better balance coevolutionary signals and physicochemical properties.

We compared the performance of RF2-PPI with the ColabFold pipeline, AF3 web server widely used by the scientific community, and other methods. Using area under the precision and recall curve (AUCPR) as a metric, RF2-PPI (orange in Fig. 2D) showed a 1.7-fold improvement over our previous method, RF2-lite (gold in Fig. 2D). When combined with omicMSAs, RF2-PPI (red in Fig. 2D) slightly outperforms the ColabFold pipeline (grey in Fig. 2D) in its ability to identify true PPIs. More importantly, RF2-PPI combined with omicMSA shows substantially improved recall at high precision (80% or 90%), making it well-suited for generating high-quality PPI predictions. AFmm (cyan in Fig. 2D) and AF3 (green in Fig. 2D), despite superior performance in modeling 3D structures of protein complexes, show substantially worse performance than RF2-PPI in distinguishing true PPIs from random pairs. The inferior performance of AFmm and AF3 is also related to the low signal-to-noise ratio (1:1000) associated with de novo PPI screens; at a signal-to-noise ratio of

1:10, AFmm and AF3 will outperform AF2 and other methods (fig. S31). The best performance for a de novo PPI screen is achieved when AF2 is used in combination with omicMSAs (blue in Fig. 2D).

We further evaluated the performance of RF2-PPI and AF2 in distinguishing PPIs derived from human protein complexes with experimentally determined structures (data S2). The availability of 3D structural data enabled us to stratify these interactions by interface size: large interfaces typically correspond to strong, permanent interactions, whereas small interfaces are more often associated with weak, transient PPIs (*40*). We observed that the performance of both RF2-PPI and AF2 correlates with interface size; strong interactions with large interfaces are more reliably predicted (Fig. 2E). However, weak and transient interactions, which are challenging to detect experimentally (*40*), also pose a challenge for these computational methods. Notably, although AF2 outperforms RF2-PPI in predicting strong interactions, RF2-PPI shows an advantage in distinguishing weak interactions from random, non-interacting pairs, likely due to its emphasis on coevolution instead of PPI interface quality.

To improve computational efficiency, we reduced the number of Evoformer-like blocks and streamlined the data processing pipeline in RF2-PPI. As a result, RF2-PPI is 50% faster than our previous model, RF2-lite, and approximately 20 times faster than AF2 (Fig. 2F). We employed both RF2-PPI and AF2 to identify and model human PPIs. RF2-PPI enabled large-scale, proteome-wide screening, while AF2 was used to generate high-quality 3D structural models for PPI candidates detected by the former.

Proteome-wide identification of human PPIs

Equipped with omicMSA and RF2-PPI, we set out to comprehensively predict the human interactome. Our analysis focused exclusively on identifying heteromeric PPIs, as homomeric interactions among human proteins have already been systematically predicted in previous studies (*41*). Moreover, heteromeric interactions are more diverse and functionally informative, making them a higher priority for expanding our understanding of cellular interaction networks and their underlying biology.

Using the 20,504 AF2 modeled monomeric human proteins from AFDB, we identified domains based on structural compactness (*36*) and evolutionary conservation (*37*). We included these domains and relatively conserved residues in the inter-domain linkers (conservation > 25% quantile of residues in domains) in our screen. Excluding regions that are poorly conserved and lack a rigid 3D structure improves the performance of our PPI screen pipeline (fig. S8). To fit within limited GPU memory, we split large proteins into segments with few inter-residue contacts and flexible relative orientation (supplementary methods M5.1). In total, we included 19,528 proteins in our screen, excluding a small fraction (4.8%) of proteins due to size, compactness, or conservation. The proteins included in our screen comprise a total of 9.5M residues, of which 6.1M (64%) are located within evolutionary domains. In addition, we included 1.0M relatively conserved residues from inter-domain linkers. Although these residues represent only 29% (1.0/(9.5–6.1)) of all inter-domain linker regions, our analysis of human protein complexes with structure templates in PDB suggests that they are expected to account for 56% of PPIs mediated by inter-domain linkers (fig. S23).

We began with an unbiased systematic search for PPIs across the 191M protein pairs formed by these 19,528 proteins. To make the search more tractable, we restricted our analysis to protein pairs localized within the same cellular compartment (CC), as well as poorly characterized proteins with unknown subcellular localization, based on UniProt keyword annotations. We analyzed 53.8M pairs sharing CC annotations and 57.4M pairs involving proteins without CC annotations sequentially through Direct Coupling Analysis (DCA) (*42*), RF2-PPI, and AF2 (Fig. 3A, middle). Although DCA, a statistical method to detect coevolution, shows considerably worse performance (Fig. 3B, left) than the state-of-the-art DL tools, it is 5-fold faster than RF2-PPI and we used it to select 43.6M (40%) pairs. Based on our benchmark, most (over 80%) pairs showing high interaction probability by RF2-PPI or AF2 will pass this DCA pre-filter (table S4). We applied RF2-PPI on the 43.6M pre-filtered protein pairs (supplementary methods M5.4) to predict their interaction probabilities ($RF_{IntProb}$). Of these, 1.9M pairs with $RF_{IntProb}$ above 0.3 (Fig. 3B, middle) advanced to the next stage, where AF2 was used to generate a 3D model and estimate another interaction probability ($AF_{IntProb}$) for each pair. Pairs with high $AF_{IntProb}$ (Fig. 3B, right) were selected, resulting in 6,763 predicted PPIs at an estimated precision of 80%, or a more stringent subset of 4,379 pairs at an estimated precision of 90% (supplementary methods M5.6).

We complemented the above de novo pipeline by carrying out a second screen incorporating the extensive information about human genetic associations and physical interactions derived primarily from high throughput experiments. Genetic associations between proteins do not imply direct interactions and many experimental screens that identify physical interactions are plagued by false positives. Evaluating such indirect evidence and noisy experimental data with our *in silico* pipeline allows us to identify a smaller set of PPIs with high confidence. Prior evidence of interactions allowed us to use lower $RF_{IntProb}$ and $AF_{IntProb}$ cutoffs while maintaining the same level of expected precision (supplementary methods M5.6). As shown in Fig. 3, C and D, predicting PPIs between random protein pairs with high confidence requires a stringent cutoff that is typically close to 0.99. In contrast, when

prior evidence for interaction is available, $RF_{IntProb}$ or $AF_{IntProb}$ thresholds can be substantially lower while maintaining the same level of confidence.

Among the 5.23M genetically associated pairs from STRING, we predicted 9,428 PPIs at 80% precision and 7,201 at 90% precision. Similarly, from the 1.15M protein pairs with prior evidence of physical interactions from UniProt, BioGRID, and STRING, we identified 21,960 PPIs at 80% precision and 14,129 at 90% precision. Together, the de novo (unbiased) and evidence-guided pipelines allowed us to predict 29,257 PPIs at an estimated precision of 80% (data S3), corresponding to a recovery of 12–32% of the human interactome. A more stringent subset includes 17,849 predicted PPIs at 90% estimated precision (data S4), with an estimated recall of 8–22% (supplementary methods M5.6). We focused on this set of high-confidence predictions for our subsequent in-depth analyses, but we share the full set of predictions with the scientific community at http://prodata.swmed.edu/humanPPI.

Integration of predicted PPIs with human protein knowledge bases

We compared our predictions against PPIs from other databases (Fig. 3E and fig. S25). We reasoned that the fraction of PPIs from another source that can be identified by our *in silico* screen ($f_{AI|DB}$) reflects the precision of PPIs in that database, which can be computed as $pre_{DB} = f_{AI|DB}/rec_{AI}$. Based on our estimates (Fig. 3F and table S8), PPIs with orthologous PDB templates are almost always true. Comprehensive PPI databases such as STRING (physical interactions only) and BioGRID contain a large number of PPIs at low precision (6–23%). However, the smaller high-confidence subsets provided by each database have higher precision (34%–100%). PPIs in three widely used resources, UniProt, HuRI (*43*), and Bioplex (*44*), show only moderate precision (13–57%). While genetic associations in STRING and homologous (but not orthologous) structural templates from the PDB may support the existence of PPIs, they do not provide direct evidence of physical interactions and exhibit low precision (1–3%).

Compared to PPIs documented in databases, our predicted PPIs—similar to those supported by orthologous PDB templates—show a much stronger tendency to share the same biological process (BP; Fig. 3G, left) and cellular component (CC; Fig. 3G, right) annotations in UniProt. Confident physical interactions curated by STRING also exhibit a high degree of biological process and cellular component concordance, although this may partly reflect STRING's use of functional criteria in selecting high-confidence PPIs. These observations support the high accuracy of our predictions and suggest that they can be used to infer the functions and subcellular localizations of poorly characterized proteins by linking them to well-annotated interaction partners.

We computed 3D structures of predicted PPIs using AF2 and AFmm and selected the best models (supplementary methods M6.2). We compared the structural properties of predicted PPIs against 9,725 human PPIs with orthologous templates in the PDB. Similar to PDB complexes, most (71%) predicted PPIs are primarily mediated by globular domains. PPIs mediated by inter-domain linkers (flexible helices or IDRs) are less frequent among our predictions compared to those observed in PDB complexes (Fig. 3H). This likely stems from our exclusion of poorly conserved regions in inter-domain linkers, which tend to increase false positives in our prediction pipeline.

It is important to note that PPIs mediated by IDRs typically involve small interfaces and are inherently more difficult to predict using RF2-PPI or AF2 (Fig. 2E). These interactions also often evade detection by experimental structure determination methods (*45*). Consequently, they are underrepresented both in our predicted PPIs and in experimentally resolved structures. In reality, IDR-mediated interactions are expected to constitute a substantial portion of the human interactome, and accurately identifying them and characterizing their structural features remains a major challenge for both computational and experimental approaches.

A higher fraction of predicted PPIs involve transmembrane helices (TMHs) at the interfaces than experimentally determined protein complexes. We wondered if the hydrophobic nature of TMHs makes them prone to false positives because DL networks might have learned to pack hydrophobic TMHs against each other even if they do not coevolve or interact in vivo. We tested the performance of our pipeline on pairs of transmembrane proteins (TMPs) and found that it predicts a larger fraction (magenta versus blue curves in Fig. 3I) of their interactions at a high precision, including those mediated by single-TMH proteins. These observations suggest that our pipeline is suitable for identifying interactions between TMPs, which are hard to detect experimentally (*46*).

Our 3D models of the high-confidence (90% precision) PPI predictions nearly triple the number of human PPIs with high-quality 3D structures. Thousands of post-translational modification (PTM) and functional (active and substrate/cofactor binding) sites map to the experimentally determined and predicted PPI interfaces (Fig. 3J, left and middle), representing a moderate fraction of all such annotated sites in human proteins (8.1% for PTMs and 12.7% for functional sites). Our models thus offer a rich resource for uncovering mechanistic insights into protein function. Furthermore, about 16k (15.3%) disease-associated single amino acid variations (SAVs) can be (supplementary methods M6.2) (*47–49*) mapped to the predicted or experimentally determined PPI interfaces (Fig. 3J, right). Among our high-confidence

predictions, 4,950 PPIs contain SAVs at their interfaces. These findings underscore disruption of PPIs as a common mechanism for genetic disorders and establish our predictions as a valuable resource for elucidating the molecular basis of human diseases.

Functional insights from predicted PPIs

Among the 17,849 predicted PPIs in our high-confidence set, 3,037 (17%) have orthologous PDB templates, another 11,181 (63%) are reported in PPI databases (UniProt, BioGRID, and STRING physical), and the remaining 3,631 pairs (20%) are not present in these established source of PPIs (Fig. 4A). Many of these uncharacterized PPIs are indirectly supported by other evidence such as genetic interactions in STRING and homologous (not orthologous) templates in the PDB. Our literature search showed some have been experimentally tested but are not yet propagated into these databases. Based on a recent catalog of poorly characterized proteins (*50*), 1,413 predicted PPIs involve proteins of unknown function wherein linking these proteins to well-characterized binding partners provides a shortcut to investigate their function (data S5). For our in-depth analysis, we prioritized 2,700 PPIs that are absent from PPI databases and lack homologous PDB templates (data S6). We classified predicted PPIs into functional categories according to UniProt keywords, and found categories enriched (P-value < 0.05) among these least known PPIs (Fig. 4B). Focusing on several of these categories, we highlight how our predictions can uncover unrecognized functional insights and help elucidate mechanisms of disease (Fig. 4, C to H, and supplementary results R4).

G protein-coupled receptors (GPCRs)

GPCRs are a large family of TMPs that detect extracellular signals and mediate cellular responses. They play vital roles in human physiology and are important drug targets (*51*). We detect interactions not present in current PPI databases between GPCRs and their putative downstream signaling molecules, ligands, or modulators. For example, we predict an interaction between G-protein coupled receptor 143 (GPR143) and BLOC-3 complex member HPS1 (Fig. 4C). GPR143 is an atypical GPCR localized in endolysosomes and melanosomes, while HPS1 is a component of a guanine nucleotide exchange factor (GEF) complex that activates Ras-related GTPase RAB32 and RAB38. The GPR143-HPS1 interaction is intriguing because both proteins are expressed in melanosomes, involved in melanosome biogenesis, and associated with albinism (*52*). Notably, four SAVs in GPR143 associated with Ocular Albinism Type 1 are located within the predicted GPR143-HPS1 interface (magenta box in Fig. 4C), supporting our prediction and suggesting that disruptions in this interaction may contribute to disease pathogenesis.

We also predict an interaction between C5a anaphylatoxin chemotactic receptor 2 (C5AR2) and neutrophil elastase (ELANE), a serine protease (Fig. 4D). The former is a receptor for C5a, a potent inflammatory mediator that modulates immune responses, particularly in myeloid cells such as neutrophils (*53*). The latter is also predominantly found in neutrophils and can inhibit C5a-dependent chemotaxis and neutrophil function (*54*). This predicted C5AR2-ELANE interaction suggests that ELANE is a negative modulator of C5AR2, supported by their association with neutrophils and opposite roles in chemotaxis. Specifically, ELANE appears to inhibit C5a-dependent neutrophil responses, which are mediated by C5a anaphylatoxin chemotactic receptor 1 (C5AR1) and C5AR2. Five SAVs in ELANE associated with Cyclical Neutropenia, a hematologic disorder, are located in the predicted ELANE-C5AR2 interface (magenta box in Fig. 4D), supporting our prediction.

Immune system proteins

The evolutionary arms race between hosts and pathogens drives hosts to deploy a wide array of mechanisms to defend against infectious organisms. 263 of the least known PPIs (Fig. 4B) involve immunity proteins, many of which mediate innate immunity signaling pathways. For instance, we predict an intriguing interaction between a deubiquitinase, OTU domain-containing protein 4 (OTUD4), and an E3 ubiquitin ligase, ankyrin repeat and SOCS box protein 8 (ASB8) (Fig. 4E). OTUD4 removes ubiquitin chains from mitochondrial antiviral-signaling protein (MAVS), an innate immune adapter that detects viral RNA, preventing its degradation (*55*); in contrast, ASB8 introduces polyubiquitins on the downstream protein kinase TBK1, promoting its degradation (*56*). We speculate that an interaction between these two proteins may help maintain a balance between activation and termination of innate immunity signaling.

Other uncharacterized PPIs we predict involve proteins with roles in the differentiation and activation of specialized immune cells. For instance, killer cell lectin-like receptor subfamily G member 1 (KLRG1) is a TM receptor that inhibits the activity of natural killer cells and effector T cells (*57*). The extracellular domain of KLRG1 is predicted to interact with toll-like receptor 3 (TLR3), a key player in recognizing double-stranded RNA and inducing immune responses (*58*). Mutations associated with a rare inherited immunodeficiency disorder causing susceptibility to Herpes simplex viruses 1 encephalitis, are located at the interface between KLRG1 and TLR3 (magenta box in Fig. 4F), supporting our prediction. This interaction suggests that part of KLRG1's inhibitory role might lie in deactivating TLR3's function as a pattern recognition receptor.

Mitochondrial targeting proteins

More than half of the previously unidentified interactions

(1,556 / 2,700) involve TMPs. These TMPs are frequently located in cellular organelles, such as mitochondria, ER, Golgi apparatus, or lysosomes (Fig. 4B). Many of the PPIs we predict in the mitochondria are associated with the assembly of mitochondrial complexes that form the respiratory electron transport chain. For example, we identified an interaction between cytochrome c oxidase subunit 2 (MT-CO2), a subunit of complex IV, and cytochrome c oxidase assembly protein COX20 (COX20), a known complex IV assembly factor (Fig. 4G). Our predictions also reveal several proteins of unknown function that may serve as assembly factors for key mitochondrial complexes. Notably, we predict an interaction between the uncharacterized protein C8orf82 and the electron transfer flavoprotein (ETF) complex, composed of ETFA and ETFB. This prediction suggests that C8orf82 may function as a chaperone or scaffold to facilitate ETF complex assembly, or a regulatory subunit modulating ETF activity under specific metabolic conditions. Disease-associated SAVs at the predicted PPI interfaces (magenta boxes in Fig. 4, G and H) support the functional relevance of these interactions and underscore their potential to provide insights into the molecular mechanisms of human disease.

Higher order oligomeric complexes

By integrating our predictions and PPIs derived from orthologous PDB templates, we identify 404 predicted multi-protein complexes, wherein each component interacts with at least two other components (supplementary methods M6.3 and data S7). We classify PPIs in these complexes into three categories: (1) with experimental structures; (2) confident entries of PPI databases; and (3) no or weak experimental support. We focus on putative complexes dominated by PPIs lacking confident experimental support (3rd category). Predicting their existence and modeling their 3D structures open new directions for future studies.

For example, tubulin polyglutamylase (TPG) is an enzyme complex responsible for adding polyglutamate chains to the C-terminal tail of tubulin, a modification important for the regulation of microtubule functions (*59*). In addition to the catalytic subunit TTLL1, the TPG complex is known to contain four other subunits: TPGS1, TPGS2, LRRC49 and NICN1 (*60*). We predict three additional subunits of TPG: TBC1D19, CSTPP1, and SANBR, each interacting with multiple known TPG components (Fig. 5A). TBC1D19 is the central component of the complex, interacting with five other subunits (TTLL1, LRRC49, NICN1, TPGS1, and SANBR). TBC1D19 and SANBR directly interact with the catalytic subunit TTLL1 via interfaces away from its ATP binding and catalytic sites, possibly affecting the activity of TTLL1 by allosteric regulation.

Another predicted complex consisting of NOL10, UTP25, DDX47, ESF1, ABT1, and DDX49, could link the regulation of basal transcription with ribosome biogenesis (Fig. 5B). ABT1 and its interacting partner ESF1 are suggested to regulate basal transcription (*61*), while UTP25 and DDX47 are implicated in ribosome biogenesis (*62*). These two processes must be coordinated to ensure that transcribed mRNAs can be efficiently translated into proteins. The central component, NOL10, is a poorly characterized nucleolar protein made of an N-terminal beta-propeller domain and C-terminal flexible helices. The connection between this complex and the ribosome is also supported by its predicted interactions with ribosomal proteins via ESF1, DDX47, and NOL10; the N-domain of NOL10 is even part of an experimental ribosome structure (PDB: 7MQ8). Future studies of this putative complex may reveal additional insights about the coordination between the fundamental processes of transcription and translation.

We predict several complexes composed of proteins involved in the biogenesis of cilia and flagella (Fig. 5C and fig. S40), two cellular organelles that share molecular machineries during their formation. The flagellum enables sperm to swim toward and fertilize the egg, linking these complexes to sperm motility and reproduction. The first predicted complex includes CATIP, RIIAD1, MDH1B, CFAP91, AKAP14, AK9, and MORN5, among which CFAP91, an intrinsically disordered protein, serves as a scaffold to assemble all but one (RIIAD1) other protein. Involvement of this complex in flagella biogenesis and sperm motility is supported by several facts: 1) CFAP91 and CATIP participate in cell skeleton organization during cilia/flagella biogenesis (*63*, *64*); 2) AK9 affects sperm mobility (*65*); and 3) defects in CATIP and AK9 are associated with asthenozoospermia (*64*, *65*).

Additional components of existing protein complexes

We predict 139 additional components of known human protein complexes (supplementary methods M6.3 and data S8) in the complex portal database (*66*). Each additional component is predicted to interact with at least two subunits of a known complex and primarily interact (>50% of predicted PPIs) with this complex. For example, LYAR, a nuclear protein involved in rRNA processing (*67*), is predicted to interact with two telomere maintenance complex (TMC) proteins, DKC1 and GAR1 (Fig. 5D). Further components of TMC include NOP10, NHP2, and WRAP53, and they participate in the processing and trafficking of the RNA template used to synthesize telomeric DNA (*68*). Multiple components of the TMC, including DKC1, GAR1, NOP10, and NHP2, have been shown to participate in rRNA processing and ribosome biosynthesis (*68*). The binding of LYAR to these components may enable the TMC to process rRNA instead of telomeric RNA templates.

We identify an additional subunit to the Glycosylphosphatidylinositol-N-acetylglucosaminyltransferase (GPI-GnT) complex, which catalyzes the first step of GPI anchor

biosynthesis (*69*). The GPI-GnT complex consists of seven subunits: PIGA, PIGC, PIGH, PIGP, PIGQ, PIGY, and DPM2 (*69*), and its 3D structure remains unresolved. ARV1, a TMP found to be related to GPI biosynthesis (*70*), is predicted to interact with PIGC and PIGQ. Our 3D model of the GPI-GnT complex with ARV1 (Fig. 5E) shows that a hydrophobic helix (residues 422-442) in the catalytic subunit, PIGA, is positioned in parallel to the membrane and docked onto a heterotrimer of the 2-pass TMPs: PIGP, PIGY, and DPM2. ARV1 (5 TMHs) is positioned between PIGC (8 TMHs) and PIGQ, and several hydrophobic helices of PIGQ lie in parallel to the membrane. These subunits form a ring-like structure with a large cavity in the middle of the membrane that could facilitate substrate binding or transport of the product.

Finally, we predict an additional component of the MKS transition zone complex (TZC) located between the basal body and axoneme of cilia and flagella (*71*). Predicted PPIs between TZC components enabled us to model a large complex (Fig. 5F) spanning the extracellular (TCN2 and TMEM231), membrane (TMEM17, TMEM107, TMEM138, and TMEM237), and cytoplasmic space (MKS1, B9D1, B9D2, and CC2D2A). TMEM138 has been suggested to localize to the transition zone based on immunofluorescence assays and is associated with the same ciliopathies as other transition zone proteins (*72*). We further predict TMEM138 as a subunit of TZC, sandwiched between TMEM17 and TMEM237 in our computed model. We included one copy of each TZC subunit in our model; however, it is likely that these proteins will homo-oligomerize to form a much larger complex that constitutes the transition zone.

Conclusions

Despite major advances in monomeric and oligomeric protein structure prediction using DL networks, distinguishing the 74k-200k true PPIs among the 200M total pairs of human proteins has remained a grand challenge. Here, we tackle this challenge by leveraging the largest sequence and structure datasets available and focusing on two key innovations: 1) strengthening the coevolutionary signals between interacting proteins using 7-fold deeper MSAs directly assembled from genomic data and 2) developing RF2-PPI, a new, efficient DL network trained on 17-fold larger datasets derived from predicted DDIs, to differentiate true PPIs from random protein pairs. These improvements enabled us to systematically identify human PPIs on a proteome-wide scale, predicting 18k interactions with an expected precision of 90%. These predictions include more than 3.6k previously unknown PPIs, particularly among transmembrane proteins that are challenging to characterize experimentally. These predictions should offer numerous valuable insights into human biology and diseases, as illustrated by our biological vignettes.

There remain many future challenges for understanding the human 3D interactome. We can only confidently predict a small fraction of human PPIs (12–32%), and capturing the weaker and transient interactions, frequently mediated by IDRs, remain challenging for computational methods. We expect the power of our approach will continue to expand as more sequence and structure data become available and DL techniques advance. By seamlessly integrating these rapidly advancing computational approaches with rigorous experimental studies, we are poised to accelerate the comprehensive characterization of the human interactome, thereby unlocking new insights into human biology and diseases.

Materials and methods are available in the supplementary materials.

REFERENCES AND NOTES

1. T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Wanamaker, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, M. Vidal, A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014). doi:10.1016/j.cell.2014.10.050 Medline
2. V. S. Rao, K. Srinivas, G. N. Sujini, G. N. S. Kumar, Protein-protein interaction detection: Methods and analysis. *Int. J. Proteomics* **2014**, 147648 (2014). doi:10.1155/2014/147648 Medline
3. K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A.-L. Barabási, M. Vidal, An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009). doi:10.1038/nmeth.1280 Medline
4. A. Bateman, M.-J. Martin, S. Orchard, M. Magrane, R. Agivetova, S. Ahmad, E. Alpi, E. H. Bowler-Barnett, R. Britto, B. Bursteinas, H. Bye-A-Jee, R. Coetzee, A. Cukura, A. Da Silva, P. Denny, T. Dogan, T. G. Ebenezer, J. Fan, L. G. Castro, P. Garmiri, G. Georghiou, L. Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, P. Jokinen, V. Joshi, D. Jyothi, A. Lock, R. Lopez, A. Luciani, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, M. Menchi, A. Mishra, K. Moulang, A. Nightingale, C. S. Oliveira, S. Pundir, G. Qi, S. Raj, D. Rice, M. R. Lopez, R. Saidi, J. Sampson, T. Sawford, E. Speretta, E. Turner, N. Tyagi, P. Vasudev, V. Volynkin, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.-C. Blatter, J. Bolleman, E. Boutet, L. Breuza, C. Casals-Casas, E. de Castro, K. C. Echioukh, E. Coudert, B. Cuche, M. Doche, D. Dornevil, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, G. Keller, A. Kerhornou, V. Lara, P. Le Mercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. B. Neto, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, M. Pozzato, M. Pruess, C. Rivoire, C. Sigrist, K. Sonesson, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L.-S. Yeh, J. Zhang, P. Ruch, D. Teodoro; UniProt Consortium, UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021). doi:10.1093/nar/gkaa1100 Medline
5. R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-

Aryamontri, K. Dolinski, M. Tyers, 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 Medline
6. D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, C. von Mering, 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). doi:10.1093/nar/gkac1000 Medline
7. D. Petrey, H. Zhao, S. J. Trudeau, D. Murray, B. Honig, PrePPI: A structure informed proteome-wide database of protein-protein interactions. *J. Mol. Biol.* **435**, 168052 (2023). doi:10.1016/j.jmb.2023.168052 Medline
8. H. Zhao, D. Petrey, D. Murray, B. Honig, ZEPPI: Proteome-scale sequence-based evaluation of protein-protein interaction models. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2400260121 (2024). doi:10.1073/pnas.2400260121 Medline
9. M. Gao, D. Nakajima An, J. M. Parks, J. Skolnick, AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022). doi:10.1038/s41467-022-29394-2 Medline
10. E. W. Bell, J. H. Schwartz, L. Freddolino, Y. Zhang, PEPPI: Whole-proteome protein-protein interaction prediction through structure and sequence similarity, functional association, and machine learning. *J. Mol. Biol.* **434**, 167530 (2022). doi:10.1016/j.jmb.2022.167530 Medline
11. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi:10.1038/s41586-021-03819-2 Medline
12. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021). doi:10.1126/science.abj8754 Medline
13. I. R. Humphreys, J. Zhang, M. Baek, Y. Wang, A. Krishnakumar, J. Pei, I. Anishchenko, C. A. Tower, B. A. Jackson, T. Warrier, D. T. Hung, S. B. Peterson, J. D. Mougous, Q. Cong, D. Baker, Essential and virulence-related protein interactions of pathogens revealed through deep learning. bioRxiv 2024.04.12.589144 [Preprint] (2024); https://doi.org/10.1101/2024.04.12.589144.
14. I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde, V. G. Stancheva, X.-H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernández, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. C. Fromme, T. L. Hendrickson, Q. Cong, D. Baker, Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021). doi:10.1126/science.abm4805 Medline
15. Q. Cong, I. Anishchenko, S. Ovchinnikov, D. Baker, Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019). doi:10.1126/science.aaw6718 Medline
16. D. F. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. S. Dunham, P. Albanese, A. Keller, R. A. Scheltema, J. E. Bruce, A. Leitner, P. Kundrotas, P. Beltrao, A. Elofsson, Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* **30**, 216–225 (2023). doi:10.1038/s41594-022-00910-8 Medline
17. J. Zhang, J. Pei, J. Durham, T. Bos, Q. Cong, Computed cancer interactome explains the effects of somatic mutations in cancers. *Protein Sci.* **31**, e4479 (2022). doi:10.1002/pro.4479 Medline
18. I. R. Humphreys, J. Zhang, M. Baek, Y. Wang, A. Krishnakumar, J. Pei, I. Anishchenko, C. A. Tower, B. A. Jackson, T. Warrier, D. T. Hung, S. B. Peterson, J. D. Mougous, Q. Cong, D. Baker, Protein interactions in human pathogens revealed through deep learning. *Nat. Microbiol.* **9**, 2642–2652 (2024). doi:10.1038/s41564-024-01791-x Medline
19. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022). doi:10.1093/nar/gkab1061 Medline
20. R. Guigó, Genome annotation: From human genetics to biodiversity genomics. *Cell Genom.* **3**, 100375 (2023). doi:10.1016/j.xgen.2023.100375 Medline
21. W. De Coster, M. H. Weissensteiner, F. J. Sedlazeck, Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021). doi:10.1038/s41576-021-00367-3 Medline
22. H. Li, Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023). doi:10.1093/bioinformatics/btad014 Medline
23. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005). doi:10.1186/1471-2105-6-31 Medline
24. D. P. Wall, H. B. Fraser, A. E. Hirsh, Detecting putative orthologs. *Bioinformatics* **19**, 1710–1711 (2003). doi:10.1093/bioinformatics/btg213 Medline
25. S. V. Rajagopala, P. Sikorski, A. Kumar, R. Mosca, J. Vlasblom, R. Arnold, J. Franca-Koh, S. B. Pakala, S. Phanse, A. Ceol, R. Häuser, G. Siszler, S. Wuchty, A. Emili, M. Babu, P. Aloy, R. Pieper, P. Uetz, The binary protein-protein interaction landscape of *Escherichia coli*. *Nat. Biotechnol.* **32**, 285–290 (2014). doi:10.1038/nbt.2831 Medline
26. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019). doi:10.1186/s12859-019-3019-7 Medline
27. S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, D. Baker, Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017). doi:10.1126/science.aah4043 Medline
28. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022). doi:10.1038/s41592-022-01488-1 Medline
29. H. Alexander, S. K. Hu, A. I. Krinos, M. Pachiadaki, B. J. Tully, C. J. Neely, T. Reiter, Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *mBio* **14**, e0167623 (2023). doi:10.1128/mbio.01676-23 Medline
30. S. E. Acuner Ozbabacan, H. B. Engin, A. Gursoy, O. Keskin, Transient protein-protein interactions. *Protein Eng. Des. Sel.* **24**, 635–648 (2011). doi:10.1093/protein/gzr025 Medline
31. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034 [Preprint] (2022); https://doi.org/10.1101/2021.10.04.463034.
32. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024). doi:10.1038/s41586-024-07487-w Medline
33. M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, M. Mirdita, J. Yeo, O. Kovalevskiy, K. Tunyasuvunakool, A. Laydon, A. Žídek, H. Tomlinson, D. Hariharan, J. Abrahamson, T. Green, J. Jumper, E. Birney, M. Steinegger, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024). doi:10.1093/nar/gkad1011 Medline
34. N. Sen, M. S. Madhusudhan, A structural database of chain–chain and domain–

domain interfaces of proteins. *Protein Sci.* **31**, e4406 (2022). doi:10.1002/pro.4406
35. R. Verma, S. B. Pandit, Unraveling the structural landscape of intra-chain domain interfaces: Implication in the evolution of domain-domain interactions. *PLOS ONE* **14**, e0220336 (2019). doi:10.1371/journal.pone.0220336 Medline
36. J. Zhang, R. D. Schaeffer, J. Durham, Q. Cong, N. V. Grishin, DPAM: A domain parser for AlphaFold models. *Protein Sci.* **32**, e4548 (2023). doi:10.1002/pro.4548 Medline
37. T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunić, A. Marchler-Bauer, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023). doi:10.1093/nar/gkac993 Medline
38. M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, M. Steinegger, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024). doi:10.1038/s41587-023-01773-0 Medline
39. M. Baek, I. Anishchenko, I. R. Humphreys, Q. Cong, D. Baker, F. DiMaio, Efficient and accurate prediction of protein structure using RoseTTAFold2. bioRxiv 2023.05.24.542179 [Preprint] (2023); https://doi.org/10.1101/2023.05.24.542179.
40. J. R. Perkins, I. Diboun, B. H. Dessailly, J. G. Lees, C. Orengo, Transient protein-protein interactions: Structural, functional, and network properties. *Structure* **18**, 1233–1243 (2010). doi:10.1016/j.str.2010.08.007 Medline
41. H. Schweke, M. Pacesa, T. Levin, C. A. Goverde, P. Kumar, Y. Duhoo, L. J. Dornfeld, B. Dubreuil, S. Georgeon, S. Ovchinnikov, D. N. Woolfson, B. E. Correia, S. Dey, E. D. Levy, An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999–1010.e15 (2024). doi:10.1016/j.cell.2024.01.022 Medline
42. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009). doi:10.1073/pnas.0805923106 Medline
43. K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, D. Choi, A. G. Coté, M. Daley, S. Deimling, A. Desbuleux, A. Dricot, M. Gebbia, M. F. Hardy, N. Kishore, J. J. Knapp, I. A. Kovács, I. Lemmens, M. W. Mee, J. C. Mellor, C. Pollis, C. Pons, A. D. Richardson, S. Schlabach, B. Teeking, A. Yadav, M. Babor, D. Balcha, O. Basha, C. Bowman-Colin, S.-F. Chin, S. G. Choi, C. Colabella, G. Coppin, C. D'Amata, D. De Ridder, S. De Rouck, M. Duran-Frigola, H. Ennajdaoui, F. Goebels, L. Goehring, A. Gopal, G. Haddad, E. Hatchi, M. Helmy, Y. Jacob, Y. Kassa, S. Landini, R. Li, N. van Lieshout, A. MacWilliams, D. Markey, J. N. Paulson, S. Rangarajan, J. Rasla, A. Rayhan, T. Rolland, A. San-Miguel, Y. Shen, D. Sheykhkarimli, G. M. Sheynkman, E. Simonovsky, M. Taşan, A. Tejeda, V. Tropepe, J.-C. Twizere, Y. Wang, R. J. Weatheritt, J. Weile, Y. Xia, X. Yang, E. Yeger-Lotem, Q. Zhong, P. Aloy, G. D. Bader, J. De Las Rivas, S. Gaudet, T. Hao, J. Rak, J. Tavernier, D. E. Hill, M. Vidal, F. P. Roth, M. A. Calderwood, A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020). doi:10.1038/s41586-020-2188-x Medline
44. E. L. Huttlin, R. J. Bruckner, J. Navarrete-Perea, J. R. Cannon, K. Baltier, F. Gebreab, M. P. Gygi, A. Thornock, G. Zarraga, S. Tam, J. Szpyt, B. M. Gassaway, A. Panov, H. Parzen, S. Fu, A. Golbazi, E. Maenpaa, K. Stricker, S. Guha Thakurta, T. Zhang, R. Rad, J. Pan, D. P. Nusinow, J. A. Paulo, D. K. Schweppe, L. P. Vaites, J. W. Harper, S. P. Gygi, Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022–3040.e28 (2021). doi:10.1016/j.cell.2021.04.011 Medline
45. S. Maiti, A. Singh, T. Maji, N. V. Saibo, S. De, Experimental methods to study the structure and dynamics of intrinsically disordered regions in proteins. *Curr. Res. Struct. Biol.* **7**, 100138 (2024). doi:10.1016/j.crstbi.2024.100138 Medline
46. Y. Bagheri, A. A. Ali, M. You, Current methods for detecting cell membrane transient interactions. *Front Chem.* **8**, 603259 (2020). doi:10.3389/fchem.2020.603259 Medline
47. J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, R. G. Schneider, A. W. Senior, J. Jumper, D. Hassabis, P. Kohli, Ž. Avsec, Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023). doi:10.1126/science.adg7492 Medline
48. H. Gao, T. Hamp, J. Ede, J. G. Schraiber, J. McRae, M. Singer-Berk, Y. Yang, A. S. D. Dietrich, P. P. Fiziev, L. F. K. Kuderna, L. Sundaram, Y. Wu, A. Adhikari, Y. Field, C. Chen, S. Batzoglou, F. Aguet, G. Lemire, R. Reimers, D. Balick, M. C. Janiak, M. Kuhlwilm, J. D. Orkin, S. Manu, A. Valenzuela, J. Bergman, M. Rousselle, F. E. Silva, L. Agueda, J. Blanc, M. Gut, D. de Vries, I. Goodhead, R. A. Harris, M. Raveendran, A. Jensen, I. S. Chuma, J. E. Horvath, C. Hvilsom, D. Juan, P. Frandsen, F. R. de Melo, F. Bertuol, H. Byrne, I. Sampaio, I. Farias, J. V. do Amaral, M. Messias, M. N. F. da Silva, M. Trivedi, R. Rossi, T. Hrbek, N. Andriaholinirina, C. J. Rabarivola, A. Zaramody, C. J. Jolly, J. Phillips-Conroy, G. Wilkerson, C. Abee, J. H. Simmons, E. Fernandez-Duque, S. Kanthaswamy, F. Shiferaw, D. Wu, L. Zhou, Y. Shao, G. Zhang, J. D. Keyyu, S. Knauf, M. D. Le, E. Lizano, S. Merker, A. Navarro, T. Bataillon, T. Nadler, C. C. Khor, J. Lee, P. Tan, W. K. Lim, A. C. Kitchener, D. Zinner, I. Gut, A. Melin, K. Guschanski, M. H. Schierup, R. M. D. Beck, G. Umapathy, C. Roos, J. P. Boubli, M. Lek, S. Sunyaev, A. O'Donnell-Luria, H. L. Rehm, J. Xu, J. Rogers, T. Marques-Bonet, K. K.-H. Farh, The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023). doi:10.1126/science.abn8197 Medline
49. J. Jänes, M. Müller, S. Selvaraj, D. Manoel, J. Stephenson, C. Gonçalves, A. Lafita, B. Polacco, K. Obernier, K. Alasoo, M. C. Lemos, N. Krogan, M. Martin, L. R. Saraiva, D. Burke, P. Beltrao, Predicted mechanistic impacts of human protein missense variants. bioRxiv 2024.05.29.596373 [Preprint] (2024); https://doi.org/10.1101/2024.05.29.596373.
50. J. J. Rocha, S. A. Jayaram, T. J. Stevens, N. Muschalik, R. D. Shah, S. Emran, C. Robles, M. Freeman, S. Munro, Functional unknomics: Systematic screening of conserved genes of unknown function. *PLOS Biol.* **21**, e3002222 (2023). doi:10.1371/journal.pbio.3002222 Medline
51. A. S. Hauser, M. M. Attwood, M. Rask-Andersen, H. B. Schiöth, D. E. Gloriam, Trends in GPCR drug discovery: New agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017). doi:10.1038/nrd.2017.178 Medline
52. A. Sitaram, M. S. Marks, Mechanisms of protein delivery to melanosomes in pigment cells. *Physiology* **27**, 85–99 (2012). doi:10.1152/physiol.00043.2011 Medline
53. X. X. Li, J. D. Lee, C. Kemper, T. M. Woodruff, The complement receptor C5aR2: A powerful modulator of innate and adaptive immunity. *J. Immunol.* **202**, 3339–3348 (2019). doi:10.4049/jimmunol.1900371 Medline
54. T. Tralau, U. Meyer-Hoffert, J.-M. Schröder, O. Wiedow, Human leukocyte elastase and cathepsin G are specific inhibitors of C5a-dependent neutrophil enzyme release and chemotaxis. *Exp. Dermatol.* **13**, 316–325 (2004). doi:10.1111/j.0906-6705.2004.00145.x Medline
55. T. Liuyu, K. Yu, L. Ye, Z. Zhang, M. Zhang, Y. Ren, Z. Cai, Q. Zhu, D. Lin, B. Zhong, Induction of OTUD4 by viral infection promotes antiviral responses through deubiquitinating and stabilizing MAVS. *Cell Res.* **29**, 67–79 (2019). doi:10.1038/s41422-018-0107-6 Medline
56. Y. Guo, R. Li, Z. Tan, J. Shi, Y. Fu, Y. Song, M. Zhu, L. Zhang, J. Huang, E3 ubiquitin ligase ASB8 negatively regulates interferon via regulating TBK1/IKKi homeostasis. *Mol. Immunol.* **121**, 195–203 (2020). doi:10.1016/j.molimm.2020.03.011 Medline
57. S. M. Borys, A. K. Bag, L. Brossay, D. O. Adeegbe, The yin and yang of targeting KLRG1^{+} Tregs and effector cells. *Front. Immunol.* **13**, 894508 (2022). doi:10.3389/fimmu.2022.894508 Medline
58. M. Matsumoto, H. Oshiumi, T. Seya, Antiviral responses induced by the TLR3 pathway. *Rev. Med. Virol.* **21**, 67–77 (2011). doi:10.1002/rmv.680 Medline
59. M. Genova, L. Grycova, V. Puttrich, M. M. Magiera, Z. Lansky, C. Janke, M. Braun, Tubulin polyglutamylation differentially regulates microtubule-interacting proteins. *EMBO J.* **42**, e112101 (2023). doi:10.15252/embj.2022112101 Medline
60. N. Peng, F. Nakamura, Microtubule-associated proteins and enzymes modifying tubulin. *Cytoskeleton* **80**, 60–76 (2023). doi:10.1002/cm.21748 Medline
61. T. Oda, A. Fukuda, H. Hagiwara, Y. Masuho, M.-A. Muramatsu, K. Hisatake, T. Yamashita, ABT1-associated protein (ABTAP), a novel nuclear protein conserved from yeast to mammals, represses transcriptional activation by ABT1. *J. Cell. Biochem.* **93**, 788–806 (2004). doi:10.1002/jcb.20114 Medline
62. K. Dörner, C. Ruggeri, I. Zemp, U. Kutay, Ribosome biogenesis factors-from names to functions. *EMBO J.* **42**, e112699 (2023). doi:10.15252/embj.2022112699 Medline
63. G. Martinez, J. Beurois, D. Dacheux, C. Cazin, M. Bidart, Z.-E. Kherraf, D. R.

Robinson, V. Satre, G. Le Gac, C. Ka, I. Gourlaouen, Y. Fichou, G. Petre, E. Dulioust, R. Zouari, N. Thierry-Mieg, A. Touré, C. Arnoult, M. Bonhivers, P. Ray, C. Coutton, Biallelic variants in *MAATS1* encoding CFAP91, a calmodulin-associated and spoke-associated complex protein, cause severe astheno-teratozoospermia and male infertility. *J. Med. Genet.* **57**, 708–716 (2020). doi:10.1136/jmedgenet-2019-106775 Medline

64. M. Arafat, A. Harlev, I. Har-Vardi, E. Levitas, T. Priel, M. Gershoni, C. Searby, V. C. Sheffield, E. Lunenfeld, R. Parvari, Mutation in *CATIP* (C2orf62) causes oligoteratoasthenozoospermia by affecting actin dynamics. *J. Med. Genet.* **58**, 106–115 (2020). doi:10.1136/jmedgenet-2019-106825 Medline
65. Y. Sha, W. Liu, S. Li, L. V. Osadchuk, Y. Chen, H. Nie, S. Gao, L. Xie, W. Qin, H. Zhou, L. Li, Deficiency in AK9 causes asthenozoospermia and male infertility by destabilising sperm nucleotide homeostasis. *EBioMedicine* **96**, 104798 (2023). doi:10.1016/j.ebiom.2023.104798 Medline
66. B. H. M. Meldal, H. Bye-A-Jee, L. Gajdoš, Z. Hammerová, A. Horácková, F. Melicher, L. Perfetto, D. Pokorný, M. R. Lopez, A. Türková, E. D. Wong, Z. Xie, E. B. Casanova, N. Del-Toro, M. Koch, P. Porras, H. Hermjakob, S. Orchard, Complex Portal 2018: Extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.* **47**, D550–D558 (2019). doi:10.1093/nar/gky1001 Medline
67. K. Izumikawa, H. Ishikawa, H. Yoshikawa, S. Fujiyama, A. Watanabe, H. Aburatani, H. Tachikawa, T. Hayano, Y. Miura, T. Isobe, R. J. Simpson, L. Li, J. Min, N. Takahashi, LYAR potentiates rRNA synthesis by recruiting BRD2/4 and the MYST-type acetyltransferase KAT7 to rDNA. *Nucleic Acids Res.* **47**, 10357–10372 (2019). doi:10.1093/nar/gkz747 Medline
68. S. Grill, J. Nandakumar, Molecular mechanisms of telomere biology disorders. *J. Biol. Chem.* **296**, 100064 (2021). doi:10.1074/jbc.REV120.014017 Medline
69. T. Kinoshita, Biosynthesis and biology of mammalian GPI-anchored proteins. *Open Biol.* **10**, 190290 (2020). doi:10.1098/rsob.190290 Medline
70. Y.-S. Liu, Y. Wang, X. Zhou, L. Zhang, G. Yang, X.-D. Gao, Y. Murakami, M. Fujita, T. Kinoshita, Accumulated precursors of specific GPI-anchored proteins upregulate GPI biosynthesis with ARV1. *J. Cell Biol.* **222**, e202208159 (2023). doi:10.1083/jcb.202208159 Medline
71. J. Gonçalves, L. Pelletier, The ciliary transition zone: Finding the pieces and assembling the gate. *Mol. Cells* **40**, 243–253 (2017). doi:10.14348/molcells.2017.0054 Medline
72. C. Li, V. L. Jensen, K. Park, J. Kennedy, F. R. Garcia-Gonzalo, M. Romani, R. De Mori, A.-L. Bruel, D. Gaillard, B. Doray, E. Lopez, J.-B. Rivière, L. Faivre, C. Thauvin-Robinet, J. F. Reiter, O. E. Blacque, E. M. Valente, M. R. Leroux, MKS5 and CEP290 dependent assembly pathway of the ciliary transition zone. *PLOS Biol.* **14**, e1002416 (2016). doi:10.1371/journal.pbio.1002416 Medline
73. J. Zhang, I. Humphreys, J. Pei, J. Kim, C. Choi, R. Yuan, J. Durham, S. Liu, H.-J. Choi, M. Baek, D. Baker, Q. Cong, RoseTTAFold2-PPI: A fast deep learning method for large-scale protein-protein interaction screening. Zenodo (2025); https://doi.org/10.5281/zenodo.15314316.
74. J. Zhang, I. Humphreys, J. Pei, J. Kim, C. Choi, R. Yuan, J. Durham, S. Liu, H. J. Choi, M. Baek, D. Baker, Q. Cong, A dataset for predicting protein-protein interactions in humans. Dryad (2025); https://doi.org/10.5061/dryad.15dv41p84.
75. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017). doi:10.1038/nbt.3988 Medline
76. S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, R. D. Finn, HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018). doi:10.1093/nar/gky448 Medline
77. T. A. Thanaraj, F. Clark, J. Muilu, Conservation of human alternative splice events in mouse. *Nucleic Acids Res.* **31**, 2544–2552 (2003). doi:10.1093/nar/gkg355 Medline
78. B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021). doi:10.1038/s41592-021-01101-x Medline
79. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995). doi:10.1016/S0022-2836(05)80134-2 Medline
80. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, CATH—A hierarchic classification of protein domain structures. *Structure* **5**, 1093–1109 (1997). doi:10.1016/S0969-2126(97)00260-8 Medline
81. G. Apic, J. Gough, S. A. Teichmann, Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325 (2001). doi:10.1006/jmbi.2001.4776 Medline
82. C. Chothia, J. Gough, Genomic and structural aspects of protein evolution. *Biochem. J.* **419**, 15–28 (2009). doi:10.1042/BJ20090122 Medline
83. H. Cheng, R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B.-H. Kim, N. V. Grishin, ECOD: An evolutionary classification of protein domains. *PLOS Comput. Biol.* **10**, e1003926 (2014). doi:10.1371/journal.pcbi.1003926 Medline
84. L. Holm, Using Dali for protein structure comparison. *Methods Mol. Biol.* **2112**, 29–42 (2020). doi:10.1007/978-1-0716-0270-6_3 Medline
85. T. Paysan-Lafosse, A. Andreeva, M. Blum, S. R. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, F. Llinares-López, L. Meng-Papaxanthos, L. J. Colwell, N. V. Grishin, R. D. Schaeffer, D. Clementel, S. C. E. Tosatto, E. Sonnhammer, V. Wood, A. Bateman, The Pfam protein families database: Embracing AI/ML. *Nucleic Acids Res.* **53**, D523–D534 (2025). doi:10.1093/nar/gkae997 Medline
86. R. D. Schaeffer, J. Zhang, L. N. Kinch, J. Pei, Q. Cong, N. V. Grishin, Classification of domains in predicted structures of the human proteome. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2214069120 (2023). doi:10.1073/pnas.2214069120 Medline
87. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2 Medline
88. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011). doi:10.1038/nmeth.1818 Medline
89. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919 (1992). doi:10.1073/pnas.89.22.10915 Medline
90. F. Morcos, T. Hwa, J. N. Onuchic, M. Weigt, Direct coupling analysis for protein contact prediction. *Methods Mol. Biol.* **1137**, 55–70 (2014). doi:10.1007/978-1-4939-0366-5_5 Medline
91. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (2018). doi:10.1111/j.2517-6161.1995.tb02031.x
92. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003). doi:10.1101/gr.1239303 Medline
93. B. H. M. Meldal, L. Perfetto, C. Combe, T. Lubiana, J. V. Ferreira Cavalcante, H. Bye-A-Jee, A. Waagmeester, N. Del-Toro, A. Shrivastava, E. Barrera, E. Wong, B. Mlecnik, G. Bindea, K. Panneerselvam, E. Willighagen, J. Rappsilber, P. Porras, H. Hermjakob, S. Orchard, Complex Portal 2022: New curation frontiers. *Nucleic Acids Res.* **50**, D578–D586 (2022). doi:10.1093/nar/gkab991 Medline
94. J. Szymborski, A. Emad, RAPPPID: Towards generalizable protein interaction prediction with AWD-LSTM twin networks. *Bioinformatics* **38**, 3958–3967 (2022). doi:10.1093/bioinformatics/btac429 Medline
95. S. Sledzieski, R. Singh, L. Cowen, B. Berger, D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst.* **12**, 969–982.e6 (2021). doi:10.1016/j.cels.2021.08.010 Medline
96. Y. S. Ko, J. Parkinson, C. Liu, W. Wang, TUnA: An uncertainty-aware transformer model for sequence-based protein-protein interaction prediction. *Brief. Bioinform.* **25**, bbae359 (2024). doi:10.1093/bib/bbae359 Medline
97. J. Bernett, D. B. Blumenthal, M. List, Cracking the black box of deep sequence-based protein-protein interaction prediction. *Brief. Bioinform.* **25**, bbae076 (2024). doi:10.1093/bib/bbae076 Medline
98. J. Chen, N. Sawyer, L. Regan, Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.* **22**, 510–515 (2013). doi:10.1002/pro.2230 Medline
99. D. Xiong, Y. Qiu, J. Zhao, Y. Zhou, D. Lee, S. Gupta, M. Torres, W. Lu, S. Liang, J. J. Kang, C. Eng, J. Loscalzo, F. Cheng, H. Yu, A structurally informed human protein-protein interactome reveals proteome-wide perturbations caused by disease mutations. *Nat. Biotechnol.* **43**, 1510–1524 (2025). doi:10.1038/s41587-024-02428-4 Medline

100. D. C. Jeffares, B. Tomiczek, V. Sojo, M. dos Reis, A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol. Biol.* **1201**, 65–90 (2015). doi:10.1007/978-1-4939-1438-8_4 Medline
101. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, B. M. Neale, M. J. Daly, D. G. MacArthur; Genome Aggregation Database Consortium, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi:10.1038/s41586-020-2308-7 Medline
102. Y. P. Conley, J. Jakobsdottir, T. Mah, D. E. Weeks, R. Klein, L. Kuller, R. E. Ferrell, M. B. Gorin, *CFH*, *ELOVL4*, *PLEKHA1* and *LOC387715* genes and susceptibility to age-related maculopathy: AREDS and CHS cohorts and meta-analyses. *Hum. Mol. Genet.* **15**, 3206–3218 (2006). doi:10.1093/hmg/ddl396 Medline

ACKNOWLEDGMENTS

The authors thank Nick V. Grishin, Jian Zhou, Rohith Krishna, Frank Dimaio, Edin Muratspahic, and Helen H. Hobbs for inspiring discussions and insightful suggestions. We also thank Luki Goldschmidt and Aaron Guillory for computing resource management, and Lynda Stewart and Lance Stuart for logistical support. **Funding:** QC is a Southwestern Medical Foundation Endowed Scholar and DB is a Howard Hughes Medical Institute Investigator. This research is supported by I-2095-20220331 and V-I-0004-20230731 from the Welch Foundation; 1K99AI180984-01A1, 5-R01-HL-159946-03, and contract No. 75N93022C00036 from NIH, Department of Health and Human Services; Bill & Melinda Gates Foundation Investments INV-010680 and INV-043758; Novo Nordisk A/S contract RDC 2022-002902; NRF/MSIT (RS-2023-00210147, RS-2024-00397865, RS-2024-00440824, and RS-2024-00407331); IITP/MSIT (RS-2023-00220628); KBSI(NFEC)/MSIT (RS-2024-00401698). We acknowledge NSF ACCESS (MED230034 and BIO250039), TACC Lonestar6 (MCB24018), and Perlmutter grant NERSC award BER-ERCAP0022018 for providing the high-performance computing resources that were critical to the success of this research. **Author contributions:** Conceptualization: DB, QC; Formal analysis: JZ, JP, QC; Result interpretation: JZ, IRH, JP, JK, CC, RY, JD, SL, HJC, QC; Funding acquisition: DB, QC; Investigation: JZ, IRH, JP, QC; Methodology: JZ, IRH, JP, MB, QC; Project administration: QC; Supervision: QC; Visualization: JZ, IRH, JP, QC; Writing – original draft: JZ, IRH, JP, QC; Writing – review & editing: JZ, IRH, JP, JK, CC, RY, JD, SL, HC, MB, DB, QC. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** We present our predictions and intermediate results at http://prodata.swmed.edu/humanPPI, which allows users to easily navigate through our findings and accelerate future discoveries. We deposited the RF2-PPI network at Zenodo (*73*) and https://github.com/CongLabCode/RoseTTAFold2-PPI. We share our omicMSAs, training datasets, trained weights of RF2-PPI, predicted PPIs, and 3D models of human protein complexes at https://conglab.swmed.edu/humanPPI/humanPPI_download.html and Dryad (*74*). To use our omicMSA to build protein complexes with AF2 or AFmm, please use our colab notebook at: https://colab.research.google.com/drive/1suhoIB5q6xn0APFHJE8c1eMiCuv9gCk_.

SUPPLEMENTARY MATERIALS

Submitted 27 September 2024; accepted 20 August 2025
Published online 25 September 2025
10.1126/science.adt1630

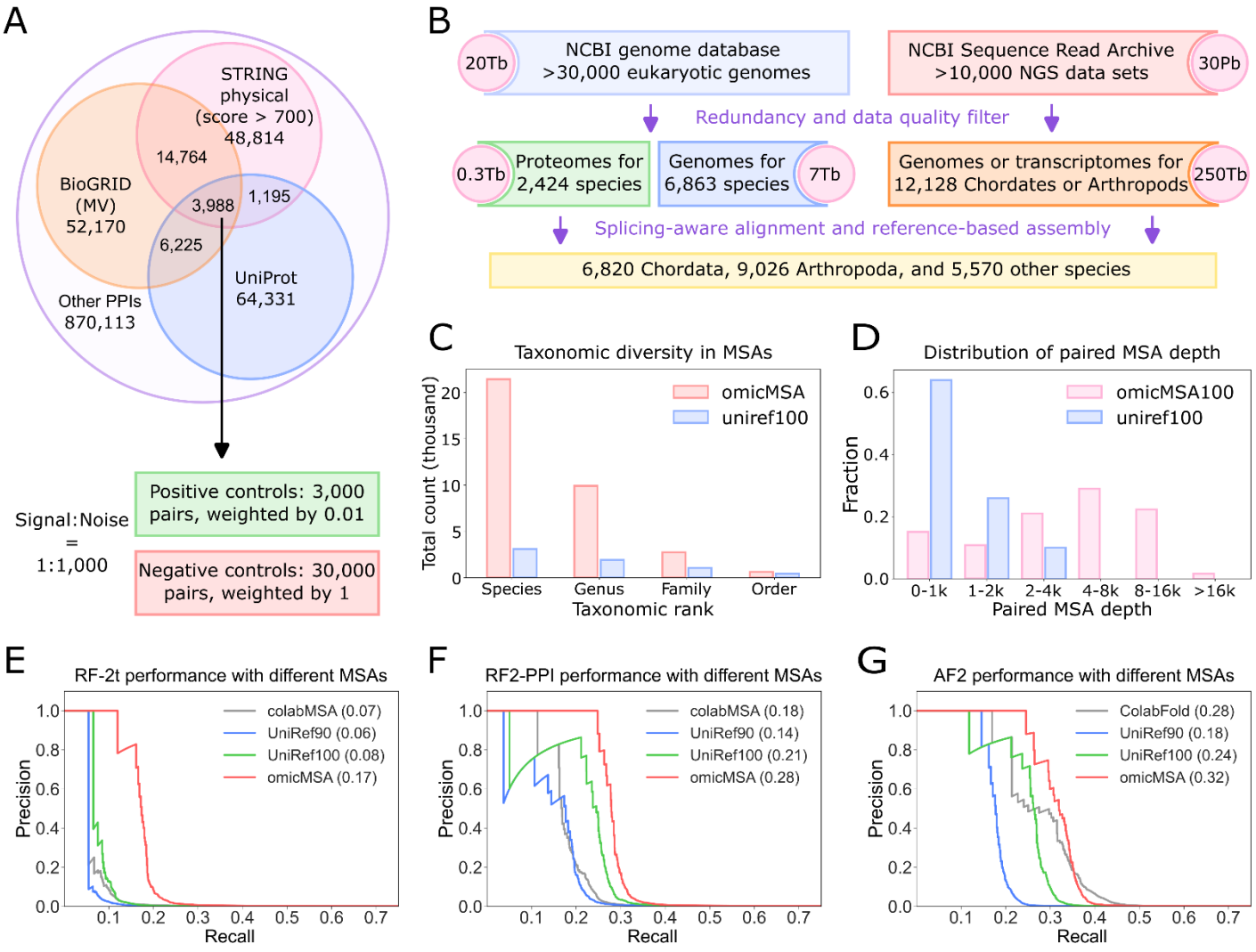

Fig. 1. OmicMSAs enrich coevolutionary signals and improve PPI identification accuracy. (**A**) Top: Venn diagram showing the intersections of confident PPIs across three databases, including STRING physical interactions showing a combined STRING score above 700, BioGRID multi-validated PPIs, and UniProt PPIs. The intersection of the three PPI sources was used to randomly select 3,000 positive controls. 30,000 random protein pairs showing no previous evidence of interaction are used as negative controls. Each positive pair is weighted by 0.01 for performance evaluation to reflect a signal-to-noise ratio of 1:1000. (**B**) Strategies to obtain omicMSAs by assembling protein sequences from genomic data at NCBI. Boxes show dataset statistics, ovals indicate dataset size (T: terabytes, P: petabytes), and purple text describes the data processing procedures. (**C**) Taxonomic diversity of omicMSAs derived from eukaryotic genomic data and eukaryotic protein sequences in UniRef100. (**D**) Distribution of pMSA depth (filtered at 100% identity) for protein pairs in the positive and negative control sets. (**E** to **G**) Precision-recall curves at a signal-to-noise ratio of 1:1000. The positive and negative pairs are ranked reversely by interaction probabilities computed by (E) RF-2t, the DL network we used previously to screen yeast PPIs (*14*), (F) RF2-PPI, and (G) AF2; at each rank, the precision and recall are calculated as $precision = \frac{TP \cdot wp}{(TP \cdot wp + FP)}$ and $recall = \frac{TP}{P}$. *TP* and *FP* are numbers of positive and negative controls, respectively, above this rank; *P* is the total number of positive controls; *wp* = 0.01.

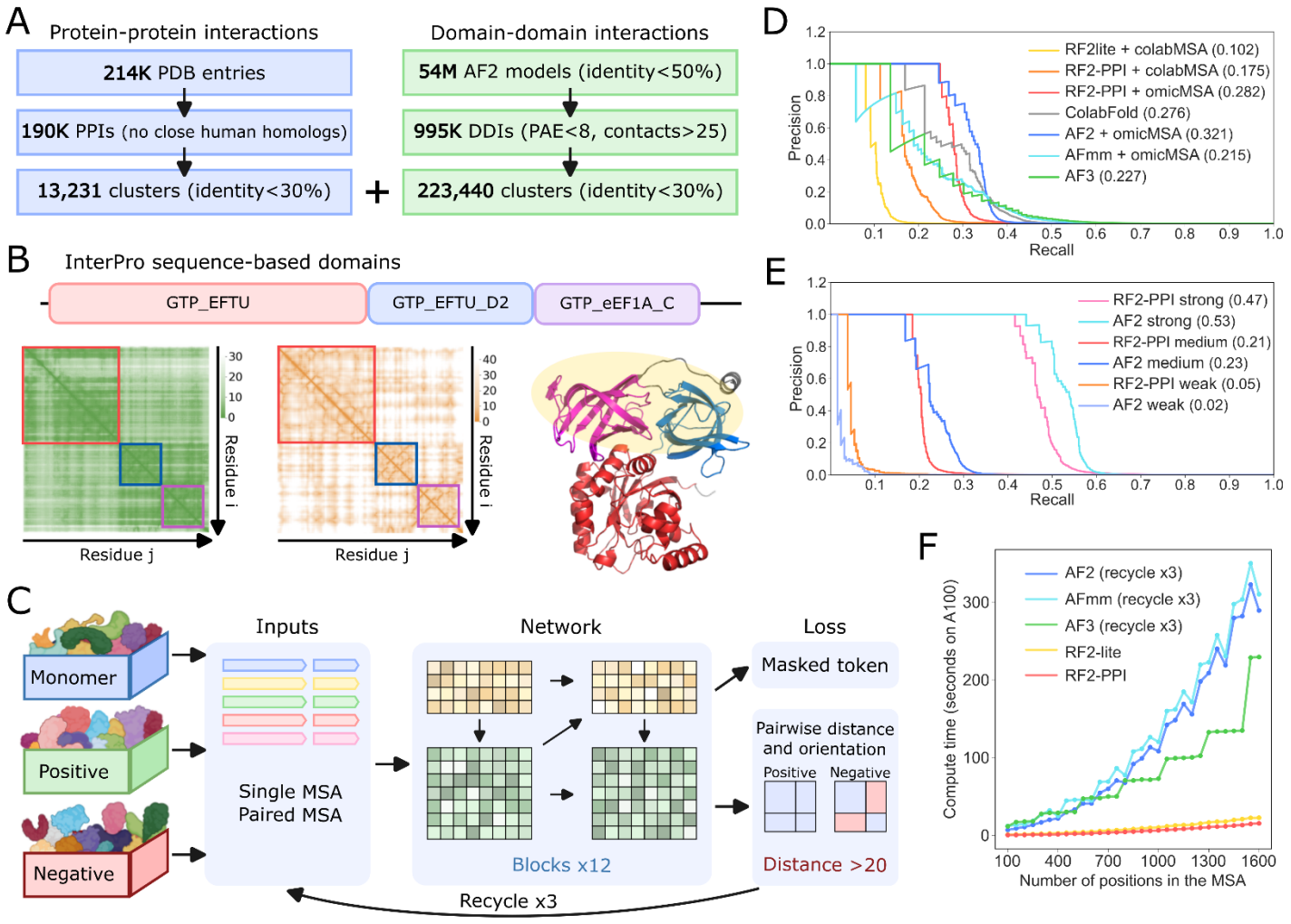

Fig. 2. DDIs distilled from AFDB enable a DL network for rapid PPI prediction. (**A**) Statistics of the PPI and DDI training datasets. Close homologs of human protein pairs were excluded to prevent potential information leakage, and only confident DDIs, measured by low inter-domain PAE, were included. (**B**) Illustration of our domain segmentation protocol. Top: domains annotated by InterPro; left: PAE matrix; middle: inter-residue distance matrix; right: 3D structure. In both matrices, the parsed domains are marked by boxes colored the same as in the 3D structure. The two domains in yellow shade belong to the DDI training set, while their interaction with the 3rd domain did not pass our inter-domain PAE cutoff. (**C**) Architecture and training routine for RF2-PPI. RF2-PPI contains 12 Evoformer-like blocks, takes only MSAs as inputs, and is trained using masked token, pairwise distance, and pairwise orientation losses. (**D**) Precision-recall curves for different methods in their ability to distinguish true PPIs from false ones at the signal-to-noise ratio of 1:1000. These curves were computed as described in Fig. 1. (**E**) Precision-recall curves for human PPIs with PDB structures stratified by interface size, defined as the average number of interface residues (distance to another protein < 6 Å) of the two proteins. Interactions were classified as weak (≤10 residues), medium (11–50 residues), or strong (>50 residues). PPIs in each category are considered as positive controls, respectively, and the negatives are the same as in Fig. 1. Precision and recall are calculated as $precision = \frac{TP \cdot wp}{(TP \cdot wp + FP)}$ and $recall = \frac{TP}{P}$. The weight for positives, *wp*, is introduced to adjust the signal-to-noise ratio to 1:1000; it is calculated by $\frac{P \cdot wp}{N} = \frac{1}{1000}$, where *N* is the size of the negative control set, i.e., 30000, and *P* is the size of a positive control set. (**F**) Compute time as a function of input protein size for different methods.

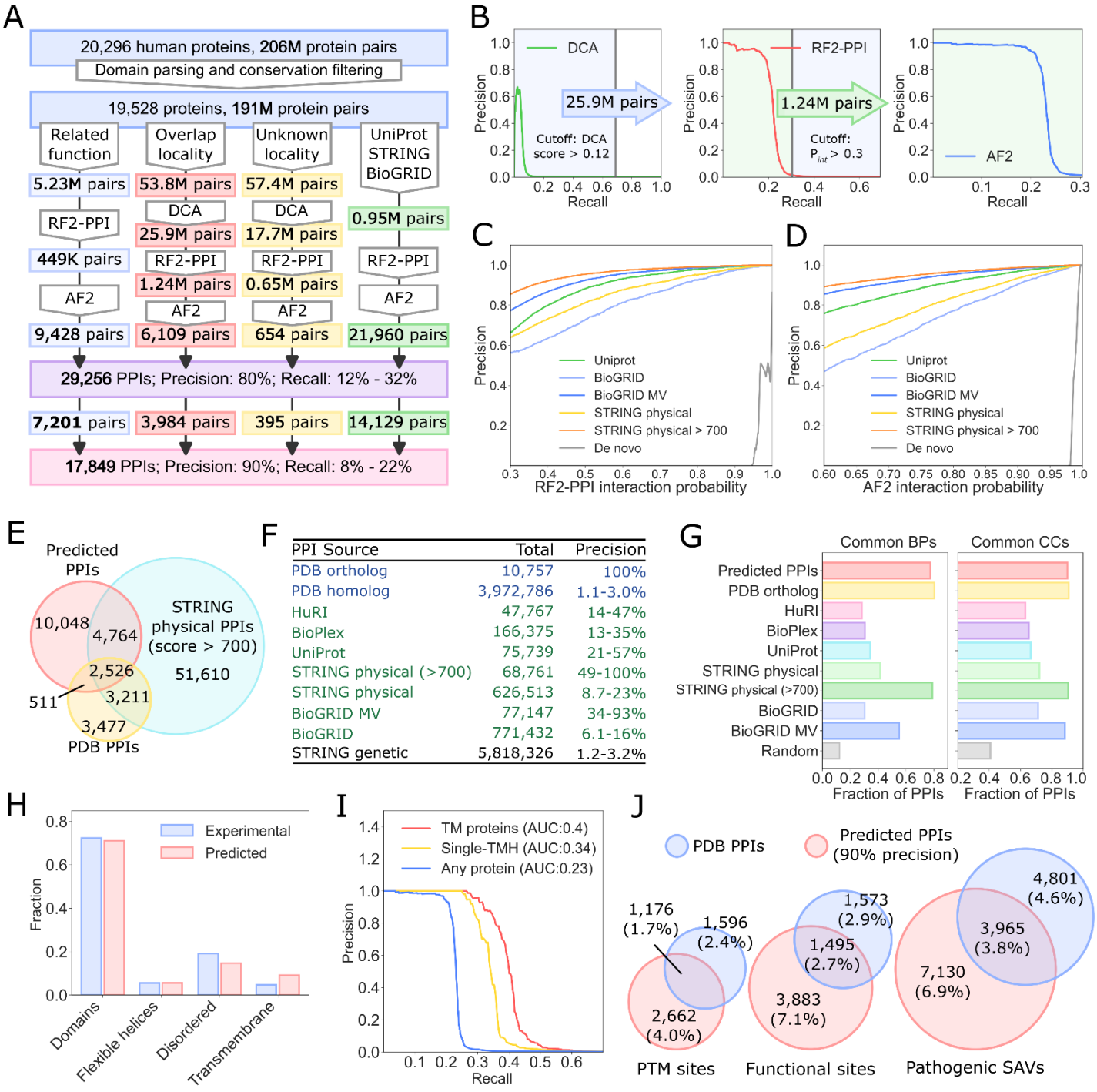

Fig. 3. Proteome-wide PPI screen pipeline and statistics. (**A**) PPI screen strategies. Key methods are shown in white connectors and the numbers of selected pairs after key steps are shown in colored boxes. (**B**) Precision and recall curves (see supplementary methods M5.6 for details) of our de novo pipeline applied to colocalized protein pairs. The grey line indicates the threshold (corresponding score labeled by the line) at each stage. The number of pairs selected at each stage is shown in the arrow. (**C** and **D**) Estimated precision as a function of (C) RF2-PPI and (D) AF2 interaction probability for protein pairs from different sources. The de novo set includes random pairs. (**E**) A Venn diagram showing the overlap between our predictions and PPIs from other sources. (**F**) The precision of PPIs in other databases. PDB ortholog, with orthologous PDB templates; PDB homolog, with homologous PDB templates. (**G**) Predicted PPIs show a much higher chance to participate in the same biological processes (BP) and localize in the same cellular components (CC) than human PPIs in most other databases. (**H**) Fraction of different interface types in predicted and experimentally determined PPIs. Domains: >50% interface residues are in domains for both; flexible helices: >50% interface residues are in helical inter-domain linkers for at least one; disordered regions: >50% interface residues are in inter-domain loops for at least one; transmembrane: >25% interface residues belong to TMHs for at least one. (**I**) Performance of our de novo pipeline for TMPs evaluated using positive and negative controls of different types of proteins. (**J**) Venn diagrams showing the numbers and proportions of PTM sites (left), functional sites (including active and cofactor/substrate binding sites; middle), and disease-associated SAVs (right) at PPI interfaces identified in experimental structures or predicted in this study.

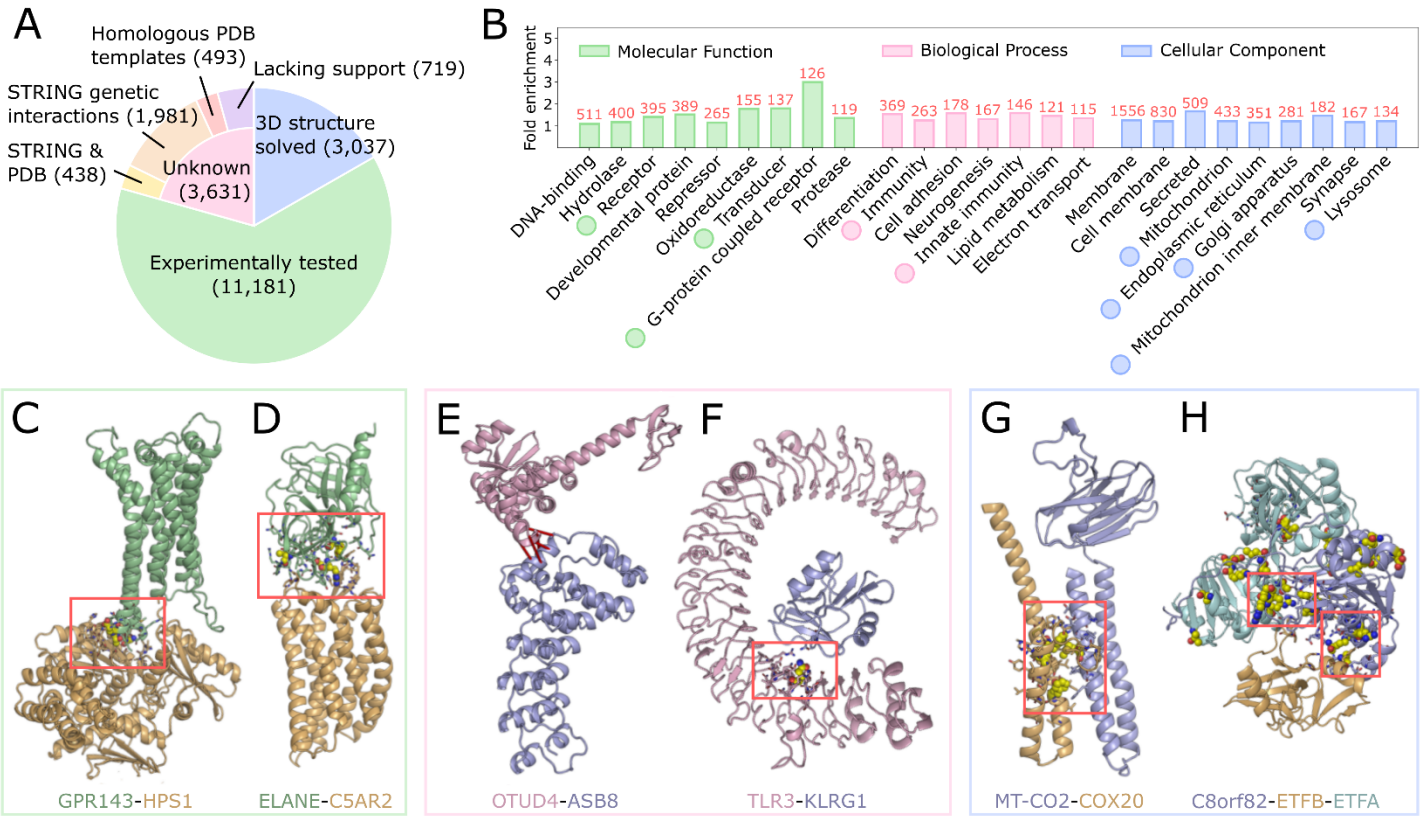

Fig. 4. Functional insights revealed by predicted PPIs. (**A**) A pie chart showing the fraction of predicted PPIs with experimental evidence. 3D structure solved: with orthologous PDB template; experimentally tested: present in UniProt, STRING (physical), or BioGRID databases, but 3D structures not resolved; the remaining unknown PPIs are further partitioned by whether they show evidence of genetic interactions, have homologous templates, or both. (**B**) Functional categories (UniProt keywords) enriched (P-value < 0.01) with least known PPIs lacking experimental support and structural templates (numbers labeled above the bars). Dots by the labels mark categories discussed in the main text. (**C** and **D**) PPIs involving GPCRs. (**E** and **F**) PPIs related to immunity. (**G** and **H**) Mitochondrial PPIs. In (C) to (H), yellow spheres indicate interface residues associated with disease-causing SAVs; otherwise, red bars connect residues with their most confidently predicted inter-protein contacts.

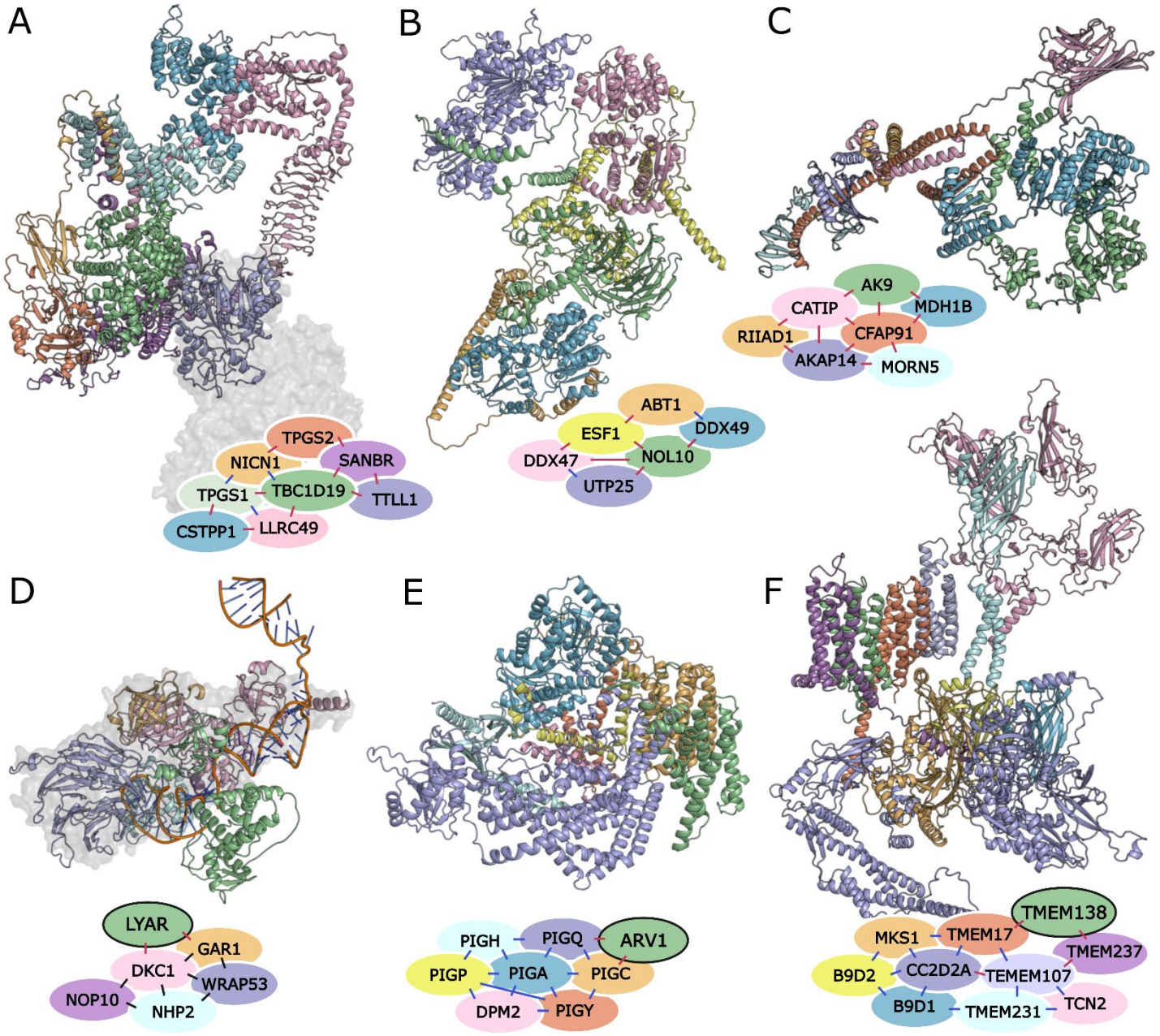

Fig. 5. Predicted higher-order oligomeric protein complexes. Each panel contains the 3D structure of a complex and a network graph. Nodes in the graph are labeled by gene names and colored the same as the corresponding subunits in the 3D structure. Edges in the graph represent PPIs and are colored as follows: black, PPIs with experimental structures; blue, confident PPIs from databases; red: uncharacterized PPIs or with weak experimental support. Examples of predicted oligomeric complexes are shown in (A) to (C). (Examples of predicted additional components of known complexes are shown in (D) to (F). (**A**) The tubulin polyglutamylase complex. An experimental structure (PDB: 5jh7) of the catalytic subunit and its substrate (tubulins) is shown as surface representation and colored in gray. (**B**) A complex connecting basal transcription regulation with ribosome biosynthesis. (**C**) A complex related to cilia/flagella biogenesis. (**D**) Adding LYAR to the telomere maintenance complex; an experimental structure of this complex (PDB: 8ouf) is shown as surface representation and colored in gray. (**E**) Adding ARV1 to a GPI anchor biosynthesis complex. (**F**) Adding TMEM138 to the transition zone complex.