Human protein interactome structure prediction at scale with Boltz-2

Alexander M. Ille[1,2], Christopher Markosian[1,2], Stephen K. Burley[3,4,5,6,7,*], Renata Pasqualini[1,2,9,*], and Wadih Arap[1,8,9,*]

[1]Rutgers Cancer Institute, Newark, NJ, USA
[2]Division of Cancer Biology, Department of Radiation Oncology, Rutgers New Jersey Medical School, Newark, NJ, USA
[3]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[4]Rutgers Artificial Intelligence and Data Science Collaboratory, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[5]Rutgers Cancer Institute, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
[6]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[7]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California-San Diego, La Jolla, San Diego, CA, USA
[8]Division of Hematology/Oncology, Department of Medicine, Rutgers New Jersey Medical School, Newark, NJ, USA
[9]These authors jointly supervised the work.

*Correspondence to: Stephen K. Burley (sburley@proteomics.rutgers.edu), Renata Pasqualini (renata.pasqualini@rutgers.edu), or Wadih Arap (wadih.arap@rutgers.edu)

ABSTRACT

In humans, protein-protein interactions mediate numerous biological processes and are central to both normal physiology and disease. Extensive research efforts have aimed to elucidate the human protein interactome, and comprehensive databases now catalog these interactions at scale. However, structural coverage of the human protein interactome is limited and remains challenging to resolve through experimental methodology alone. Recent advances in artificial intelligence/machine learning (AI/ML)-based approaches for protein interaction structure prediction present opportunities for large-scale structural characterization of the human interactome. One such model, Boltz-2, which is capable of predicting the structures of protein complexes, may serve this objective. Here, we present *de novo* computed models of 1,394 binary human protein interaction structures predicted using Boltz-2 based on biochemically determined interaction data sourced from the IntAct database. We assessed the predicted interaction structures through different confidence metrics, which consider both overall structure and the interaction interface. These analyses indicated that prediction confidence tended to be greater for smaller complexes, while increased multiple sequence alignment (MSA) depth tended to improve prediction confidence. Additionally, we examined annotated protein domains and found that 679 of the predicted structural complexes contained a variety of domains with putative interaction involvement on the basis of interaction interface proximity. Furthermore, our analyses revealed intricate interaction networks within the context of biological function and cancer. This work demonstrates the utility of Boltz-2 for *in silico* structural modeling of the human protein interactome, highlighting both strengths and limitations, while also providing a novel view of broad functional contextualization. Ultimately, such modeling is expected to yield broad structural insights with relevance across multiple domains of biomedical research.

INTRODUCTION

Interactions between biological macromolecules are fundamental to life, mediating numerous processes in both physiologic and pathological contexts. In particular, protein-protein interactions (PPIs) have been extensively studied for the elucidation of a of numerous biological functions [1]. Multiple large-scale initiatives have focused on systematic identification and annotation of PPIs based on experimental (e.g., biochemical) evidence [2-7]. These efforts have produced data resources, which provide comprehensive characterization of the 'interactomes' of many species including human. In parallel, databases dedicated to sequence annotation and protein structure determination, namely UniProt [8] and the Protein Data Bank (PDB) [9], respectively, have also been critical to systematic profiling of PPIs.

In recent years, metagenomic sequence data and three-dimensional (3D) structure data from open-access repositories [8-11] have enabled development of artificial intelligence/machine learning (AI/ML)-based approaches for protein structure prediction [12-15], including for prediction of structures of biomolecular interactions [14, 16, 17]. AI/ML models trained to leverage the relationship between sequence and 3D structure [18, 19], including those using multiple sequence alignments (MSAs) [12, 14, 16], have exhibited unprecedented accuracy [20, 21] for prediction of both individual structures and those of multiprotein complexes from amino acid sequence alone. AlphaFold-Multimer [17], an initial approach for interaction structure prediction, motivated previous work aimed toward structurally resolving the human interaction network [17]. Since then, other AI/ML models, including AlphaFold3 [16] and more recently Boltz-2 [22], have been developed and exhibit substantially improved performance in predicting the structures of protein complexes.

Boltz-2 has drawn particular interest given its ability to predict binding affinities of small molecules interacting with proteins [22]. Moreover, Boltz-2 performs robustly in terms of predicting structures of protein-protein, protein-DNA, and protein-RNA complexes, as evaluated against other AI/ML models (including but not limited to AlphaFold3) on a range of complexes freely available from the PDB (following the training cut-off date) [22]. Herein, we leverage the protein-protein structure prediction capabilities of Boltz-2 for high-throughput modeling of n = 1,394 binary human protein interaction structures, based on biochemically determined PPIs sourced from the IntAct database. The human protein interactome was the focus of the current study due to direct relevance in biomedical research. We assessed predictions using different confidence metrics, considering both overall complex structure and the interaction interface of protein pairs, and found that prediction confidence tended to be greater for smaller complexes while increased MSA depth tended to improve prediction confidence. We also analyzed the predicted complexes for the presence interaction interface-proximal Pfam domains, which revealed putative interaction networks with biological function contextualization. These findings demonstrate the utility of Boltz-2, highlighting both strengths and limitations, for capturing the broad structural landscape of the

human protein interactome and exemplifying opportunities for large-scale structural and functional insights with applicability across the biomedical sciences.

RESULTS

Large-scale prediction of human protein interaction structures

To construct a representative dataset of the human interactome suitable for structural prediction, we coupled biochemically determined PPIs with their corresponding amino acid sequences. PPIs were sourced from the IntAct database [2] and sequences from UniProt [8], with a binary arrangement of interactor A paired with interactor B **(Figure 1a)**. PPIs were filtered to include only interactions in which both protein interactors were human and designated as having direct physical binding as determined by pull-down assay using purified proteins. Boltz-2 [22] was then used to perform structural predictions of these binary interactions (see Methods). Predictions were performed for protein pairs with individual interactor sequence lengths up to 1,000 residues, providing a dataset consisting of a total of n = 1,394 predicted protein interaction structures. For example, the predicted complex structure with the greatest confidence score is presented in **Figure 1b**.

With the scope of better characterizing the set of predicted interaction structures, we explored sequence-level attributes of the interacting proteins **(Supplementary Table 1)**. Bivariate analysis of the sequence lengths across the protein interactor pairs used for structural prediction revealed a relatively uniform distribution **(Figure 1c)**. This pattern did not indicate any strong bias toward specific sequence length combinations other than most sequence pairs having lengths greater than 100 amino acid residues. For additional biological context, the interactor pair sequences used for structural prediction were queried against the Pfam database [23, 24] to identify co-occurrence of functional protein domains **(Figure 1d)**. In total, there were 2,391 unique domain co-occurrences, highlighting rich biological variability among the predicted structures. The most abundant co-occurring domains were PF00385 (chromatin organization modifier domain) with PF13923 (zinc finger, C3HC4 type (RING finger)) and PF00385 (chromatin organization modifier domain) with PF16207 (RAWUL domain RING finger- and WD40-associated ubiquitin-like), each present in 15 protein interactor pairs. Notably, the domain PF02991 (autophagy protein Atg8 ubiquitin like) was present in 173 protein interactor pairs. It should be noted that the presence of these domains does not imply interaction involvement in the interactor pairs in which they are found—*i.e.* these domains may or may not be located at the interaction interface, which is addressed in more detail later on.

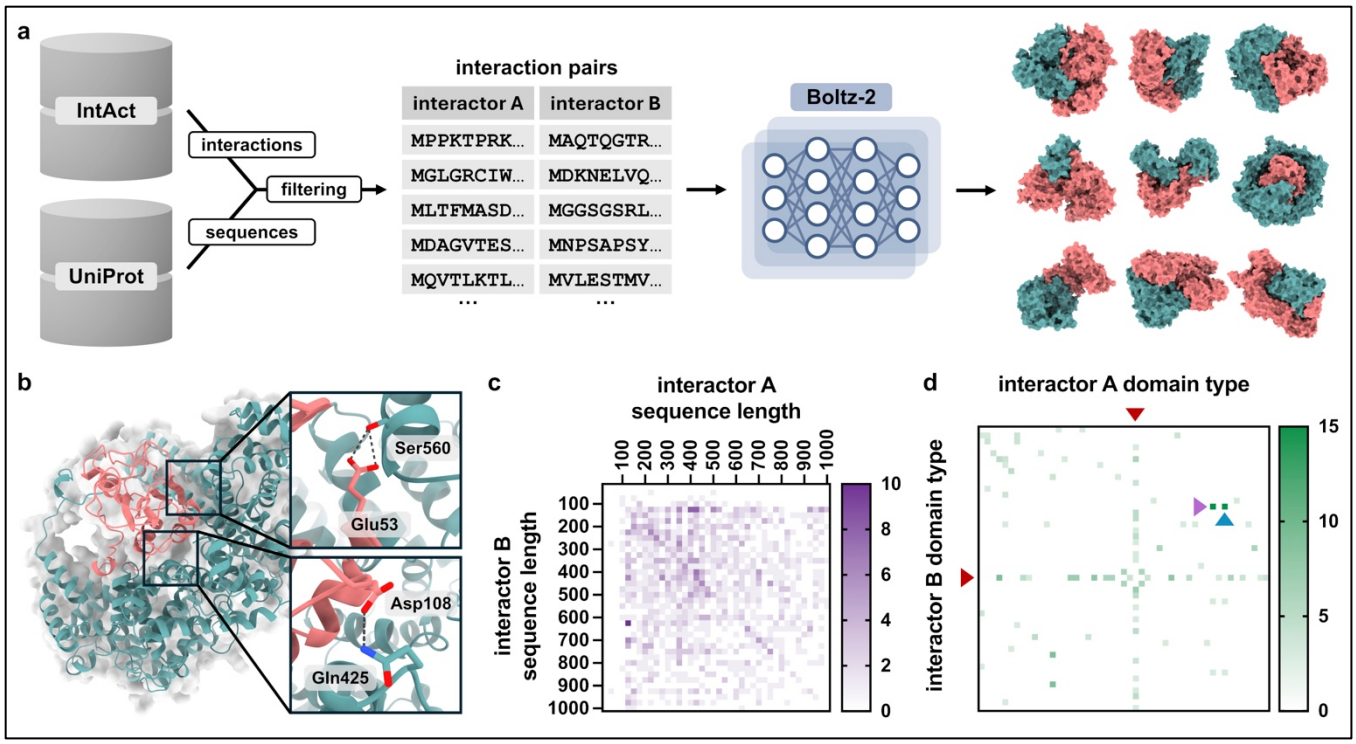

Figure 1. (a) Overview of protein-protein interaction structure prediction workflow, leveraging human biochemical interaction data and amino acid sequence data from the IntAct and UniProt databases, respectively, for binary complex prediction with Boltz-2. **(b)** Boltz-2 structure interaction prediction with the highest confidence score (0.938) among all interaction structures predicted in the current study. Interactor protein A is depicted in teal, interactor protein B is depicted in red, and dashed lines represent hydrogen bonds. The interaction is that of nuclear cap-binding complex protein subunit 1 and protein subunit 2. Insets show examples of predicted interacting residues between the two proteins. **(c)** Bivariate distribution of sequence lengths of protein interactor pairs, with an upper limit of 1,000 amino acid residues per protein. Data shown as increasing intervals of 25 amino acids, with the scale bar representing number of protein interactor pairs. **(d)** Co-occurrence matrix of Pfam domains across protein interactor pairs. Red arrowheads indicate the most abundant domain, PF02991 (autophagy protein Atg8 ubiquitin like), n = 215 co-occurrences. Purple arrowhead indicates co-occurrence of domains PF13923 (zinc finger, C3HC4 type (RING finger)) and PF00385 (chromatin organization modifier domain), n = 15 co-occurrences, while blue arrowhead indicates co-occurrence of domains PF16207 (RAWUL domain RING finger- and WD40-associated ubiquitin-like) and PF00385 (chromatin organization modifier domain), n = 15 co-occurrences. Data shown as bins with a minimum of three co-occurrences per bin, with the scale bar representing number of domain co-occurrences.

Prediction confidence and sequence-dependent effects

To gain insight into the quality of the predicted interaction structures, we used confidence metrics for assessment of overall structure as well as the interaction interface **(Supplementary Table 2)**. Predicted local distance difference test (pLDDT) [12, 25] and predicted template modelling (pTM) [12, 26] scores were used for evaluation of overall structure. In brief, pLDDT scores residue-level confidence while pTM

scores confidence in regional topology. Interaction interface-restricted versions of pLDDT and pTM were used for assessment of interaction-specific structural confidence [16, 17, 22]. The Boltz-2 confidence score, which combines overall pLDDT with interface pTM, was also included for evaluation [22] (see Methods). These metrics each exhibited a wide range of values across the dataset of n = 1,394 predicted interaction structures **(Figure 2a)**. The distribution of these values suggests greater residue-level confidence (as indicated by pLDDT metrics) and lower confidence in regional topology (as indicated by pTM metrics), while the combined confidence score had a median value of 0.583, indicating that more than half of the structures were predicted with moderate confidence.

We also considered whether sequence length and MSA depth might influence prediction confidence. The relationship between combined sequence length of interaction structures (defined here as the geometric mean of the two interactor proteins) and the overall confidence score indicates that structures with shorter sequences were predicted somewhat more confidently, though this negative correlation was relatively weak (r = -0.261) **(Figure 2b)**. Comparing MSA depth with the overall confidence score revealed a weak-to-moderate positive correlation (r = 0.353), suggesting that increased MSA depth may improve prediction confidence **(Figure 2c)**. Further analyses comparing MSA depth with overall **(Figure 2d)** and interface **(Figure 2e)** pLDDT and pTM metrics showed varying degrees of prediction confidence improvement with increased MSA depth, although relatively limited in correlation strength. The interaction network of protein complexes within the first (upper) quartile of the confidence metric **(Figure 1a)**, additionally limited to proteins involved in at least two binary interactions, is depicted in **Figure 3a**, along with examples of corresponding predicted structures **(Figure 3b & c)**.

We next performed a limited evaluation of prediction accuracy using experimentally resolved structures as a reference. Given that the training cutoff for Boltz-2 was 06/01/2023, we searched the PDB [9] for structures deposited after the cutoff date for entries consisting of binary protein complexes corresponding to those predicted. While only three such structures were identified, PDB IDs 9B4Y, 8X8A [27], and 8T1H [28], they provide a means for experimental comparison. These experimentally determined complex structures were compared to the corresponding interaction structures predicted by Boltz-2 and by AlphaFold3 **(Supplementary Figure 1)**. DockQ [29] scores and interface RMSD were computed between experimentally determined and predicted complex structures, with the predicted structures generated by both models ranging in scores from medium quality (DockQ of 0.49 – 0.80) **(Supplementary Figure 1a)**, acceptable quality (DockQ of 0.23 – 0.49) **(Supplementary Figure 1b)**, and incorrect (DockQ of 0 – 0.23) **(Supplementary Figure 1c)**. Taken together, these results highlight various strengths and limitations of large-scale protein interaction structure prediction with Boltz-2, including favorable residue-level prediction confidence contrasted by relative uncertainty in regional topology prediction. Furthermore, the availability of additional sequence context through increased MSA depth, though not found to be a requirement for high prediction confidence, may have a moderately positive effect.

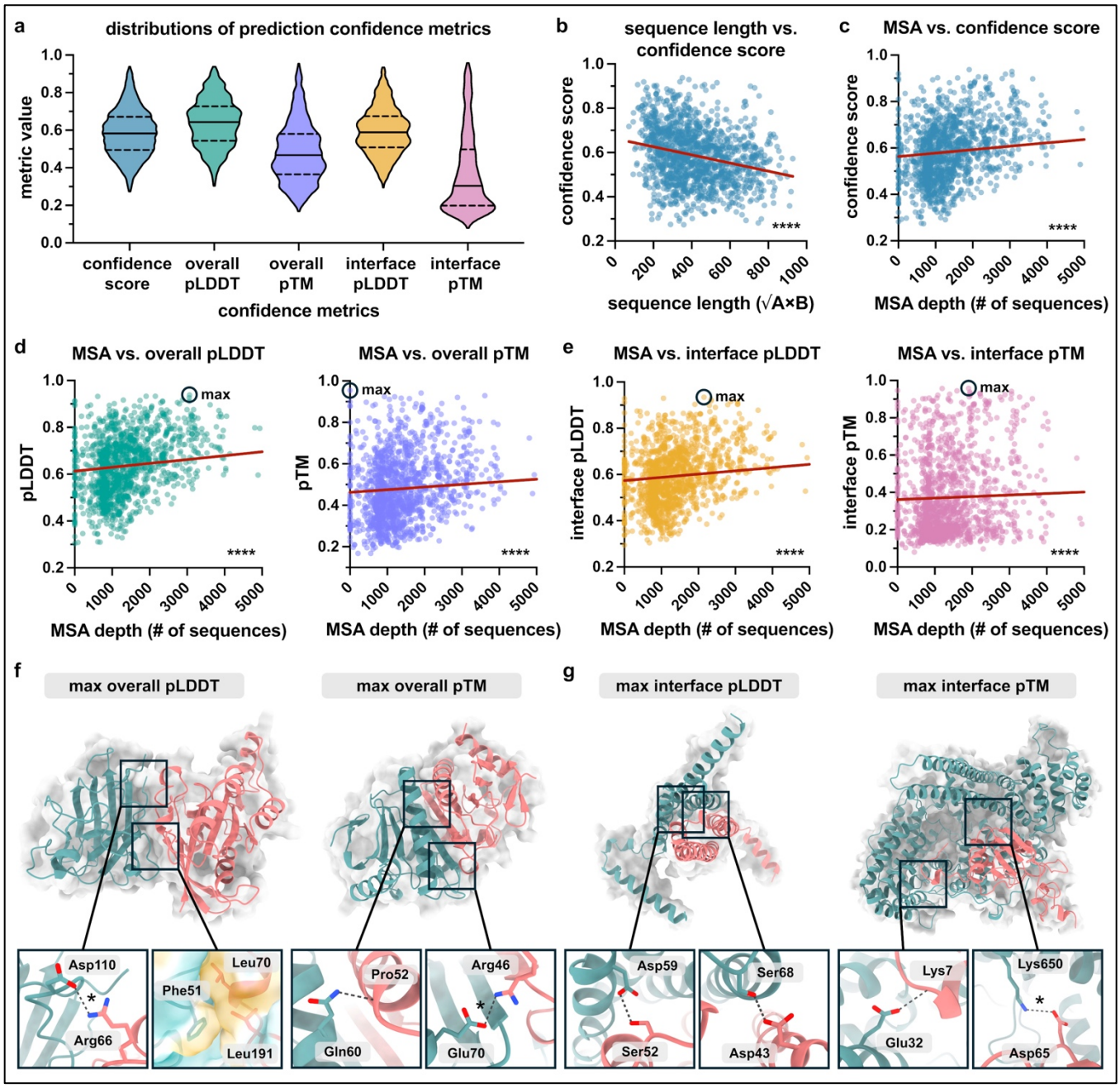

Figure 2. (a) Prediction confidence metric distributions across all protein interaction structures predicted (n = 1,394). All metrics were determined by Boltz-2 on a normalized scale of 0 to 1, enabling direct comparison (see Methods). Solid lines represent medians and dashed lines represent first and third quartiles. **(b)** Relationship between combined sequence length (the geometric mean of interactor sequence A and interactor sequence B) and confidence score of the predicted interaction structures. Spearman correlation r = -0.261, ****p <0.0001. **(c)** Relationship between MSA depth and confidence score of the predicted interaction structures. Spearman correlation r = 0.353, ****p <0.0001. Predicted interaction structures with MSA depth greater than 5000 (n = 25 structures) are not graphically depicted here, but were included in statistical analyses for panels c-e. **(d)** Relationship between MSA depth and entire interaction complex (overall) pLDDT and pTM. Spearman correlation r = 0.369 for overall pLDDT, Spearman correlation r = 0.266 for overall pTM, ****p <0.0001 **(e)** Relationship between MSA depth and interaction interface-aggregated pLDDT and pTM. Spearman

correlation r = 0.331 for interface pLDDT, Spearman correlation r = 0.139 for interface pTM, ****p <0.0001. Lines determined by linear regression are depicted in red in panels b-e. **(f-g)** Predicted interaction structures which exhibited the greatest overall (f) and interface (g) metrics. Interactor protein A is depicted in teal and interactor protein B is depicted in red. Dashed lines represent hydrogen bonds, dashed lines with an asterisk represent salt bridges, and yellow surfaces represent hydrophobic regions.

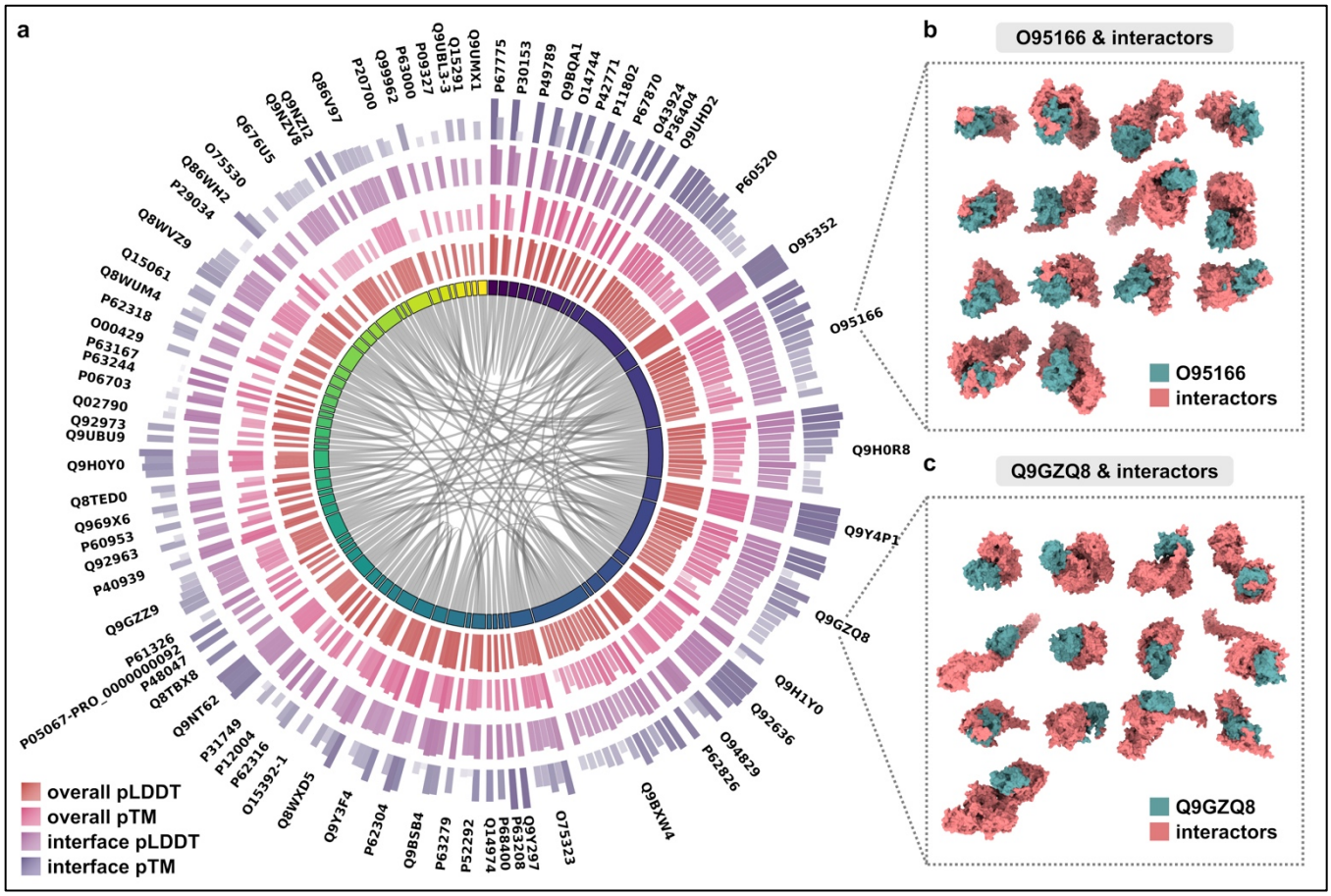

Figure 3. (a) Chord diagram depicting the interaction network between proteins with a minimum of two other interactors where the interaction confidence score is > 0.6709 (third quartile, **Figure 2a**). Individual proteins are plotted along the inner circumference of the circle, with differing colors for each protein. Grey connecting lines (chords) represent interactions between proteins. The four outer layers consisting of scalar bars denote (1) overall pLDDT, (2) overall pTM, (3) interface pLDDT, and (4) interface pTM. **(b-c)** Predicted interaction structures for the two proteins with the greatest number of interactors from (a). This includes n = 14 binary interactions for O95166 (b) and n = 13 binary interactions for Q9GZQ8 (c).

Interaction interface domains, biological function, and oncogenic involvement

The presence and co-occurrence of Pfam domains [23, 24] uncovered during our sequence-level analyses **(Figure 1d)** motivated us to explore the putative interaction involvement of these domains

based on the 3D structural complexes predicted with Boltz-2. In order to identify Pfam domains at the interaction interface, we employed a standard < 5 Å distance cutoff [30, 31] between any atom in a given domain present in one interactor protein and any atom in the entirety of the other interactor protein. Under this proximity criteria, a Pfam domain was considered to be putatively involved in a protein-protein interaction. Analysis of the *n* = 1,394 predicted interaction structures revealed that 679 structures contained Pfam domains within 5 Å of the interaction interface compared to 793 structures containing Pfam domains without any proximity restriction **(Supplementary Table 3)**. The prevalence of Pfam domains varied considerably **(Figure 4a, Supplementary Table 4)**, with the autophagy protein Atg8 ubiquitin like domain (PF02991) found in the greatest number of interaction structures within 5 Å of the interaction interface. Individual examples of prevalent domains from within the predicted interaction structures are shown in **Figure 4b**.

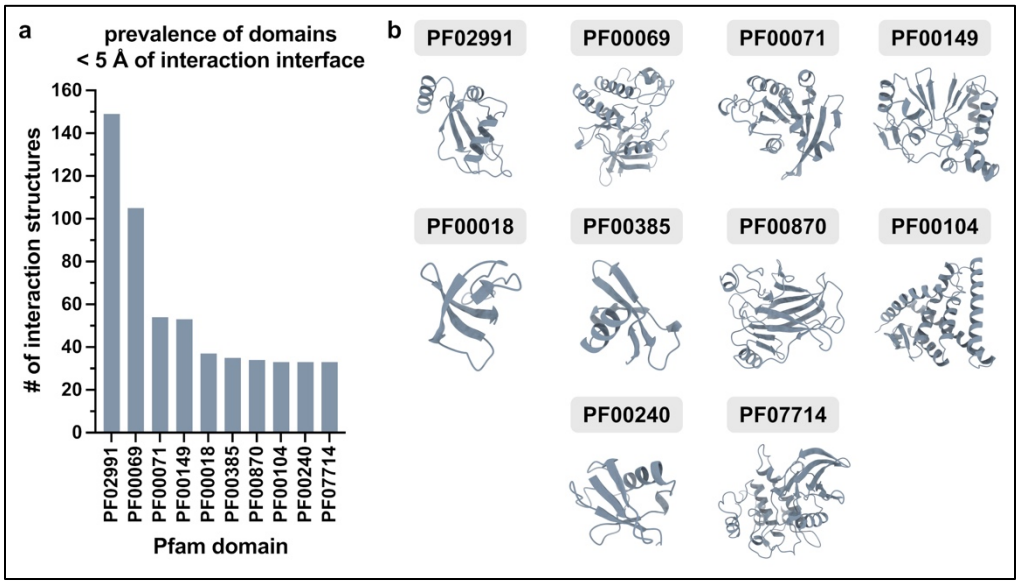

Figure 4. (a) Top 10 most prevalent Pfam domains by number of interaction pairs, restricted to only include domains found within 5 Å of the protein-protein interaction interface. These include: autophagy protein Atg8 ubiquitin like domain (PF02991), present in *n* = 149 interaction structures; protein kinase domain (PF00069), *n* = 105 interaction structures; Ras family domain (PF00071), *n* = 54 interaction structures; calcineurin-like phosphoesterase domain (PF00149), *n* = 53 interaction structures; SH3 domain (PF00018), *n* = 37 interaction structures; chromatin organization modifier domain (PF00385), *n* = 35 interaction structures; P53 DNA-binding domain (PF00870), *n* = 34 interaction structures; ligand-binding domain of nuclear hormone receptor (PF00104), *n* = 33 interaction structures; ubiquitin family domain (PF00240), *n* = 33 interaction structures; protein tyrosine and serine/threonine kinase domain (PF07714), *n* = 33 interaction structures. The quantification of number of interactor pairs containing a given domain is non-redundant, *i.e.* multiple occurrences of the same domain (either in interactor A or interactor B) do not cumulatively inflate the count. **(b)** Examples of the top 10 most prevalent domains from (a) as found within predicted interaction structures based on Pfam domain sequence annotation.

We also analyzed the co-occurrence of Pfam domains that were present within 5 Å of the interaction interface **(Figure 5a-d)**. The autophagy protein Atg8 ubiquitin like domain (PF02991) had the greatest level of co-occurrence with other domains, aligning with the high prevalence of this domain across interaction structures **(Figure 4a)**, while domains PF13923 with PF00385 and PF16207 with PF00385 exhibited the most frequent mutual co-occurrence. While the structural interaction interface proximity-restricted domain co-occurrences were less prevalent overall, the general pattern of co-occurrence resembled that observed from sequence-level analysis **(Figure 1d)**. Additionally, we sought to infer a broad contextualization of biological function based on the predicted structural interactions paired with interaction interface proximity-restricted domain annotation. As a means for high-level functional categorization, we utilized UniProt biological process keywords—a controlled vocabulary augmented with Gene Ontology (GO) annotation [32, 33] **(Supplementary Table 3)**. Six broad functional groupings of the domain-based interactions were compiled, including: apoptosis, cell division, differentiation, DNA repair, immunity, and protein transport. These groupings reveal intricate domain-based interaction networks across the predicted structural complexes **(Figure 5e)**. Furthermore, these groupings varied in Pfam domain prevalence and overlap **(Supplementary Figure 2)**, indicating potential structural involvement of specific domains within and across biological functions.

Another aspect we considered was domain-based pathological involvement of the predicted protein complexes. We focused on involvement in cancer, specifically oncogenic proteins. Filtering the predicted structures by the UniProt disease keyword proto-oncogene (KW-0656) revealed an interaction network consisting of n = 129 potentially oncogenic interactions **(Supplementary Table 5, Figure 6)**. Furthermore, the majority of proteins involved in these interactions have genetic variants, as annotated in the Single Nucleotide Polymorphism Database (dbSNP) [34]. This prediction-informed network sheds light on the structural basis of protein interactions which may be involved in cancer, as well genetic variants of possible relevance. Collectively, these analyses yield an extensive body of novel data which position protein-protein interaction profiles within functional and disease contexts on the basis of large-scale structural information. While this structural basis is predicted rather than experimental, and should be interpreted with appropriate caution, it provides an *in silico* modality for exploring putative structure-function relationships with vast coverage.

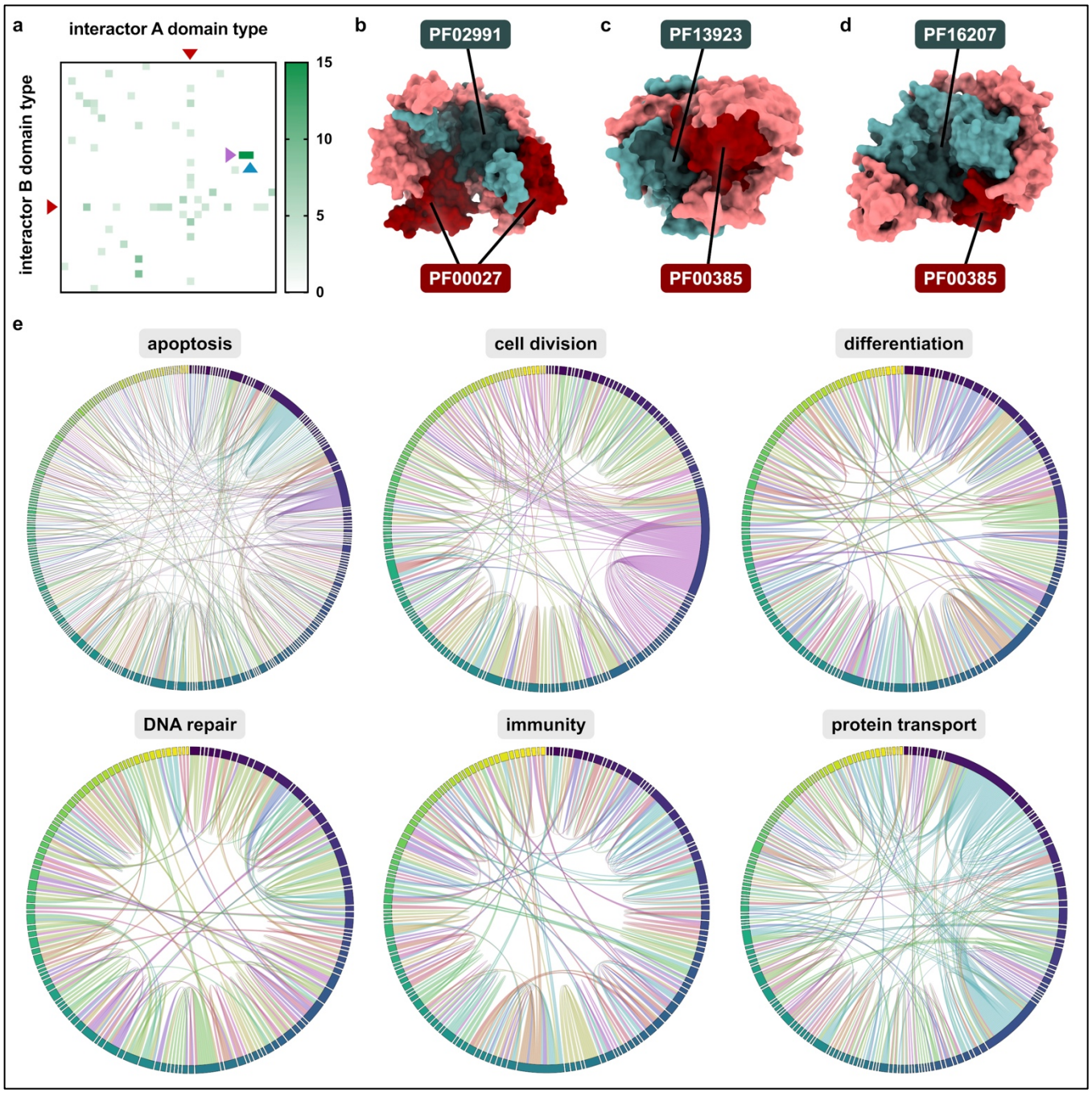

Figure 5. (a) Co-occurrence matrix of Pfam domains within 5 Å of the interaction interface between protein interactor pairs. Red arrowheads indicate the most abundant domain PF02991 (*n* = 122 interaction interface co-occurrences), purple arrowhead indicates co-occurrence of domains PF13923 and PF00385 (*n* = 15 interaction interface co-occurrences), and blue arrowhead indicates co-occurrence of domains PF16207 and PF00385 (*n* = 15 interaction interface co-occurrences). Data shown as bins with a minimum of three protein interactor pairs per bin, with the scale bar representing number of protein interactor pairs. **(b-d)** Examples of predicted interaction structures which include co-occurrences of domains PF02991 with PF00027 (b), PF13923 with PF00385 (c), and PF16207 with PF00385 (d). **(e)** Chord diagrams illustrating domain-based interaction networks grouped by UniProt biological process keyword annotation. For each diagram, individual proteins are plotted along the circumference of the circle and connecting lines (chords) represent interactions with other proteins involving domains < 5 Å of the interaction interface. Differing colors along the circle circumference

represent unique proteins while differing colors of the chords represent unique domains. UniProt biological process keywords used for functional grouping include KW-0053 (apoptosis), KW-0132 (cell division), KW-0221 (differentiation), KW-0234 (DNA repair), KW-0391 (immunity), KW-0653 (protein transport), with at least one protein in the interaction pair annotated with the designated keyword required for group inclusion.

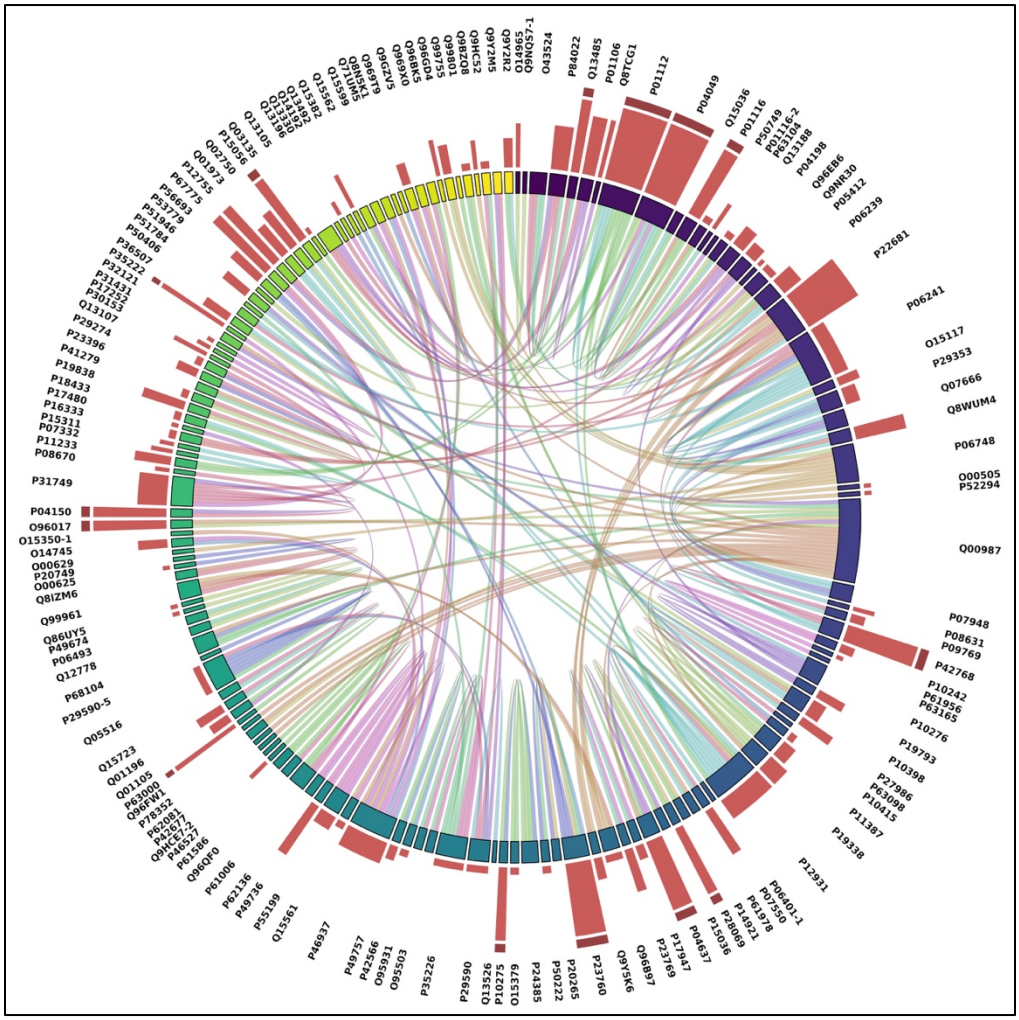

Figure 6. Chord diagram depicting an oncology-focused interaction network based on domains proximal to the interaction interface. As in **Figure 5b**, the circumference of the circle contains individual proteins (differing by color) with connecting lines (chords) representing interactions between proteins which involve domains within 5 Å of the interaction interface. Interaction pairs were included on the basis of at least one protein being annotated with the UniProt disease keyword proto-oncogene (KW-0656). The red outer layer consisting of scalar bars indicates the number of genetic variants documented for each protein ranging from 0 to 10 variants, with variant counts > 10 denoted by dark red bars.

DISCUSSION

This work demonstrates the utility of Boltz-2 for structural modeling of the human interactome through large-scale prediction of binary interaction structures based on biochemical data from IntAct in

combination with sequence data from UniProt. Although we predicted structures of a constrained representative subset of human PPIs, the scalability of this approach supports application to larger datasets. Our analyses revealed that over half of the structures were predicted with moderate confidence, with residue-level confidence appearing to outperform confidence in regional topology. Furthermore, we found that increased MSA depth was associated with greater prediction confidence, though only moderately. Another trend we observed, albeit weakly correlated, was decreased prediction confidence with increased sequence length, indicating a potential limitation for the prediction of larger protein complexes.

Notably, in the limited evaluation comparing predicted complex structures to those experimentally determined, both Boltz-2 and AlphaFold3 performed strongly and struggled on the same cases. This observation could indicate that certain structural classes or interface types remain inherently challenging to predict. Novel training strategies and architecture refinements may serve to improve predictive accuracy for currently challenging targets. Nevertheless, caution should be exercised when interpreting predicted complex structures generated with either computational tool, and confidence metrics may be useful indicators in this regard.

Our investigation of annotated domains and functional contextualization were also insightful. The finding that 679 of the 1,394 (48.7 %) predicted structural complexes contained Pfam domains within 5 Å of the interaction interface highlights the importance of domains in protein-protein interactions [35, 36]. The domain-based interactions inferred in the current work, although predictive, provide a broadened representation of interaction involvement given the substantially increased structural coverage afforded through structural prediction. Furthermore, the elucidation of domain-based interaction networks within functional groupings indicated interaction intricacy and variability in relation to biological function. Similarly, prediction-informed network analysis of oncogenic proteins revealed both complexity and putative involvement of genetic variants. These large-scale functional and disease-relevant characterizations are enabled by increased structural coverage, and may serve as a basis for establishing novel structure-function relationships as well as characterizing pathologic processes.

Looking ahead, it is reasonable to propose that Boltz-2 could be used to model multi-protein complexes of the human interactome, going beyond binary interactions. Boltz-2 is capable of performing predictions of structures with more than two chains, and such modeling has previously been explored for protein complexes with AlphaFold-Multimer [37]. Moreover, large-scale modeling of interactions with disease-relevant mutations as well as interactions with pathogenic proteins would be particularly informative for understanding mechanisms of disease and guide targeted therapeutic development. There are multiple additional avenues of modeling that merit further research. The current work suggests Boltz-2 is a valuable tool for human interactome protein structure prediction and related work, though potential limitations related to protein complex size and topological accuracy remain important considerations.

These AI/ML-based modeling approaches are anticipated to aid multiple areas of biomedical research through broadened characterization of structural interactions.

METHODS

Curation and processing of interaction and sequence data

The *Homo sapiens* species-specific dataset, which lists binary interaction pairs by UniProt ID, was downloaded from the IntAct database in tabular format [2]. The dataset was filtered as follows: (1) removal of all interaction pairs for which one of the two interacting proteins were non-human; (2) exclusive retention of interaction pairs which contained the “Direct interaction” designation; and (3) exclusive retention of interaction pairs which determined by pull down as the experimental interaction detection method. Corresponding sequences of interaction pairs were obtained by UniProt ID using the UniProt website REST API [8]. Corresponding Pfam domains were obtained using the InterPro Rest API [23]. Twelve interaction pairs were excluded from data analysis. Three interaction pairs were excluded due to their sequences containing the amino acid selenocysteine. Additionally, nine interaction pairs (P54278 with P40692; Q9UBS5 with P46459; Q8N8A2 with P35125-3; Q9UIF7 with P43246; Q13563 with Q13563; Q92622 with Q8NEB9; P11142 with P34932; Q04759 with Q8IVH8; and Q9UNQ0 with P22413) for which Boltz-2 failed to generate predictions despite multiple attempts, were also excluded. With all of the above curation and exclusion, a total of n = 1,394 binary interaction pairs was analyzed in the current work.

Interaction structure prediction and analyses

Structure predictions were performed using Boltz-2 [22] version 2.1.1 in the Google Colab environment. Hardware configurations were set to A100 GPU and High-RAM. The paired interaction protein sequences were used as input for protein complex prediction. Default Boltz-2 settings were used, except for MSA server use set to True, as follows:

diffusion_samples = 1
recycling_steps = 3
sampling_steps = 200
step_scale = 1.638
use_msa_server = True

msa_server_url = https://api.colabfold.com
max_msa_seqs = 8192
subsample_msa = False
msa_pairing_strategy = greedy

With MSA server use, Boltz-2 relies on the MMseqs2 server [38-40] for automatic MSA generation. Output from Boltz-2 used for analyses include predicted structures in CIF format, MSAs, and confidence metrics (confidence score, overall pLDDT, overall pTM, interface pLDDT, interface pTM), as described in the Boltz-2 documentation available on GitHub (https://github.com/jwohlwend/boltz). The Boltz-2 confidence score is calculated as follows: confidence score = 0.8 × overall pLDDT + 0.2 × interface pTM.

For comparison to experimental structures, the RCSB PDB Search API [9] was used to search for structures corresponding to the UniProt IDs of the n = 1,394 predicted structures that were deposited after the Boltz-2 model training cutoff date of 06/01/2023 [22]. Search criteria were set to include only structures determined by X-ray crystallography or CryoEM, without mutations, protein only, and containing exactly two chains (corresponding to the two interactor proteins), which retrieved a total of three structures **(Supplementary Figure 1)**. Corresponding AlphaFold3 structures were predicted using the Google DeepMind AlphaFold server (https://alphafoldserver.com/) [16]. DockQ scores and interface RMSD were determined using DockQ v2 [29]. All other molecular analyses, including hydrogen bond and salt bridge determination, hydrophobicity determination, and structure visualization, were performed using ChimeraX [41]. In ChimeraX, hydrogen bonds were determined with the H-bonds tool using default settings (*i.e.,* radius = 0.075 Å and without relaxed distance/angle criteria) and selection of salt bridge only setting allowed for salt bridge determination. Hydrophobic surfaces were determined using the automated molecular lipophilicity potential (MLP) command.

Pfam domain and interaction network analyses

Pfam domain annotations were curated based on interactor protein UniProt IDs using the InterPro REST API [23]. For a given interactor protein, this included retrieval of Pfam domain IDs and their corresponding start and end positions within the UniProt-obtained protein sequence. To identify Pfam domains less than 5 Å of the interaction interface, the Biopython Bio.PDB package [42] was used to map domain residues within the predicted structural complexes and compute minimum interatomic distances between each domain and the partner interactor protein chain. UniProt biological process keywords corresponding to each interactor protein UniProt ID were obtained using the UniProt website REST API [23]. Chord diagrams of interaction networks were generated using the pyCirclize Python package (https://github.com/moshi4/pyCirclize), with individual proteins plotted as sectors (nodes) along the

circumference of the circle and chords (connecting lines) representing interactions with other proteins. The six biological function networks **(Figure 5e)** were compiled by filtering the entire binary interaction dataset **(Supplementary Table 3)** by UniProt biological process keyword ID, specifically: (1) KW-0053 for apoptosis; (2) KW-0132 for cell division; (3) KW-0221 for differentiation; (4) KW-0234 for DNA repair; (5) KW-0391 for immunity; and (6) KW-0653 for protein transport. The oncology-focused network **(Figure 6)** was compiled by filtering the entire binary interaction dataset **(Supplementary Table 3)** by the UniProt disease keyword ID proto-oncogene (KW-0656). Interactions depicted in the chord diagrams in **Figure 5e** and **Figure 6** are strictly based exclusively on domains < 5 Å of the interaction interface. Genetic variant dbSNP [34] annotations were retrieved using the UniProt website REST API [23].

Statistical Analyses and Graphical Representation

Statistical analyses of medians, quartiles, Spearman correlations, linear regression, as well as graphical representations shown in Figures 1 and 2, were performed using GraphPad Prism 10. In Figure 2c-e, n = 25 interaction structures which had MSA depths greater than 5000 sequences are not shown on the graphs but were not excluded from statistical analyses. The complete set of data used for statistical analyses is presented in **Supplementary Table 1**.

Data Availability

All predicted complex structure models described herein are freely available online at https://github.com/structural-interactome/human-interactome.

ACKNOWLEDGMENTS

This work was supported by the Levy-Longenbaugh Donor-Advised Fund (PIs: R.P. and W.A.). RCSB Protein Data Bank core operations are jointly funded by the National Science Foundation (DBI-2321666, PI: S.K.B.), the US Department of Energy (DE-SC0019749, PI: S.K.B.), and the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of General Medical Sciences of the National Institutes of Health (R01GM157729, PI: S.K.B.).

We are grateful to Saro Passaro, Gabriele Corso, Jeremy Wohlwend et al. [22] from the Massachusetts Institute of Technology (MIT) for the development and open-source availability of Boltz-2, which was central to the current work. We would also like to acknowledge the European Bioinformatics Institute (EMBL-EBI) Molecular Networks team for open access to the IntAct database [2] and the UniProt

Consortium [8]. Molecular analyses and visualization were performed using UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from the National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

AUTHOR CONTRIBUTIONS

A.M.I., C.M., S.K.B., R.P., and W.A. conceptualization; A.M.I. and C.M. methodology; A.M.I. investigation; A.M.I., C.M., S.K.B., R.P., and W.A. formal analysis; A.M.I. and C.M. writing—original draft; S.K.B., R.P., and W.A. writing—review & editing; S.K.B., R.P., and W.A. funding acquisition; S.K.B., R.P. and W.A. overall project supervision.

COMPETING INTERESTS

A.M.I. is a founder and partner of North Horizon, which is engaged in the development of artificial intelligence-based software. R.P. and W.A. are founders and equity shareholders of PhageNova Bio. R.P. is Chief Scientific Officer and a paid consultant of PhageNova Bio. R.P. and W.A are founders and equity shareholders of MBrace Therapeutics. R.P. and W.A. serve as paid consultants for MBrace Therapeutics. R.P. and W.A. have Sponsored Research Agreements (SRAs) in place with PhageNova Bio, MBrace Therapeutics, and Alnylam Pharmaceuticals; this study falls outside of the scope of these SRAs. These arrangements are managed in accordance with the established institutional conflict-of-interest policies of Rutgers, The State University of New Jersey. C.M. and S.K.B. declare no competing interests.

REFERENCES

1. Greenblatt, J.F., B.M. Alberts, and N.J. Krogan, Discovery and significance of protein-protein interactions in health and disease. *Cell*, 2024. **187**(23): p. 6501-6517. doi:10.1016/j.cell.2024.10.038
2. Del Toro, N., et al., The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res*, 2022. **50**(D1): p. D648-D653. doi:10.1093/nar/gkab1006
3. Giurgiu, M., et al., CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res*, 2019. **47**(D1): p. D559-D563. doi:10.1093/nar/gky973
4. Luck, K., et al., A reference map of the human binary protein interactome. *Nature*, 2020. **580**(7803): p. 402-408. doi:10.1038/s41586-020-2188-x
5. Drew, K., J.B. Wallingford, and E.M. Marcotte, hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol Syst Biol*, 2021. **17**(5): p. e10016. doi:10.15252/msb.202010016
6. Oughtred, R., et al., The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*, 2021. **30**(1): p. 187-200. doi:10.1002/pro.3978
7. Szklarczyk, D., et al., The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res*, 2025. **53**(D1): p. D730-D737. doi:10.1093/nar/gkae1113
8. The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res*, 2024. doi:10.1093/nar/gkae1010
9. wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*, 2019. **47**(D1): p. D520-D528. doi:10.1093/nar/gky949
10. Burley, S.K. and H.M. Berman, Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. *Structure*, 2021. **29**(6): p. 515-520. doi:10.1016/j.str.2021.04.010
11. Arita, M., I. Karsch-Mizrachi, and G. Cochrane, The international nucleotide sequence database collaboration. *Nucleic Acids Res*, 2021. **49**(D1): p. D121-D124. doi:10.1093/nar/gkaa967
12. Jumper, J., et al., Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021. **596**(7873): p. 583-589. doi:10.1038/s41586-021-03819-2
13. Lin, Z., et al., Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023. **379**(6637): p. 1123-1130. doi:10.1126/science.ade2574
14. Baek, M., et al., Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021. **373**(6557): p. 871-876. doi:10.1126/science.abj8754
15. Burley, S.K., W. Arap, and R. Pasqualini, Predicting Proteome-Scale Protein Structure with Artificial Intelligence. *N Engl J Med*, 2021. **385**(23): p. 2191-2194. doi:10.1056/NEJMcibr2113027
16. Abramson, J., et al., Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024. **630**(8016): p. 493-500. doi:10.1038/s41586-024-07487-w
17. Evans, R., et al., Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2022. doi:10.1101/2021.10.04.463034
18. Anfinsen, C.B., Principles that govern the folding of protein chains. *Science*, 1973. **181**(4096): p. 223-30. doi:10.1126/science.181.4096.223

19. Ille, A.M., et al., From sequence to protein structure and conformational dynamics with artificial intelligence/machine learning. *Struct Dyn*, 2025. **12**(3): p. 030902. doi:10.1063/4.0000765

20. Kryshtafovych, A., et al., Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins*, 2021. **89**(12): p. 1607-1617. doi:10.1002/prot.26237

21. Kryshtafovych, A., et al., Critical assessment of methods of protein structure prediction (CASP)-Round XV. *Proteins*, 2023. **91**(12): p. 1539-1549. doi:10.1002/prot.26617

22. Passaro, S., et al., Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv*, 2025: p. 2025.06.14.659707. doi:10.1101/2025.06.14.659707

23. Blum, M., et al., InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res*, 2025. **53**(D1): p. D444-D456. doi:10.1093/nar/gkae1082

24. Paysan-Lafosse, T., et al., The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res*, 2025. **53**(D1): p. D523-D534. doi:10.1093/nar/gkae997

25. Mariani, V., et al., lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 2013. **29**(21): p. 2722-8. doi:10.1093/bioinformatics/btt473

26. Zhang, Y. and J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins*, 2004. **57**(4): p. 702-10. doi:10.1002/prot.20264

27. Zhang, Y., et al., Decoding the molecular mechanism of selective autophagy of glycogen mediated by autophagy receptor STBD1. *Proc Natl Acad Sci U S A*, 2024. **121**(37): p. e2402817121. doi:10.1073/pnas.2402817121

28. Rochon, K., et al., Structural basis for regulated assembly of the mitochondrial fission GTPase Drp1. *Nat Commun*, 2024. **15**(1): p. 1328. doi:10.1038/s41467-024-45524-4

29. Mirabello, C. and B. Wallner, DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 2024. **40**(10). doi:10.1093/bioinformatics/btae586

30. Salamanca Viloria, J., et al., An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass. *Sci Rep*, 2017. **7**(1): p. 2838. doi:10.1038/s41598-017-01498-6

31. Elez, K., A. Bonvin, and A. Vangone, Distinguishing crystallographic from biological interfaces in protein complexes: role of intermolecular contacts and energetics for classification. *BMC Bioinformatics*, 2018. **19**(Suppl 15): p. 438. doi:10.1186/s12859-018-2414-9

32. Bairoch, A., et al., The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 2005. **33**(Database issue): p. D154-9. doi:10.1093/nar/gki070

33. Gene Ontology, C., et al., The Gene Ontology knowledgebase in 2023. *Genetics*, 2023. **224**(1). doi:10.1093/genetics/iyad031

34. Phan, L., et al., The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Res*, 2025. **53**(D1): p. D925-D931. doi:10.1093/nar/gkae977

35. Lee, H., et al., Domain-mediated interactions for protein subfamily identification. *Sci Rep*, 2020. **10**(1): p. 264. doi:10.1038/s41598-019-57187-z

36. Beltran, A., et al., Site-saturation mutagenesis of 500 human protein domains. *Nature*, 2025. **637**(8047): p. 885-894. doi:10.1038/s41586-024-08370-4

37. Burke, D.F., et al., Towards a structurally resolved human protein interaction network. *Nat Struct Mol Biol*, 2023. **30**(2): p. 216-225. doi:10.1038/s41594-022-00910-8

38. Steinegger, M. and J. Soding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 2017. **35**(11): p. 1026-1028. doi:10.1038/nbt.3988

39. Mirdita, M., et al., ColabFold: making protein folding accessible to all. *Nat Methods*, 2022. **19**(6): p. 679-682. doi:10.1038/s41592-022-01488-1

40. Kim, G., et al., Easy and accurate protein structure prediction using ColabFold. *Nat Protoc*, 2024. doi:10.1038/s41596-024-01060-5

41. Meng, E.C., et al., UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci*, 2023. **32**(11): p. e4792. doi:10.1002/pro.4792

42. Cock, P.J., et al., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 2009. **25**(11): p. 1422-3. doi:10.1093/bioinformatics/btp163