A predicted structural interactome reveals binding interference from intrinsically disordered regions

Junhui Peng[1], Li Zhao[1]

[1] Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA
*Correspondence to: jpeng@rockefeller.edu; lzhao@rockefeller.edu

Abstract

Proteins function through dynamic interactions with other proteins in cells, forming complex networks fundamental to cellular processes. While high-resolution and high-throughput methods have significantly advanced our understanding of how proteins interact with each other, the molecular details of many important protein-protein interactions are still poorly characterized, especially in non-mammalian species, including *Drosophila*. Recent advancements in deep learning techniques have enabled the prediction of molecular details in various cellular pathways at the network level. In this study, we used AlphaFold2 multimer to examine and predict protein-protein interactions from both physical and functional datasets in *Drosophila*. We found that functional associations contribute significantly to high-confidence predictions. Through detailed structural analysis, we also found the importance of intrinsically disordered regions in the predicted high-confidence interactions. Our study highlights the importance of disordered regions in protein-protein interactions and demonstrates the importance of incorporating functional interactions in predicting physical interactions between proteins. We further compiled an interactive web interface to present the predictions, facilitating functional exploration, comparative analysis, and the generation of mechanistic hypotheses for future studies.

Introduction

Cells are dynamic environments where thousands of proteins interact simultaneously at any given time, and these protein-protein interactions (PPIs) are essential for various biological functions. PPIs can range from stable, permanent associations – such as those within stable protein complexes and molecular machinery – to transient, short-lived interactions with a wide range of affinities and lifetimes, as seen in signal transduction pathways (1–4). Understanding the organization and dynamics of the cellular PPI network is crucial for a comprehensive picture of cellular functions. Over the past decades, diverse high-throughput methods have been developed and used to study these interactions in different organisms, presenting a systematic overview of how proteins work together in cells (5). For instance, affinity purification mass spectrometry (6) has been instrumental in identifying protein complexes. Yeast two-hybrid systems (7) and other in vitro techniques such as cross-linking mass spectrometry (8) have also been used to map PPIs at high throughputs. More recently, in vivo proximity labeling methods such as BioID (9) and TurboID (10) have emerged as powerful tools for capturing dynamic protein interactions within living cells.

Drosophila is an important model species for biological and biomedical research. Despite significant efforts to study *Drosophila* protein-protein interactions both in vivo and in vitro, many uncertainties remain in existing PPI databases, such as false positives (11), inconsistency between databases (12), etc. The combination of complex tissues and small body sizes poses unique challenges in biochemistry compared to other model organisms such as rodents, particularly due to the difficulty in obtaining sufficient amounts of protein samples from specific tissues for high-quality experiments such as affinity purification mass spectrometry. Additionally, methods like yeast two-hybrid may introduce errors due to system differences (13). Therefore, it is essential to evaluate these databases to gain a deeper understanding of the reliability and biases of previous methods (11, 14). Furthermore, while much is known about the interactions of ordered structural components, characterizing the interaction details involving disordered proteins or disordered regions still remains challenging (1, 15–21).

AlphaFold2 Multimer (22, 23) has provided a new tool for evaluating direct and indirect protein interactions, potentially distinguishing true or stable interactions from experimental artifacts (18, 24–26), enhancing our understanding of the structure and function of interaction networks. From an evolutionary biological perspective, AlphaFold2 Multimer is also a powerful tool for studying the interactions and changes in interactions of homologous proteins across different lineages (27).

In this study, we combine AlphaFold2 multimer and the STRING database (14) to examine and predict the structural details of protein-protein interactions in *D. melanogaster*. We aim to systematically and comprehensively evaluate different sources of protein-protein interactions within the *Drosophila* genome, including physical interactions inferred from experiments or homology and functional associations derived from co-expression, co-existence in the same pathways, text-mining, genomic context, etc. Our study provides a resource for the structural details of possible physical interactions in the *Drosophila* genome, and highlights the importance of incorporating functional PPIs in future high-throughput computational studies. We also reveal the important and prevalence roles of intrinsically disordered regions in the *Drosophila* PPI network. We further compiled our analysis and results into an interactive web interface for users to explore, validate, and expand protein interaction hypotheses.

Material and Methods

Extracting representative PPIs from the STRING database

We first collected all PPIs with supporting evidence scores greater than 900 from the STRING database, version 11.5 (14). We downloaded and extracted 107,946 PPIs with the highest STRING confidence scores (≥0.9). We used MMseqs2 (28) to cluster the proteins involved in these PPIs at 50% sequence identity and extracted PPIs from the representative proteins to further removed redundant protein sequences with high sequence similarities. This reduced the number of PPIs to 34,940 for structural predictions. Functional enrichment analysis was performed in the STRING database.

Infer disordered, coil, and conditionally folded regions

We obtained the monomer structures of *D. melanogaster* proteome from the AlphaFold structural database, last accessed March 14, 2024 (29). We used AlphaFold-disorder (30) to compute the intrinsic structural disorder (ISD) of protein monomers. In the monomers, we define a region to be disordered if consecutive five or more residues were predicted with ISD<0.5. Regions with short structured segments (<3 consecutive structured residues) embedded within disordered stretches were also considered disordered. Disordered regions with 30 or more residues were considered as long intrinsically disordered regions (long IDRs) (31), and those between 5 to 30 residues were short disordered regions (short IDRs).

We used DSSP (32) version 4.4 to obtain the secondary structures of the protein monomers. We further grouped the eight-code secondary structures defined in DSSP into three states (33) by converting H (α-helix), G (3_{10}-helix), and I (π-helix) to H (helix), E (β-strand) and B (bridge) to E (strand), and others to C (coil). We also extracted coil regions from the monomer structures where consecutive 5 or more residues were predicted as coils (C) in AlphaFold structural database. Similarly, regions with short helix or strand segments (fewer than 3 residues) embedded within coil stretches were also considered part of the coil regions. We further define conditionally folded (CF) regions as the regions that are predicted with both high PLDDT (>70) and high structural disorder (ISD > 0.5) in their monomer states (30).

Structure prediction and structural analysis of D. melanogaster PPIs

We used the full version of AlphaFold2 multimer (version 2.3.0) to predict the structures of the 34,940 *D. melanogaster* PPIs. Finally, we were able to obtain 27,711 PPI predictions. To obtain the interaction interfaces of each predicted PPI, we defined two residues – each from one protein in the predicted PPI structure – to be interacting if any pair of their non-hydrogen atoms were within 7 Å of each other. We excluded noisy, suspicious interactions when the predicted aligned error (PAE) between the two residues was greater than 15 Å, as large PAE values indicate uncertainties in their relative orientations. We further grouped the interacting residues into different interface regions by merging neighboring residues. Interface regions that are separated by fewer than 3 residues were considered contiguous and merged.

We again used DSSP (32) version 4.4 to obtain the secondary structures in the predicted PPI structures as described above. To detect a coil-to-order transitions at coil interfaces, we compared the secondary-structure content of each interface in the monomer versus complex state using one-sided fisher-exact tests. We defined coil-to-order binding if the content of secondary structures (helix, H and strand, E) increases significantly at a p-value cutoff of 0.05. Interfaces that reside or overlap with conditional folded (CF) regions were classified as conditional folded binding, or CF binding.

Network analysis and visualization

We constructed, analyzed, and visualized the PPI network using the Python package Networkx (34). We performed Markov Clustering (MCL) analysis using a python version of MCL, accessible at https://github.com/GuyAllard/markov_clustering, commit 28787cf. In our interactive web platform (https://jhpanda.github.io/DmelPPI_web/), we used Cytoscape API (35) to show the position of each protein pair in the predicted high confidence PPI network (pDockQ ≥ 0.5).

Sequence conservation analysis

We compiled a list of 15 genomes from *D. melanogaster* to *H. salinarum* (Table S1). We used reciprocal best hit from blastp (36) to determine the orthologs of *Drosophila* proteins. We used the MAFFT-LINSI (37) version 7.429 to align the orthologous sequences. Sequence alignments were visualized by Mview (38).

Structural comparison and visualization

We used FoldSeek-Multimer (39) with default parameters to search for similar PPI structures in Protein Data Bank (40). All structures were visualized by UCSF Chimera (41).

Results

Predicted high confident D. melanogaster physical protein-protein interaction network

We downloaded all the protein-protein interactions (PPIs) with the highest STRING confidence scores (≥0.9) from the STRING database, version 11.5 (14), which include both direct physical interactions and indirect functional associations. The STRING database provides five main, non-exclusive categories of supporting evidence for the PPIs, including experimental, co-expression, curated databases, genomic context, and text-mining of published literatures. On top of these five categories, STRING database has also provided a physical subnetwork by incorporating evidence from the curated database organized in IMEx (42) and BioGRID database (43). Since it has been indicated that functionally associated proteins, such as co-expressed proteins, have higher chances to interact with each other than random protein pairs (44–46), we did not restrict our analysis to the physical subnetwork, as our goal is to systematically predict and examine all possible protein-protein interactions.

We used AlphaFold2 multimer to predict the structures of 27,711 protein pairs from the STRING database with highest STRING confidence scores (≥ 0.9), involving approximately 5000 protein-coding genes clustered at 50% protein-sequence identity (Methods). The 27,711 protein pairs can be divided into five mutually exclusive categories according to the STRING confidence scores in the five different supporting evidence categories, including 7838 in experimental, 7088 in co-expression, 6764 in curated databases, 5995 in text-mining, and 6 in genomic context (Figure S1). We used pDockQ to assess AlphaFold2 multimer prediction confidence (24, 47). Among the 27,711 protein pairs, 8,101 pairs (29.2%) were predicted to have direct physical interactions at acceptable confidence with pDockQ ≥ 0.23, and 3,621 pairs (13.1%) at high confidence with pDockQ ≥ 0.5. A

summary of the predictions can be found in Supplementary Data 1. This suggests that a relatively small percentage of the previously reported functional associations may have direct physical interactions.

We further examined whether the 3,621 predicted high-confidence PPIs (pDockQ ≥ 0.5) are supported by the STRING database curated physical subnetwork. Among these predicted high-confidence PPIs, 52% (1867 pairs) are supported by highest STRING confidence scores for physical interaction (≥ 0.9, Figure S2), ~70% (2541 pairs, which include the 1867 pairs with scores ≥ 0.9) by high STRING confidence scores (≥ 0.7, Figure S2), and ~81% (2929, which include the 2541 pairs with scores ≥ 0.7) by median STRING confidence scores (≥ 0.4, Figure S2). When further categorizing the 3,621 predicted high-confidence PPIs into five STRING database categories according to their highest STRING confidence scores, we found that 31.6% (1144) are best supported by experimental data, 13.7% (496) by co-expression, 19.7% (715) by curated database, 34.9% (1263) by text-mining, and 0.1% (3) by genomic context, specifically gene fusions. These results suggest the robustness of our AlphaFold2 Multimer predictions, with ~81% of the predicted high-confidence PPIs supported by reasonable physical interaction evidence in the STRING database. Interestingly, many of the remaining ~19% predicted interactions that lack physical evidence in STRING are supported by indirect functional associations, such as previously found co-expression and co-occurrence in cellular pathways, highlighting the importance of incorporating functional associations in predicting PPI structures.

AlphaFold2 multimer predicts core physical interaction networks

We combined network clustering and gene ontology (GO) enrichment analysis to investigate the overall picture of the highly connected network formed by the 3,621 high-confident physical PPIs (Figure 1). Overall, the 3,621 predicted high-confidence PPIs involved 2,740 proteins, with an average of 2.6 interactions per protein. We found that the PPI network was composed of one large connected subnetwork consisting of 2,630 PPIs, 11 relatively small connected subnetworks consisting of 10 to 43 PPIs, 128 very small connected subnetworks consisting of 3 to 9 PPIs, and another 228 singleton PPIs (Supplementary Data 2). The largest connected subnetwork with 2,630 PPIs involves 1,534 proteins, while the second largest subnetwork with 43 PPIs includes 32 proteins (Figure 1B). The other subnetworks are relatively small, each containing fewer than 30 proteins. The 228 isolated PPIs correspond to 456 proteins. Functional enrichment analysis of the largest connected subnetwork showed highly significant enrichment in biological processes such as transcription, translation, gene expression, and metabolism of macromolecules (Table S2), highlighting the interconnected nature of these fundamental cellular processes. In the second largest connected subnetwork, PPIs were mostly enriched in reproduction (Table S3), and the third largest was mostly enriched in signaling-related biological processes (Table S4).

We further applied Markov Clustering Algorithm to identify clusters of interacting proteins that may represent core subunits of protein complexes or core components in signaling pathways. The results showed that the top clusters represent components of major cellular complexes, such as ribosomes, the SNARE complex, the transcription initiation complex, spliceosome, the mitochondrial respiratory complexes, the mediator complex, the

ubiquitin ligase complex, the respiratory chain complex, the proteasome complex, and part of major signaling pathways, such as the CDK-dependent transcription regulation, SUMO pathway, GPCR signaling pathway, TORC2 signaling pathway, etc (Figure S1, Supplementary Data 3). This finding aligns with the results of the above GO functional enrichment analysis.

To further illustrate the predicted PPI structures, we showcase some examples where the interactions were supported by various evidence types in the STRING database (Figure 2). For instance, the CG2263/CG5706 and Fnta/Fntb were physical interactions supported by high-throughput affinity purification mass spectrometry in *Drosophila* (48). Two other examples, AP-1-2β/AP-1μ and Ost48/OstΔ were primarily functional associations derived from text-mining either in *D. melanogaster* (AP-1-2β/AP-1μ) or transferred by homology from other organisms (Ost48/OstΔ). In all cases, the predicted PPI structures closely match the homologous complexes in other organisms. Notably, in the last example of Ost48/OstΔ – where only exploratory physical interaction evidence is available in *D. melanogaster* – STRING database version 11.5 assigned a high confidence physical interaction score based on homology transfer and AlphaFold Multimer successfully predicted their interactions. This highlights that even in the absence of direct interaction evidence in *D. melanogaster*, high STRING confidence scores can be informative.

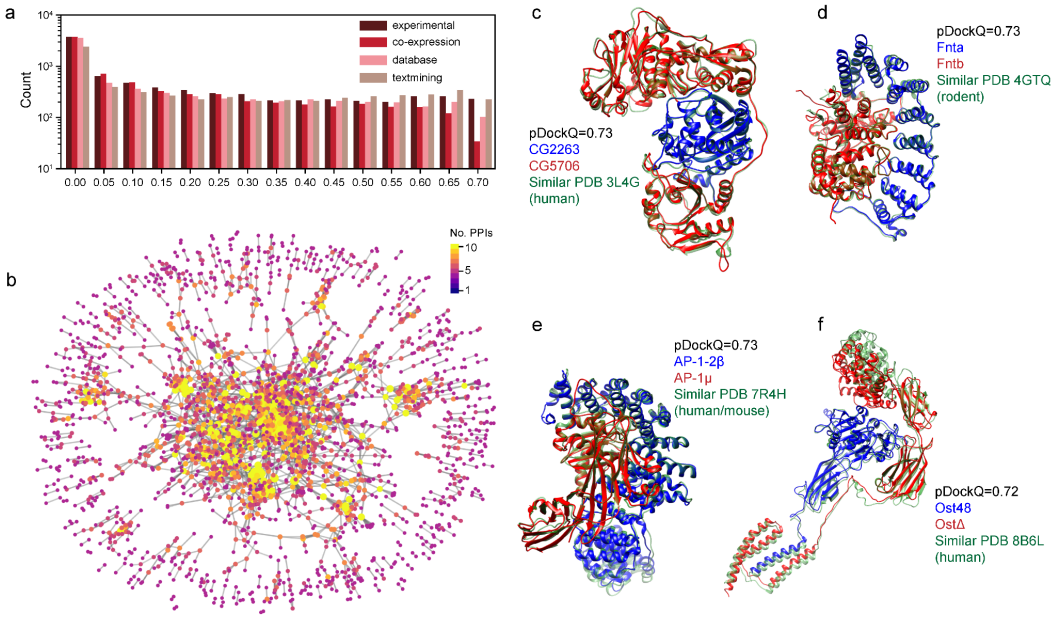

Figure 1. Overall pictures of AlphaFold2 multimer predictions of *D. melanogaster* PPIs in the STRING database. a) The number of predicted high-confidence PPIs supported by the highest evidence scores in the STRING database. The other 6 PPIs inferred from genomic context were not shown because of small sample sizes. b) The PPI network of predicted high-confidence PPIs by AlphaFold2 multimer. c), d), e), and f) Predicted PPI structures of CG2263/CG5706, Fnta/Fntb, AP-1-2β/AP-1μ, and Ost48/OstΔ, respectively. The predicted PPIs were shown in blue and red, each representing one protein. Closely matched PDB structures were shown in green. The source organisms of the PDB structures were also labeled.

Predicted high-confidence interactions with limited or no support in the STRING physical subnetwork

To gain an understanding of how indirect, functional associations – such as co-expression – contribute to the predicted high-confidence PPI network, we examined the subset protein pairs with limited or no support in the STRING physical subnetwork (STRING confidence score < 0.4). The analysis identified 692 predicted high-confidence PPIs, involving 918 proteins (Figure 2A, Supplementary Data 4). Among them, we observed a large connected subnetwork consisting of 136 proteins and 105 PPIs as well as three median sized connected subnetworks with 10 to 50 PPIs each (Figure 2A and 2B). The remaining PPIs formed 82 small connected subnetworks with 3 to 9 PPIs and 176 singleton PPIs.

To examine the biological relevance of these predicted PPIs, we manually examined several interactions. We identified potential PPIs that were either characterized after the release of the STRING database version 11.5 (14) or suggested by genetic experiments but lacked evidence of physical interaction. For example, Ran-Nup50 and Ptp61F-Stat92E were both scored as high-confidence functional associations but low-confidence physical interactions in the STRING database. However, recent high-resolution structures of the interactions between the orthologs of Ran-Nup50 in humans suggest that they directly interact (49). Our AlphaFold multimer analysis supports a high confidence interaction of Ran-Nup50 and Ptp61F-Stat92E (Figure 2C-2E). Notably, the structure of Ran-Nup50 (PDB code 7MO0) was released after the AlphaFold multimer training cutoff date (September 30, 2021) and thus was not included in model training (22). In the other example, Stat92E is a transcription factor in the JAK/STAT signalling pathway and Ptp61F is a protein tyrosine phosphatases that regulates the pathway. There is no direct evidence that Ptp61F directly interacts with Stat92E in *Drosophila*. However, indirect evidence showed that knockout of Ptp61F in *Drosophila* increases the level of phosphorylated Stat92E (50, 51), suggesting that Ptp61F potentially interacts with phosphorylated Stat92E to mediate its dephosphorylation. Overall, high-confidence PPI predictions will provide valuable hypotheses for future functional and mechanistic studies.

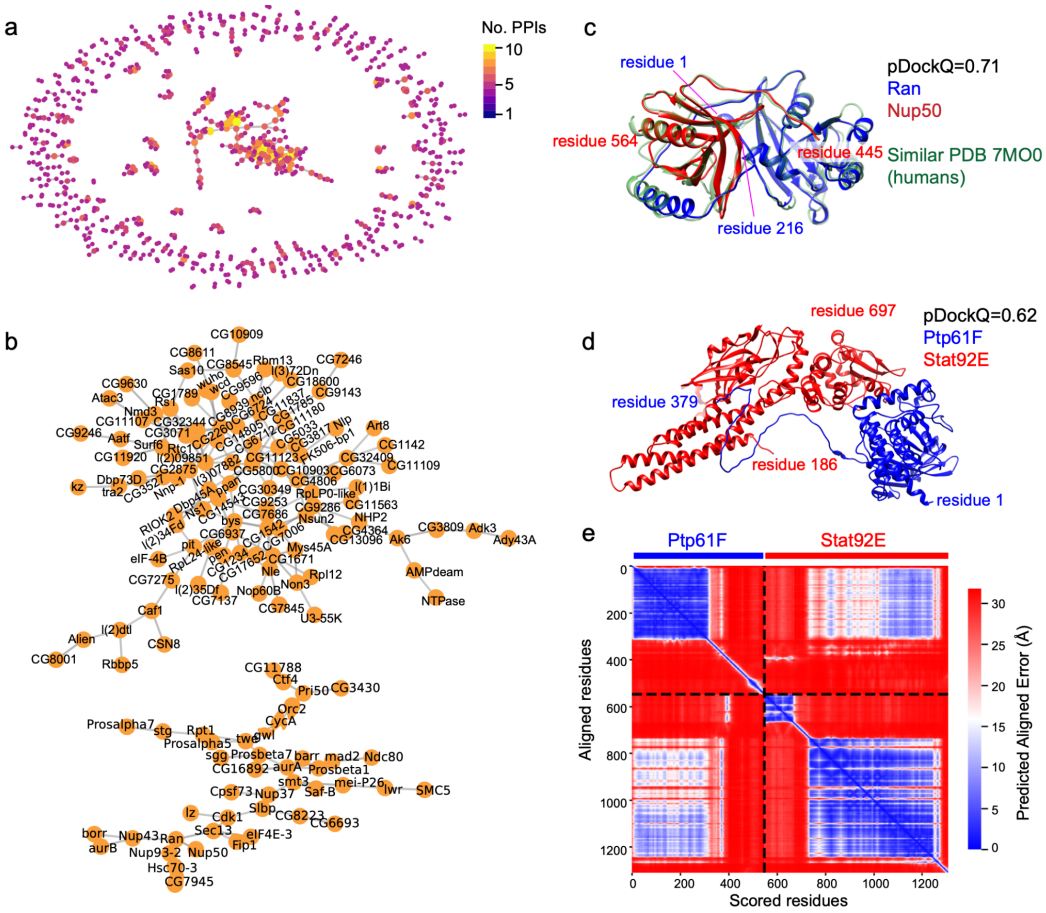

Figure 2. Predicted high-confidence interactions with limited or no support in the STRING physical subnetwork. a) The PPI network of predicted high-confidence interactions that have STRING confidence score < 0.4 in the STRING physical subnetwork. b) The two largest connected subnetworks from a). c) The predicted interactions between Ran (blue) and Nup50 (red) closely match the orthologous complex (green, PDB code 7MO0) in humans. d). The predicted interactions between Ptp61F (blue) and Stat92E (red). Flexible regions that are not in the interfaces were not shown in (c) and (d). e) The predicted aligned error map of Ptp61F-Stat92E interactions. The off-diagonal blue regions indicate that Ptp61F residues 1 to 379 might interact with Stat92E residues 186 to 697 with relatively small predicted aligned errors.

Most predicted dimers involve binding through disordered regions

To further understand the interaction details of these PPIs at the genomic scale, we analyzed the interaction interfaces of each PPI and classified them into different binding modes, including binding through ordered regions and IDRs (Methods). The latter were further classified into three different modes, which are disordered binding, coil-to-order binding, and conditionally folded (CF) binding (Methods). CF binding was hypothesized to occur via conformational selection (30, 52), while coil-to-order binding is more likely driven by induced fit, an entropy-favored mechanism in which a disordered region folds into a specific structure upon binding, adapting its shape to fit the interacting partner (53). Our analysis shows that while almost all of the predicted high confidence PPIs involve ordered regions (98.5%, 3566/3621), a comparable proportion of PPIs (98.8%, 3580/3621) involve IDRs, either entirely or in part (Figure 3A). PPIs that involve long IDRs (≥30 residues in length) also represent a

notable fraction (75.1%, 2719/3621) (Figure 3A). When classifying each PPI that involves IDR to the above mentioned three categories, we found many interactions involving conditionally folded binding (51.7%, 1873/3621) and coil-to-order binding (21.0%, 762/3621) (Figure 2A). The results highlight the important role of IDRs in mediating protein-protein interactions at the genomic scale.

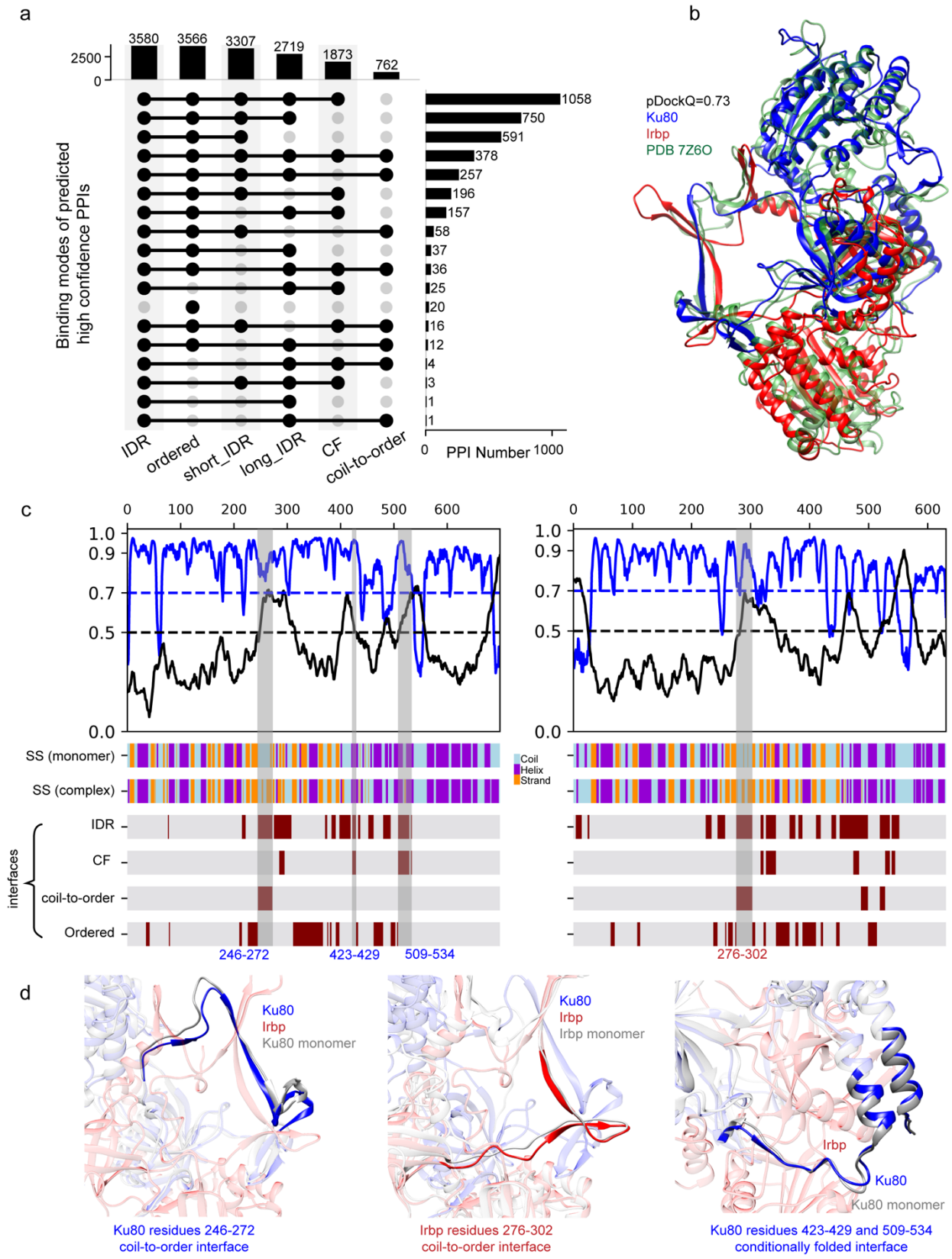

Figure 3. Many PPIs involve IDRs. a) Number of PPIs with disordered (IDR), ordered, short IDR, long IDR, conditionally folded (CF), and coil-to-order interfaces. b) An example of the predicted PPI structure of Ku80/Irbp with its similar PPI structure of their homologs in humans (PDB code 7Z6O). c). PLDDT (blue line) and intrinsic structural disorder (ISD, black

line) of Ku80 and Irbp in their monomer states. PLDDT values were scaled to 0 to 1 by dividing 100. The secondary structures of Ku80 and Irbp in monomers, SS (monomer), and complex states, SS (complex), were shown in the middle panels, with coil in lightblue, helix in purple, and strand in orange. Different interfaces were shown in bottom panels, including disordered binding interface (labeled as IDR), conditionally folded binding interface (labeled as CF), coil-to-order binding interface (labeled as coil-to-order), and ordered binding interface (labeled as ordered). The presence of one binding mode at specific residue index is indicated by darker maroon. d) Examples of disordered (left panel), coil-to-order (middle panel), and CF (right panel) interfaces in the predicted Ku80/Irbp structure. The corresponding regions in Ku80 and Irbp were shaded in (c).

Note that the three binding modes are not mutually exclusive when two proteins interact, as there might be multiple IDRs in the complex that belong to different binding modes. For example, Ku80 and Irbp are important components in DNA damage response. The predicted PPI structure was highly confident with a pDockQ of 0.73, which matches the homologous human complex structure (PDB code 7Z6O, Figure 3B). Through detailed structural analysis, we found that there are several disordered regions in both Ku80 and Irbp monomers involved in their interactions (Figure 3C). Among these disordered interfaces, residues 245-284 in Ku80 (Ku80_245-284) undergo coil-to-order transitions, where the coil secondary structure in Ku80 monomer becomes a strand in the complex (Figure 3D, left panel). Interestingly, the Irbp interface (residues 276-302) that interacts with Ku80 (residues 246-272) also undergoes a coil-to-order transition (Figure 3D, bottom panel). Besides these interfaces, we also observed that some interfaces between Ku80 and Irbp are conditionally folded regions. For example, two of the interfaces in Ku80, Ku80_423-429 and Ku80_509-534, were predicted to be disordered (Figure 3C, left panel). However, they were also predicted to have high PLDDTs (≥ 70) in the Ku80 monomer and contain helices in the monomer (Figure 3C, left panel). Such structural versatility may enable proteins such as Ku80 and Irbp to fine-tune their interactions in response to cellular conditions, particularly during processes such as DNA damage response. The complex roles and various binding modes of IDRs suggest that a single PPI can involve multiple binding mechanisms simultaneously, highlighting the structural complexity and dynamic nature of disordered regions in mediating protein-protein interactions.

Example of IDR interfaces with different binding mechanisms and evolutionary origins

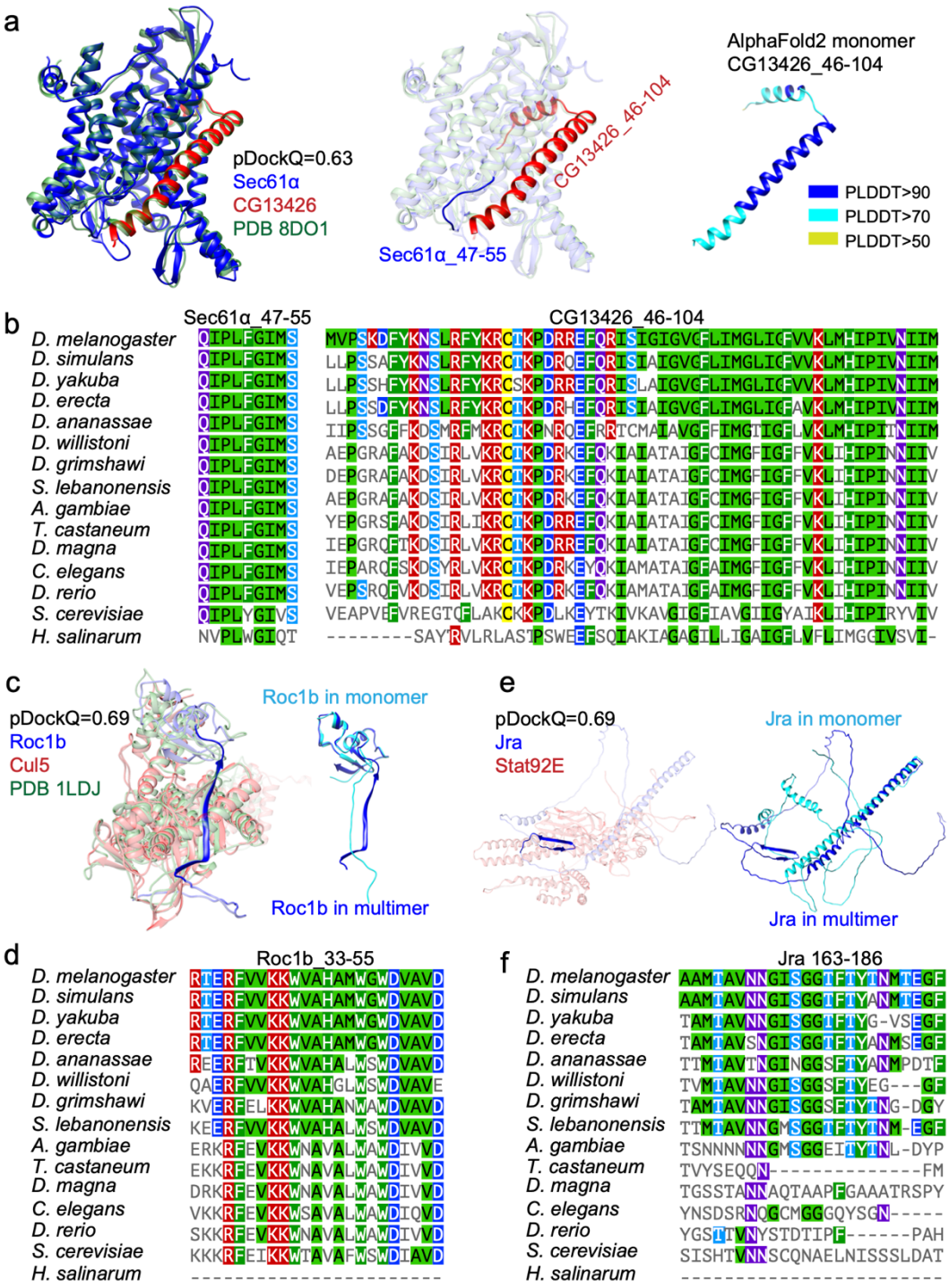

Figure 4. PPI examples illustrate different IDR interfaces with different evolutionary origins. a) interact through a coil region in Sec61α and a conditionally folded region in CG13426. b) Both interfaces of Sec61α and CG13426 are highly conserved and have very early evolutionary origins in eukaryotes. c) Roc1b undergoes coil-to-order transitions when binding to Cul5. d) The coil-to-order interface for Roc1b is highly conserved and has early evolutionary origins. e) Jra undergoes coil-to-order transitions when binding to Stat92E. f) The coil-to-order interface for Jra has a relatively recent origin and is highly conserved after its origin.

IDR interfaces can either be conserved or lineage specific. Here, we present examples where the IDR interfaces have different evolutionary origins and undergo different binding mechanisms. One example is the predicted PPI structure between Sec61α and CG13426 (Figure 4A, left panel), which resembles the human homolog structure (PDB code 1LDJ). Sec61α and CG13426 are both subunits of the translocon; their interactions are important for the transportation of peptides onto the ER membrane. The disordered interface in Sec61α is Sec61α residue 47-55, which does not involve coil-to-order transitions upon binding. The interface in CG13426, residue 46-104, is predicted to be conditionally folded, i.e., it is predicted to be disordered but with high PLDDT (>70) in the monomer states (Figure 4A, right panel). Despite these differences in binding mechanism and intrinsic disorder propensity, both interface regions are highly conserved across diverse organisms (Figure 4B).

In another example, Roc1b/Cul5 is a key interaction in the multisubunit Cullin-RING E3 ubiquitin ligase. The predicted PPI structure showed that the interface between Roc1b and Cul5 involves a long β strand (Figure 4C, left panel). However, this strand was not observed in the monomer structure of Roc1b (Figure 4C, right panel). Thus, we categorized this interface as coil-to-order interface. We showed that this interface has an early evolutionary origin and is highly conserved across different taxa (Figure 4D, left panel). A third example is a recently originated coil-to-order interface in Jra (Figure 4E, 4F). In this example, Jra residues 163 to 186 were disordered with coil structure in a monomer state (Figure 4E). It forms an antiparallel β strand structure upon binding to its partner Stat92E. The interface is relatively new in evolution, which can date back to Arthropoda species (Figure 4F).

An interactive web interface for exploring, validating, and expanding protein interaction hypotheses

Building on our predictions and analysis, we developed an interactive web platform for scientists to explore and generate new hypotheses from all the 27,711 predicted interactions. The platform is accessible at https://jhpanda.github.io/DmelPPI_web/. The homepage of the platform provides a summary of the full prediction datasets, including FBpp identifiers, gene symbols, model quality metrics, and a quick summary of prediction confidence. A search function allows users to query genes or proteins of interest by gene symbols and FBpp identifiers. Users can search a protein pair using FBpp identifiers, gene symbols, or FBgn numbers, separated by space, e.g. “Ku80 Irbp”. After selecting an interaction, users are directed to a detailed result page by clicking “View Details”. In the details page, users can view the predicted structures, predicted aligned error map, explore the interaction details such as different binding modes of the interfaces. The interaction details were only shown for predictions with high and acceptable confidence at pDockQ ≥ 0.23 and other possible interactions with ipTM ≥50 and PAE ≤ 5.0 Å. The structure view is interactive, allowing users to select and highlight interface residues with a single click. Besides, the detailed result page also provides protein annotations, the position of the protein pair in the predicted high-confidence PPI network, a link to the interaction evidence in STRING database, etc (Figure 5). The website is intended not only to present the predictions, but also to facilitate functional exploration,

comparative analysis, and the generation of mechanistic hypotheses. By integrating structural and functional information, it offers a useful starting point for both experimental follow-up and computational investigation.

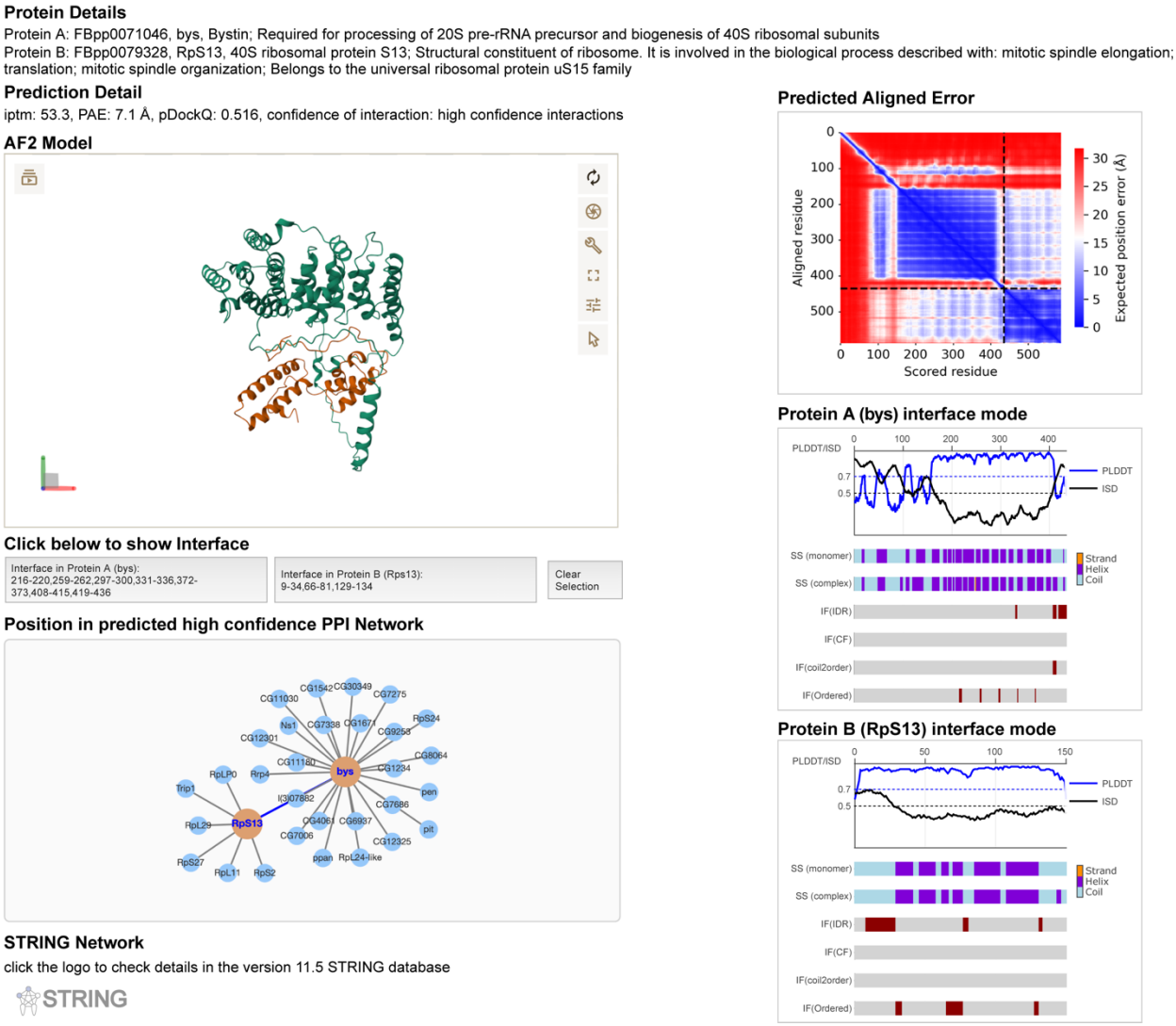

Figure 5. Snapshot of the detailed page showing interactions between bys and RpS13 in *D. melanogaster*. The detailed page includes protein annotations, prediction details, interactive structure viewer, position in predicted high confidence PPI network, link to STRING database version 11.5, predicted aligned error heatmap, and interface binding modes.

Discussion

In this work, we used AlphaFold2 Multimer to examine and predict protein-protein interactions of *D. melanogaster* proteins in the STRING dataset. Notably, we found that 81% of AlphaFold's high-confidence predictions aligned with some physical evidence in STRING (score ≥0.4, with about half ≥0.9). We also showed that the remaining 19% predictions, while lacking reasonable direct evidence of physical interaction, may still represent biologically relevant interactions. Our predictions highlight the importance of incorporating functional

associations in guiding the prediction of protein-protein interactions. In the meantime, our predictions offer a high-confidence resource for hypothesis generation in future functional and mechanistic studies.

Our study also reveals notable discrepancies between computationally or functionally inferred protein-protein interactions and direct physical interactions. Among the 27,711 high-confidence STRING pairs (score ≥ 0.9) analyzed, only 13.1% (3,621 pairs) were predicted as high-confidence physical interactions by AlphaFold2 Multimer (pDockQ ≥ 0.5). Several factors may account for this discrepancy: (1) some interactions are merely functional associations, they don't physically interact – for example, proteins that exhibit genetic interactions or co-occur in cellular pathways but do not physically interact with each other; (2) previously identified interactions from high-throughput studies may include artifacts – for example, indirect interactions that require other components, or those involving low-abundance proteins that are not reliably detected (54); and (3) AlphaFold2 Multimer combined with pDockQ overlooked predicted aligned error and may preferentially predict interactions with large interfaces (55), potentially overlooking biologically meaningful but transient, real interactions (56). Indeed, except for the total 8,108 acceptable interactions (including 3,621 high-confidence ones), there are another 2,910 interactions with minimal inter-protein predicted aligned error smaller than 10 Å or interface pTM greater than 50 (ipTM>50), indicating these interactions may be potentially biologically relevant but could be dynamic and transient (18).

One notable observation was that functional associations that show the highest STRING confidence scores in co-expression and text-mining, had similar proportions predicted as high confidence physical interactions as those supported primarily by experimental evidence (Figure 1A). This highlights the broader need for experimental validation of such associations. In particular, AlphaFold2 Multimer predictions supported by functional evidence such as text mining and lacking evidence from STRING physical subnetwork warrant further experimental investigation. Independent high-throughput experimental approaches such as cross-linking mass spectrometry or targeted mutagenesis studies could help resolve ambiguities in these cases, including whether the predicted interaction is direct or indirect, whether it occurs under physiological conditions, or whether it depends on additional binding partners.

Our results also underscore the pivotal role of intrinsically disordered regions in mediating PPIs at a genomic scale. Our findings reveal that over 98% of predicted high confidence PPIs involve at least some level of intrinsic disorder, from short IDRs (5-30 residues, 3307/3621) to long IDRs (over 30 residues, 2719/3621). This highlights IDRs as important players in cellular interactomes, possibly enabling dynamic and context-dependent binding (53, 57, 58). For instance, the Ku80–Irbp complex simultaneously employs coil-to-order and conditionally folded binding, illustrating how structural versatility fine-tunes interactions in processes like the DNA damage response.

Evolutionarily, IDR interfaces exhibit remarkable diversity, from being highly conserved to lineage specific. However, the reliance on computational predictions warrants caution about misidentification and underidentification. While computationally predicted models often align well with experimental structures, such as the case of Ku80/Irbp in this study and the results from recent studies (18, 26, 59) these predictions still require careful interpretation, especially for IDRs due to their inherently dynamic and context-dependent nature (60). By designing sophisticated, dedicated metrics, we may be able to distinguish true interactions at high precision (18, 24, 47, 55, 56, 61, 62). Future development of specialized machine learning models for predicting the structure and function of IDRs (63, 64), together with advances in high-quality experimental techniques (65), will enhance our ability to characterize the full spectrum of cellular PPIs – including both stable and transient interactions.

Data availability

All predicted structures are available in figureshare (doi.org/10.6084/m9.figshare.29319614) and the DmelPPI website (https://jhpanda.github.io/DmelPPI_web/), which will be made public upon publication/acceptance.

Code availability

All scripts and codes are available in GitHub (https://github.com/LiZhaoLab/DmelPPI).

Acknowledgements

We thank members of Zhao lab for helpful discussions. We thank The Rockefeller University High Performance Computing (HPC) Center for the support in computation.

Author Contribution

J.P. and L.Z. conceived the project, J.P. carried out all analyses with input from L.Z., and J.P. and L.Z. wrote the paper.

Competing Interests Statement

The authors declare no competing interest declaration.

Funding

This work was supported by National Institutes of Health (NIH) MIRA R35GM133780 and the Allen Distinguished Investigator Award from Paul G. Allen Family Foundation to L.Z. The work was also partly supported by the Shapiro-Silverberg Fund for the Advancement of Translational Research to J.P.

References

1. Csizmok,V., Follis,A.V., Kriwacki,R.W. and Forman-Kay,J.D. (2016) Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chem. Rev.*, **116**, 6424–6462.

2. Nooren,I.M.A. and Thornton,J.M. (2003) Diversity of protein-protein interactions. *EMBO J.*, **22**, 3486–3492.

3. Perkins,J.R., Diboun,I., Dessailly,B.H., Lees,J.G. and Orengo,C. (2010) Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure*, **18**, 1233–1243.

4. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S., *et al.* (2005) A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell*, **122**, 957–968.

5. Greenblatt,J.F., Alberts,B.M. and Krogan,N.J. (2024) Discovery and significance of protein-protein interactions in health and disease. *Cell*, **187**, 6501–6517.

6. Morris,J.H., Knudsen,G.M., Verschueren,E., Johnson,J.R., Cimermancic,P., Greninger,A.L. and Pico,A.R. (2014) Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. *Nat. Protoc.*, **9**, 2539–2554.

7. Fields,S. and Song,O. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.

8. Vasilescu,J., Guo,X. and Kast,J. (2004) Identification of protein-protein interactions using ***in vivo*** cross-linking and mass spectrometry. *PROTEOMICS*, **4**, 3845–3854.

9. Roux,K.J., Kim,D.I., Raida,M. and Burke,B. (2012) A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.*, **196**, 801–810.

10. Branon,T.C., Bosch,J.A., Sanchez,A.D., Udeshi,N.D., Svinkina,T., Carr,S.A., Feldman,J.L., Perrimon,N. and Ting,A.Y. (2018) Efficient proximity labeling in living cells and organisms with TurboID. *Nat. Biotechnol.*, **36**, 880–887.

11. Yu,J., Pacifico,S., Liu,G. and Finley,R.L. (2008) DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, **9**, 461.

12. Lehne,B. and Schlitt,T. (2009) Protein-protein interaction databases: keeping up with growing interactomes. *Hum. Genomics*, **3**, 291–297.

13. Koegl,M. and Uetz,P. (2008) Improving yeast two-hybrid screening systems. *Brief. Funct. Genomic. Proteomic.*, **6**, 302–312.

14. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P., *et al.* (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.

15. Chakrabarti,P. and Chakravarty,D. (2022) Intrinsically disordered proteins/regions and insight into their biomolecular interactions. *Biophys. Chem.*, **283**, 106769.

16. Mollica,L., Bessa,L.M., Hanoulle,X., Jensen,M.R., Blackledge,M. and Schneider,R. (2016) Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment. *Front. Mol. Biosci.*, **3**.

17. Mittag,T., Kay,L.E. and Forman-Kay,J.D. (2010) Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.*, **23**, 105–116.

18. Omidi,A., Møller,M.H., Malhis,N., Bui,J.M. and Gsponer,J. (2024) AlphaFold-Multimer accurately captures interactions and dynamics of intrinsically disordered protein regions. *Proc. Natl. Acad. Sci.*, **121**, e2406407121.

19. Orand,T. and Jensen,M.R. (2025) Binding mechanisms of intrinsically disordered proteins: Insights from experimental studies and structural predictions. *Curr. Opin. Struct. Biol.*, **90**, 102958.

20. Morris,O.M., Torpey,J.H. and Isaacson,R.L. (2021) Intrinsically disordered proteins: modes of binding with emphasis on disordered domains. *Open Biol.*, **11**, 210222.

21. Uversky,V.N. (2013) Intrinsic Disorder-based Protein Interactions and their Modulators. *Curr. Pharm. Des.*, **19**, 4191–4213.

22. Evans,R., O'Neill,M., Pritzel,A., Antropova,N., Senior,A., Green,T., Žídek,A., Bates,R., Blackwell,S., Yim,J., *et al.* (2021) Protein complex prediction with AlphaFold-Multimer. 10.1101/2021.10.04.463034.

23. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

24. Burke,D.F., Bryant,P., Barrio-Hernandez,I., Memon,D., Pozzati,G., Shenoy,A., Zhu,W., Dunham,A.S., Albanese,P., Keller,A., *et al.* (2023) Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.*, **30**, 216–225.

25. Homma,F., Huang,J. and Van Der Hoorn,R.A.L. (2023) AlphaFold-Multimer predicts cross-kingdom interactions at the plant-pathogen interface. *Nat. Commun.*, **14**, 6040.

26. Lee,C.Y., Hubrich,D., Varga,J.K., Schäfer,C., Welzel,M., Schumbera,E., Djokic,M., Strom,J.M., Schönfeld,J., Geist,J.L., *et al.* (2024) Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation. *Mol. Syst. Biol.*, **20**, 75–97.

27. Peng,J., Svetec,N., Molina,H. and Zhao,L. (2024) The Origin and Evolution of Sex Peptide and Sex Peptide Receptor Interactions. *Mol. Biol. Evol.*, **41**, msae065.

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

29. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.

30. Piovesan,D., Monzon,A.M. and Tosatto,S.C.E. (2022) Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci. Publ. Protein Soc.*, **31**, e4466.

31. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P., Brown,C.J. and Dunker,A.K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins Struct. Funct. Genet.*, **53**, 566–572.

32. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

33. Dor,O. and Zhou,Y. (2007) Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins Struct. Funct. Bioinforma.*, **66**, 838–845.

34. Hagberg,A.A., Schult,D.A. and Swart,P.J. (2008) Exploring Network Structure, Dynamics, and Function using NetworkX. In Varoquaux,G., Vaught,T., Millman,J. (eds), *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA, pp. 11–15.

35. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, **13**, 2498–2504.

36. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

37. Katoh,K. and Standley,D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.

38. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.

39. Kim,W., Mirdita,M., Levy Karin,E., Gilchrist,C.L.M., Schweke,H., Söding,J., Levy,E.D. and Steinegger,M. (2025) Rapid and sensitive protein complex alignment with Foldseek-Multimer. *Nat. Methods*, **22**, 469–472.

40. wwPDB consortium, Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Costanzo,L.D., Christie,C., Duarte,J.M., Dutta,S., *et al.* (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.

41. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

42. Orchard,S., Kerrien,S., Abbani,S., Aranda,B., Bhate,J., Bidwell,S., Bridge,A., Briganti,L., Brinkman,F.S.L., Cesareni,G., *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.

43. Oughtred,R., Rust,J., Chang,C., Breitkreutz,B., Stark,C., Willems,A., Boucher,L., Leung,G., Kolas,N., Zhang,F., *et al.* (2021) The BIOGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.*, **30**, 187–200.

44. Grigoriev,A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Res.*, **29**, 3513–3519.

45. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. and Gerstein,M. (2003) A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*, **302**, 449–453.

46. Ramani,A.K., Li,Z., Hart,G.T., Carlson,M.W., Boutz,D.R. and Marcotte,E.M. (2008) A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. *Mol. Syst. Biol.*, **4**, 180.

47. Bryant,P., Pozzati,G. and Elofsson,A. (2022) Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.*, **13**, 1265.

48. Guruharsha,K.G., Rual,J.-F., Zhai,B., Mintseris,J., Vaidya,P., Vaidya,N., Beekman,C., Wong,C., Rhee,D.Y., Cenaj,O., *et al.* (2011) A Protein Complex Network of Drosophila melanogaster. *Cell*, **147**, 690–703.

49. Bley,C.J., Nie,S., Mobbs,G.W., Petrovic,S., Gres,A.T., Liu,X., Mukherjee,S., Harvey,S., Huber,F.M., Lin,D.H., *et al.* (2022) Architecture of the cytoplasmic face of the nuclear pore. *Science*, **376**, eabm9129.

50. Buszard,B.J., Johnson,T.K., Meng,T.-C., Burke,R., Warr,C.G. and Tiganis,T. (2013) The Nucleus- and Endoplasmic Reticulum-Targeted Forms of Protein Tyrosine Phosphatase 61F Regulate *Drosophila* Growth, Life Span, and Fecundity. *Mol. Cell. Biol.*, **33**, 1345–1356.

51. Saadin,A. and Starz-Gaiano,M. (2016) Identification of Novel Regulators of the JAK/STAT Signaling Pathway that Control Border Cell Migration in the *Drosophila* Ovary. *G3 GenesGenomesGenetics*, **6**, 1991–2002.

52. Alderson,T.R., Pritišanac,I., Kolarić,Đ., Moses,A.M. and Forman-Kay,J.D. (2023) Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc. Natl. Acad. Sci.*, **120**, e2304302120.

53. Van Der Lee,R., Buljan,M., Lang,B., Weatheritt,R.J., Daughdrill,G.W., Dunker,A.K., Fuxreiter,M., Gough,J., Gsponer,J., Jones,D.T., *et al.* (2014) Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.*, **114**, 6589–6631.

54. Elhabashy,H., Merino,F., Alva,V., Kohlbacher,O. and Lupas,A.N. (2022) Exploring protein-protein interactions at the proteome level. *Structure*, **30**, 462–475.

55. Zhu,W., Shenoy,A., Kundrotas,P. and Elofsson,A. (2023) Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics*, **39**, btad424.

56. Kim,A.-R., Hu,Y., Comjean,A., Rodiger,J., Mohr,S.E. and Perrimon,N. (2024) Enhanced Protein-Protein Interaction Discovery via AlphaFold-Multimer. 10.1101/2024.02.19.580970.

57. Shimizu,K. and Toh,H. (2009) Interaction between Intrinsically Disordered Proteins Frequently Occurs in a Human Protein–Protein Interaction Network. *J. Mol. Biol.*, **392**, 1253–1265.

58. Tompa,P. and Fuxreiter,M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.*, **33**, 2–8.

59. Ibrahim,T., Khandare,V., Mirkin,F.G., Tumtas,Y., Bubeck,D. and Bozkurt,T.O. (2023) AlphaFold2-multimer guided high-accuracy prediction of typical and atypical ATG8-binding motifs. *PLOS Biol.*, **21**, e3001962.

60. Strom,J.M. and Luck,K. (2025) Bias in, bias out – AlphaFold-Multimer and the structural complexity of protein interfaces. *Curr. Opin. Struct. Biol.*, **91**, 103002.

61. Edmunds,N.S., Genc,A.G. and McGuffin,L.J. (2024) Benchmarking of AlphaFold2 accuracy self-estimates as indicators of empirical model quality and ranking: a comparison with independent model quality assessment programmes. *Bioinformatics*, **40**, btae491.

62. Schmid,E.W. and Walter,J.C. (2025) Predictomes, a classifier-curated database of AlphaFold-modeled protein-protein interactions. *Mol. Cell*, **85**, 1216-1232.e5.

63. Tesei,G., Trolle,A.I., Jonsson,N., Betz,J., Knudsen,F.E., Pesce,F., Johansson,K.E. and Lindorff-Larsen,K. (2024) Conformational ensembles of the human intrinsically disordered proteome. *Nature*, **626**, 897–904.

64. Lotthammer,J.M., Ginell,G.M., Griffith,D., Emenecker,R.J. and Holehouse,A.S. (2024) Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods*, **21**, 465–476.

65. Aspromonte,M.C., Nugnes,M.V., Quaglia,F., Bouharoua,A., DisProt Consortium, Sagris,V., Promponas,V.J., Chasapi,A., Fichó,E., Balatti,G.E., *et al.* (2024) DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Res.*, **52**, D434–D441.