Trends in
Biochemical Sciences

CellPress

Review

Recent advances in predicting and modeling protein–protein interactions

Jesse Durham,[1,2,3,6] Jing Zhang,[1,2,3,6] Ian R. Humphreys,[4,5] Jimin Pei,[1,2,3] and Qian Cong[1,2,3,*]

Protein–protein interactions (PPIs) drive biological processes, and disruption of PPIs can cause disease. With recent breakthroughs in structure prediction and a deluge of genomic sequence data, computational methods to predict PPIs and model spatial structures of protein complexes are now approaching the accuracy of experimental approaches for permanent interactions and show promise for elucidating transient interactions. As we describe here, the key to this success is rich evolutionary information deciphered from thousands of homologous sequences that coevolve in interacting partners. This covariation signal, revealed by sophisticated statistical and machine learning (ML) algorithms, predicts physiological interactions. Accurate artificial intelligence (AI)-based modeling of protein structures promises to provide accurate 3D models of PPIs at a proteome-wide scale.

Highlights

Deciphering coevolutionary signals in protein sequences and applying deep learning methods such as AlphaFold have led to breakthroughs in modeling protein structures and interactions.

The accuracy of interaction partner detection and structural modeling or protein complexes by computational methods now approaches experimental methods, and we are entering a new era where computation will play an essential role in both tasks.

We expect rapid progress in characterizing human PPIs, thus enabling biomedical applications such as interpreting pathogenic variants, developing drugs to target PPIs, and designing protein binders to regulate protein function.

We still face challenges in modeling transient and weak interactions, understanding the interactions mediated by intrinsically disordered regions (IDRs), expanding to other molecules such as polysaccharides and lipids, and moving towards modeling the entire cell.

Significance and general approaches for PPI studies

PPIs play essential roles in biological processes, and 80% of proteins have been shown to interact with other proteins to perform their primary functions [1]. Disruption of PPIs by mutations in interface residues, conformational changes, or a lack of binding partners leads to malfunction of cellular pathways and causes diseases such as cancer or neurodegeneration [2–4]. Infectious bacteria and viruses invade host cells via interactions with the host cell-surface receptors [5]. Once inside, other virulence factors interact with host proteins to evade the host immune response, exploit host replication machinery, and spread within the host [6]. Characterizing the interaction partners of proteins and determining the spatial structures of protein complexes are thus crucial for uncovering the mechanisms of genetic and infectious diseases and for developing treatment strategies. Moreover, PPI studies offer shortcuts to elucidating the function of poorly characterized proteins: the function of a protein can be deduced if the function of its binding partner is known [7]. From the perspectives of both basic science and biomedical applications, the investigation into PPIs has attracted significant attention.

Current knowledge about PPIs is mostly derived from *in vitro* and *in vivo* experiments. At the same time, advances in *in silico* methods have increased our ability to predict PPIs (see Figure I in Box 1). Two recent developments are placing computational approaches nearly on a par with experimental ones. First, the deluge of genomic data provides rich evolutionary information. Vast arrays of aligned protein sequences translated from genomes provide confident statistics on covariation between positions in multiple sequence alignments (MSAs) within and between proteins. Such covariation stems from **coevolution** (see Glossary) between residues in direct physical contact, and coevolution derived from deep MSAs can predict inter-residue contacts between proteins with unprecedented accuracy [8,9]. Second, breakthroughs in protein structure prediction, spearheaded by publicly available tools such as AlphaFold [10] and RoseTTAfold [11], can predict and model nearly any protein at near-atomic accuracy. This breakthrough opens a new era in structural biology where atomic modeling of PPIs is around the corner.

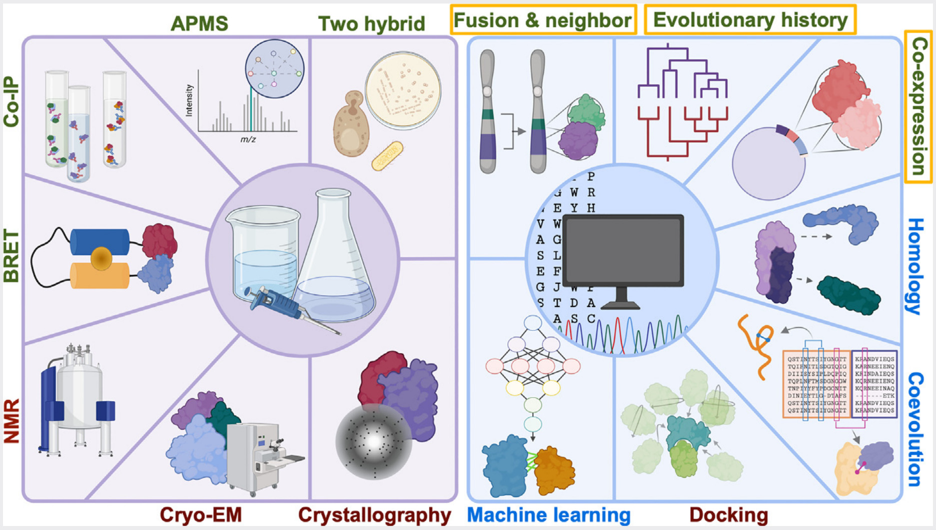
[1]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA
[2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA
[3]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA
[4]Department of Biochemistry, University of Washington, Seattle, WA, USA
[5]Institute for Protein Design, University of Washington, Seattle, WA, USA
[6]These authors contributed equally to this work.

*Correspondence:
qian.cong@utsouthwestern.edu
(Q. Cong).

Box 1. Established experimental and computational methods to study PPIs

Experimental techniques (Figure I, left) to study PPIs can be partitioned into those that detect interacting partners and those that determine the 3D structure of protein complexes. PPI detection methods include widely used low-throughput methods such as coimmunoprecipitation (Co-IP) and bioluminescence resonance energy transfer (BRET), as well as an array of high-throughput methods such as yeast two-hybrid (Y2H) screens [89] and affinity purification coupled to mass spectrometry (APMS) [90]. X-ray crystallography and, to a lesser extent, nuclear magnetic resonance (NMR) were commonly used methods to determine the structures of protein complexes. Recent resolution improvements have made cryo-electron microscopy (cryo-EM) a default method for characterizing 3D protein structures, especially for large complexes of interacting proteins [91]. Low-throughput methods are laborious and time-consuming, whereas high-throughput experiments suffer from considerable false-positive and false-negative rates [92,93]. Thus, computational methods have been developed to complement experimental approaches and accelerate PPI studies.

In silico PPI prediction methods (Figure I, right) can be roughly divided into function/evolution-based and physics-based. Function/evolution-based methods rely on the tendency of interacting proteins to (i) be encoded by genes that are nearby in the genome, especially for Prokaryotes, (ii) co-occur in the same set of species and show similar evolutionary rates across species [94], (iii) be coexpressed at the same time and tissue, (iv) interact in multiple species, allowing detection by homology [95], and (v) coevolve at residues that form direct contacts. Many evolution-based methods identify **functional interaction**s – proteins that function together but not necessarily through direct **physical interactions**. Predicted functional interactions for a wide range of organisms are summarized in the STRING database [85]. Physics-based methods model the forces governing interactions of proteins and attempt to find possible interfaces between two proteins that are complementary in their geometric and physicochemical properties. This process is frequently referred to as protein–protein docking [96]. The evolution- and physics-based methods can be combined, especially through ML techniques, to improve the accuracy of PPI prediction.



Trends in Biochemical Sciences

Figure I. Experimental (left) and computational (right) methods for PPI detection and determination of the structure of protein complexes. Techniques for detecting PPIs and determining the structures of protein complexes are labeled in green and red fonts, respectively. Techniques for PPI detection and structure determination are in blue font. Methods to detect functional interactions (not necessarily direct physical interactions) are in orange boxes.

Glossary

Coevolution: the phenomenon where two different positions (residues) in a protein or two proteins reciprocally affect each other's evolution, which usually results from direct contact between residues in the 3D structures of proteins.

Convolutional neural network (CNN): a class of neural networks that are frequently used in image processing. Neural networks are computational methods inspired by biological neural networks.

Deep learning (DL): a branch of machine learning that comprises multiple layers of neural networks.

Docking: methods that attempt to find a mutual orientation of the 3D structures of two interacting proteins that minimize an energy function over the protein–protein interaction (PPI) interface.

Empirical potentials: energy functions derived from statistical analysis of observed states in existing systems (e.g., experimentally determined protein structures). They are designed to be efficient in computing, and more frequently observed states are evaluated more favorably.

Energy function: the total energy of a particular system computed as a function of the state of the system.

Functional interaction: proteins that function together, not necessarily through physical interaction.

Intrinsically disordered regions (IDRs): regions in a protein that do not adopt a fixed or ordered 3D structure.

Paired MSA: a concatenated multiple sequence alignment (MSA) of proteins A and B used as inputs for coevolution analysis or deep learning networks such as AlphaFold. In this concatenated MSA, the homologs of protein A and protein B are paired by being placed in the same row of the MSA. The generation of a paired MSA converts the problem of modeling two proteins to a problem similar to that of modeling one protein.

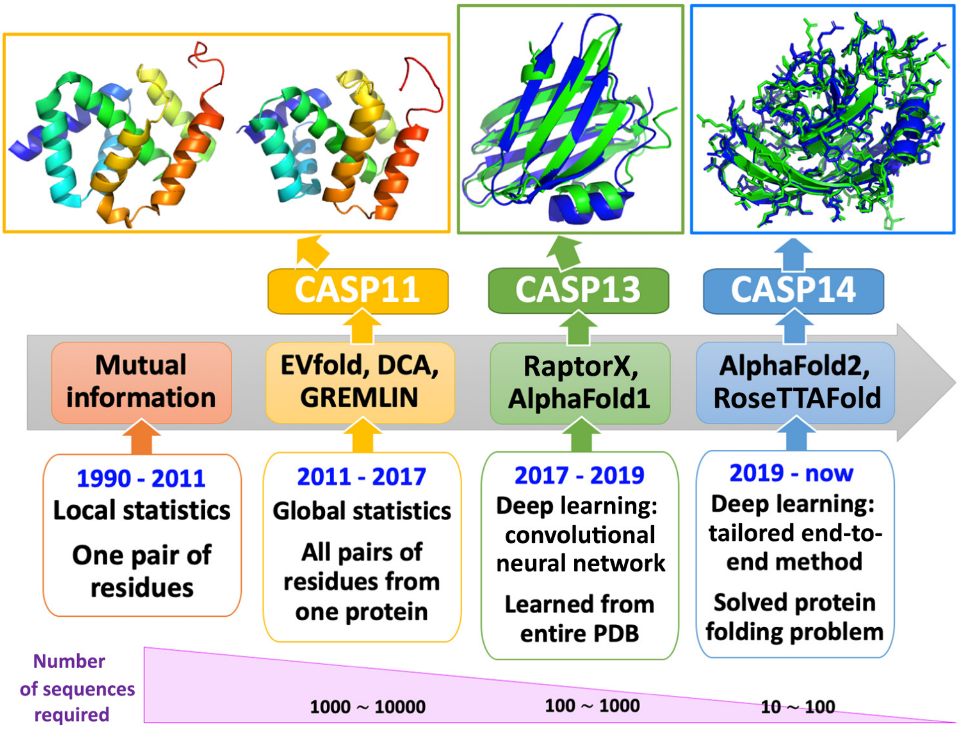
Physical interaction: proteins interact through direct binding to each other.

We review below recent developments in modeling the 3D structures of protein complexes and in predicting the interacting partners of proteins. We first describe how the breakthrough in protein structure prediction has stimulated progress in modeling PPIs. We then discuss how the methodology improvements have enabled large-scale prediction of interaction partners and interactome modeling. Finally, we highlight the biomedical applications of PPI prediction and future challenges for the community to address.

From modeling individual proteins to modeling PPIs

Modeling individual proteins is similar to modeling PPIs because folding and interactions are governed by the same physical laws. Traditionally, accurate modeling of protein 3D structures relies on homologous templates with experimentally determined structures. In the absence of such templates, *de novo* structure prediction was nearly impossible. However, the protein structure prediction community has observed remarkable progress in *de novo* structure prediction over the past decade (Figure 1).

These improvements were initially driven by statistical methods to decipher coevolutionary signals from deep MSAs. Although covariation between positions in MSAs has been discovered and analyzed for several decades [12,13], they were only proven helpful in structure prediction when global statistical methods, such as EVfold [14], direct coupling analysis (DCA) [15], and GREMLIN/CCMpred [16,17] were developed to analyze covariance of the entire protein. Owing to the tendency for coevolving residues to be close in 3D space [12,18], coevolution-based distance constraints can guide structure prediction. Such methods that accurately predict the inter-residue contacts based on coevolution signals from deep (thousands to tens of thousands of diverse sequences) MSAs have significantly improved *de novo* prediction of complex protein topologies around critical assessment of techniques for protein structure prediction (CASP) version 11 [19].



Trends in Biochemical Sciences

Figure 1. Improved protein structure prediction driven by analysis of coevolution signals using statistical and deep learning (DL) methods over the past decade. (Top) Comparisons of experimental and predicted structures illustrate the progress in each round of CASP. CASP10, correct *de novo* prediction of overall topology; CASP13, accurate prediction of protein backbone; CASP14, accurate prediction of protein side chains. (Middle) Representative methods underlying the progress of each stage; the primary ideas behind these methods are explained below. (Bottom) As the methods improve, the number of sequences necessary to ensure accurate predictions decreases. Abbreviations: CASP, critical assessment of techniques for protein structure prediction; DCA, direct coupling analysis; DL, deep learning. PDB, protein data bank.

Wang *et al*. [20] subsequently introduced **convolutional neural networks (CNNs)** to convert the covariation in MSAs into interacting probabilities between residues. After training with extensive protein sequence and structural data, CNNs can integrate coevolution and physiochemical principles of inter-residue contacts. This innovation has significantly boosted the accuracy of structure prediction and has stimulated the development of several other methods [21,22], including the first generation of AlphaFold [23,24], the best-performing method in CASP13 [25]. The ongoing progress in protein structure prediction and the revolution of **deep learning (DL)** techniques have culminated in the recent breakthrough of structure modeling to near-atomic accuracy by AlphaFold (v2) [10] in CASP14 [26].

If experimentally determined or accurately predicted 3D structures of individual proteins are available, protein–protein **docking** methods can be applied to find possible interfaces between two proteins. Docking was a computationally expensive problem, which was elegantly solved using fast Fourier transform for energy evaluation [27,28]. However, significant challenges have been the quality of **energy functions** and conformational changes upon binding. Therefore, *ab initio* docking has never been entirely accurate. Additional information has been used to guide the docking process, including experimental data such as crosslinks [29], homologous protein complexes in Protein Data Bank (PDB) [30], and coevolution between residues across the PPI interface [31]. With these additional data, docking approaches frequently find the correct poses of interacting proteins [32].

Breakthroughs in modeling individual proteins have stimulated rapid improvement in PPI modeling, and this progress is revealed through benchmarking and field-wide challenges (Box 2). First, accurate models of monomeric proteins facilitate the identification of interface residues based on physical and evolutionary properties [33]. A systematic study by Pozzati *et al*. [34] examined the utility of predicted interface residues to guide PPI predictions and they found that a partner-specific interface prediction algorithm, BIPSP [35], was the best for this purpose. Second, DL methods such as AlphaFold can integrate evidence from homologous templates and coevolution signals from MSAs, and they are expected to improve the accuracy of PPI modeling. The most recent critical assessment of predictions of interactions (CAPRI) round 50 revealed that a larger than ever fraction (70–75%) of PPI models submitted by participants were acceptable, but only a smaller fraction were highly accurate [36]. This

Box 2. Benchmarking and community challenges for PPI modeling

Whereas controls always accompany direct experiments, computations rely on benchmarks to estimate their accuracy. Databases that store PPIs detected in experimental studies, such as BioGRID [84], have been used to derive positive controls for computational methods to identify PPIs. However, one should be cautious with such data owing to the considerable rate of false positives from high-throughput experiments [97]. As computational methods become more powerful, they will instead offer an efficient way to detect true PPIs in noisy experimental data. Experimentally determined 3D structures of protein complexes have been used to benchmark methods for PPI modeling. One obstacle in benchmarking protein–protein docking methods is that the bound and unbound forms of interacting proteins frequently adopt different conformations owing to induced fit [98]. Evaluation of docking methods using the bound states of individual proteins will overestimate the accuracy. Curated databases such as the widely used Integrated PPI benchmark (currently updated to V5.5) [99] are designed to overcome this obstacle by providing both the bound and unbound states. A newer database is Dockground [100], which is periodically updated and includes decoy datasets to differentiate true complexes from incorrect complex-like associations of proteins.

Even with carefully designed benchmarks, it is easy to over-fit a method on available experimental datasets and thus reduce its performance on new datasets. Therefore, community-wide challenges where participants blindly apply their methods to unreleased experimental data are essential to evaluate the performance of different methods. Two such challenges, CASP [26] and CAPRI [36], have been crucial in revealing the successes and shortcomings in 3D modeling of proteins and protein complexes over the past few decades. These experiments allow participating groups to learn the best practices from each other, thus stimulating rapid progress in the field.

round of CAPRI was before the availability of AlphaFold (v2), and these results are expected to improve in the next round.

Research is underway to adapt state-of-the-art protein modeling tools, such as AlphaFold and RoseTTAFold [11], to PPIs. The most straightforward but effective approach is to repurpose the AlphaFold network with minor modifications and to transfer learning to data for protein complexes. AlphaFold-Multimer [37], a tool developed by the DeepMind team, correctly predicted the interface for 70% of protein complexes collected from PDB. AF2Complex by Gao *et al*. [38] integrates structural templates for monomeric proteins and enables oligomeric structure prediction without requiring **paired MSAs** of interacting proteins. In another application of AlphaFold to PPI modeling, Bryant *et al*. [39] found that optimizing alignments was important to increase accuracy, and they developed a scoring function to distinguish between correct and incorrect models that identifies ~50% of interacting proteins with only 1% false positives.

Despite the exciting progress, protein complexes are more challenging to model (unsolved problem) than monomers (nearly solved). A recent study by Hadarovich *et al*. [40] revealed that, although protein cores and interfaces are similar in amino acid composition and hydrophobicity distribution, they differ in the packing of side chains and structural compositions. These differences explain why current AI-based approaches perform very well on packing cores of individual proteins but have a higher chance of failure in modeling PPIs, especially those with smaller interfaces [41,42].

In parallel with methodology advances, computational methods have been applied to predict the 3D structure of biologically significant protein complexes on a large scale. Several resources rich in information about PPIs have been available for over a decade. For instance, Interactome INSIDER [43] is a database of predicted interfaces for PPIs detected from large-scale experiments. It combines experimental structures, homology-based modeling, and ML-based prediction of PPI interfaces and integrates such structural data with genetic variations. Primarily focused on humans, it catalogs similar PPIs in seven model organisms. The primary goal is to provide users with an advanced tool to investigate whether variants may be at PPI interfaces and thus deduce the molecular mechanisms of how such variation affects human health. A similar resource is the Interactome3D server [44], which includes experimentally detected PPIs for more organisms and generates predictions for user-provided protein pairs. Previously updated in 2020, it is a valuable resource.

Proteome-wide prediction of interacting proteins

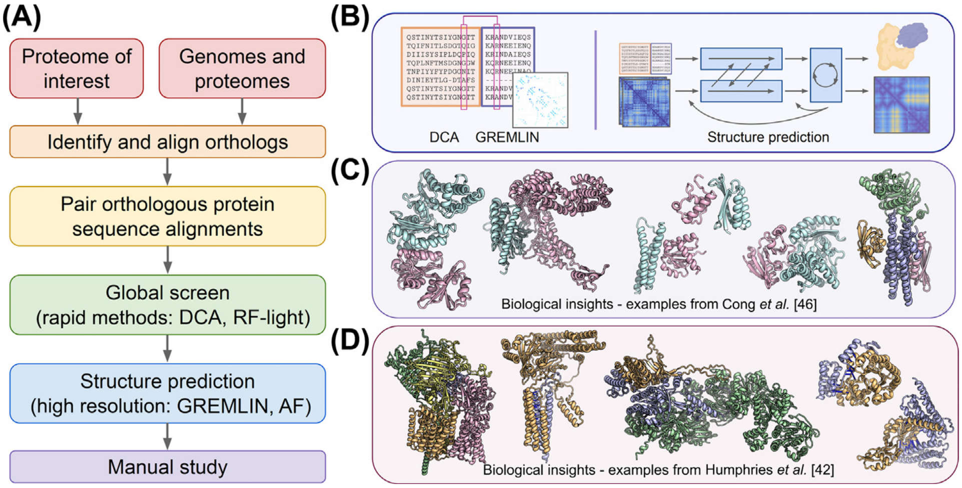
A computational method that detects whether proteins interact *in vivo* has remained a Holy Grail of bioinformatics for many years. This difficulty mainly arises because, at a proteome-wide scale, the signal-to-noise ratio for PPI identification is very low. For example, it is estimated that there are on the order of 10 000 interacting protein pairs in *Escherichia coli* [45], whereas there are 10 000 000 protein pairs (4450 proteins). Therefore, the number of true PPIs is ~1000-fold lower than that of random protein pairs, implying a signal-to-noise ratio of 1:1000. Being encoded in the same operon is nearly the best proxy for such a predictor in bacteria. In Eukaryotes this problem becomes a much greater challenge. Homology- and coevolution-based approaches have produced promising results in identifying PPIs at a proteome-wide scale. ML methods further provide a framework to integrate additional evidence and maximize the performance of PPI prediction.

The PrePPI method and the associated database [7] is a pioneering study that predicts interacting human proteins by using a combination of methods. Most PPIs are modeled by homology, and an interaction between proteins is predicted if they have homologs that have been shown to interact in experimentally determined structures. Additional evidence, such as functional similarity and coexpression, is used to score and rank these predictions. Experimental assays were carried out to validate some (>75% of 19 cases) high-scoring predicted PPIs. As a result, >300 000 predicted human PPIs are included in the PrePPI database.

Starting with the idea that interacting proteins are expected to show stronger coevolution than non-interacting pairs, Cong *et al*. developed a framework for *de novo* identification of PPIs at a proteome-wide scale (Figure 2) [46] in *E. coli*. In addition to coevolution, structure modeling further increases the accuracy of this method that can now surpass that of proteome-wide experimental screens such as yeast two-hybrid (Y2H) and affinity purification coupled to mass spectrometry (APMS) methods. Green *et al.* independently performed a similar study [47], again demonstrating the power of such approaches. The success of coevolution-based PPI screens in Prokaryotes reflects the abundance of prokaryotic sequences. A bacterial protein typically can find >10 000 non-redundant homologs in the databases.

Compared to bacteria, Eukaryotes are expected to have a larger number of unknown PPIs. However, predicting these PPIs is more difficult because of the much smaller number of sequences available for coevolution analysis. Leveraging the breakthrough in protein structure prediction, Humphreys *et al*. [42] embarked on the task of a computational proteome-wide PPI screen in *Saccharomyces cerevisiae*. The authors complemented their previous coevolution-based PPI screen framework [47] with a lightweight version of RoseTTAFold optimized for speed and the standard AlphaFold. Based on the benchmark in yeast, such an approach can recover a quarter of *bona fide* PPIs with a precision of 95% when confident PPIs and non-interacting proteins are mixed at a 1:1000 ratio. Spatial structure models were built for >100 unidentified protein assemblies and >800 known assemblies that lacked experimental 3D structures.

Burke *et al*. [48] explored the application of AlphaFold (v2) to validate and model >65 000 PPIs revealed in experimental studies. They estimated that experimental structures cover <5% of


(A)
Proteome of interest
Genomes and proteomes
Identify and align orthologs
Pair orthologous protein sequence alignments
Global screen (rapid methods: DCA, RF-light)
Structure prediction (high resolution: GREMLIN, AF)
Manual study
(B)
DCA
GREMLIN
Structure prediction
(C)
Biological insights - examples from Cong et al. [46]
(D)
Biological insights - examples from Humphries et al. [42]

Trends in Biochemical Sciences

Figure 2. Overview of coevolution-based protein–protein interaction (PPI) screens. (A) A general workflow for proteome-wide PPI screens that we have developed. RF-light is a fast version of RoseTTAFold that facilitates PPI screening. (B) Illustration of methods used in coevolution-based PPI screens. (Left) Statistical methods that predict inter-protein contacts from paired MSAs. (Right) DL networks that predict the structures of protein complexes from paired MSAs. (C) Examples of protein complexes revealed by applying our methods to *Escherichia coli*. (D) Examples of protein complexes revealed by applying our methods to *Saccharomyces cerevisiae*. Abbreviations: AF, AlphaFold; DCA, direct coupling analysis; DL, deep learning; MSA, multiple sequence alignment.

human PPIs, and learning about the remaining 95% is expected to bring significant insights into human biology. More than 1300 of their confident models of protein complexes are not similar to known structures, and they provide insights into the mechanisms of disease-causing mutations. Similarly, Zhang *et al*. [49] focused on cancer drivers and used AlphaFold to model their interactions with other proteins. These models were further used to interpret the somatic mutation landscape of cancer cells. Pei *et al*. [50] identified and modeled PPIs for human mitochondrial proteins using RoseTTAFold and AlphaFold. Constraining binding partners to a specific subcellular location reduces the noise in the PPI screen and could be a shortcut to elucidating the human PPI network.

The obstacle in human PPI detection and modeling is that state-of-the-art methods such as AlphaFold can generate confident models for only a small fraction (<5%) of PPIs detected in multiple experimental studies. In contrast to the high success rates in reproducing known PDB protein complexes by these methods, this low ratio likely arises for two reasons. First, human PPIs are dominated by weak and transient interactions such as those mediated by **intrinsically disordered regions (IDRs)** that frequently escape experimental structure determination (not in PDB) and AI-based PPI modeling. Second, compared to lower Eukaryotes such as yeast, humans underwent two rounds of whole-genome duplication [51], and duplicated proteins will likely diversify in function and adapt to different interacting partners [52]. Thus, many human proteins lack orthologs that have the same interacting partners in lower Eukaryotes. Therefore, sequences of lower Eukaryotes cannot provide valuable coevolution signals, whereas higher Eukaryotes (i.e., mammals) may not contribute sufficient coevolution signals owing to the low sequence divergence.

AlphaFold and RoseTTAFold are not trained to distinguish interacting proteins from non-interacting proteins. However, they have learned to interpret the coevolutionary signals in MSAs and can incorporate homology-based evidence provided by homologous templates when these networks are trained to predict protein 3D structures using MSAs and homologous structure templates. Therefore, they can be repurposed to model protein complexes and identify interacting proteins.

By contrast, various ML methods have been explicitly trained to distinguish true interacting partners from false positives [53] based on sequences and sequence profiles [54,55]. It is not entirely clear what features were learned by these methods, but homology is likely an essential factor. For example, two proteins are expected to interact if their homologs are already known to interact. These models are frequently trained on PPIs identified from large-scale experiments, but these datasets contain many false positives [56]. In addition, these methods are usually benchmarked by predicting true PPIs among false PPIs that are 10-fold more numerous than true PPIs, whereas the signal-to-noise ratio for a proteome-wide PPI screen is 100-fold lower. Therefore, the good performance of these methods with benchmarks will probably not translate to satisfactory performance in proteome-wide PPI prediction. However, ML provides an excellent framework for integrating homology-based, coevolution-based, physics-based, and function-based evidence. Incorporating state-of-the-art DL models such as AlphaFold into these PPI prediction models may significantly improve their accuracy.

Biomedical applications of PPI prediction

Effect of mutations on binding affinity

Livesey and Marsh [57] analyzed the distribution of genetic variants in protein 3D structures and found an enrichment of pathogenic variants at PPI interfaces. Analysis of cancer somatic mutations also revealed significant enrichment of mutations at PPI interfaces [2]. Quantitative estimates of the effects of mutations are desirable to prioritize pathogenic variants and somatic mutations for disease diagnosis and treatment. Many methods have been developed to evaluate the impact of mutations on protein–protein affinity, and these methods have been thoroughly reviewed [58]. Traditional methods such as FoldX [59] frequently rely on **empirical potentials** to calculate the impact of mutations on PPI affinity. Xiong *et al*. [60] demonstrated that adding profile scores derived from the relative frequency of amino acids in MSAs to the empirical potentials improves their performance.

Recent tools mostly utilize ML, and SKEMPI [61], a database with manually curated binding affinity changes for >7000 mutations, is perfect for training and testing these tools. For example, mCSM-PPI [62] combines graph-based structural signatures, residue-residue contacts, evolution, and energy functions by decision trees, and it performs well on CAPRI targets. Mutations frequently introduce conformational changes in surrounding residues and this is difficult to account for. Therefore, advanced methods such as mCSM-PPI still show moderate accuracy (Pearson's correlation coefficient of ~0.4) on CAPRI blind tests. Zhou *et al*. [63] showed that an end-to-end ML framework, MuPIPR, could outperform state-of-the-art methods. In contrast to other methods requiring the 3D structure of a protein complex as an input, MuPIPR only needs the sequences of interacting partners. Validating these methods with new experimental data in the future is necessary. Nevertheless, a dedicated DL model integrating the sequence, structure, and evolutionary information could perform better in this task.

Host–pathogen interactions

In the wake of pandemics, it is paramount to quickly understand the virulence mechanisms of emerging viruses. Predicting the interacting partners of viral proteins provides a shortcut to elucidating the virulence mechanism [64]. Methods to predict PPIs between host and pathogen are mostly based on homology to known protein complexes detected by sequence and structure similarities [65]. For example, using structural features, P-HIPSTer [66] predicted and modeled 282 000 virus–human interactions for ~1000 human-infecting viruses with a >76% success rate based on experimental validations. In addition, some viral proteins function by 'mimicking' the PPI interfaces through convergent evolution. Thus, in the absence of sequence homology, similarities in PPI interfaces were also used to predict interacting human partners of viral proteins [67].

ML techniques are expected to improve the prediction of host–pathogen interactions by integrating homology-based evidence and other information such as functional association and geometric/physiochemical features of the interface. Bell *et al*. [68] developed a pipeline to model interactions between human and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) proteins using an ML model. This application fetched nearly 75% of known interactions and predicted new interactions for further study. As the field progresses in characterizing the human interactome, we expect further improvement in human–pathogen PPI predictions because of improved *de novo* modeling and the expansion of high-quality 3D structure models of human endogenous PPIs that can serve as templates for predicting human–pathogen interactions.

Protein-binder design

The ability to model protein complexes and predict the impact of mutations on PPI affinity can be repurposed to design protein binders. The additional difficulty in protein-binder design is to efficiently search the vast sequence space to find the optimal points. Nevertheless, there has recently been significant progress in this direction. Dou *et al*. [69] developed a method that first generates a small set of amino acid rotamers that bind to a specific target with favorable energy, and then selects protein backbones to accommodate these favorable rotamers. Cao *et al*. [70] adapted this idea to protein-binder design. They developed a framework within Rosetta to broadly explore binding motifs for a target protein, sample backbones that can accommodate the selected motifs, and optimize the sequences. AI-based structure modeling has stimulated rapid progress in protein-binder design, and recent progress in protein design can be tracked in other reviews [71]. For example, exploiting the RoseTTAFold network, Wang *et al*. [72] developed an approach that starts with binding motifs and fills in sequence and structure to create viable protein scaffolds. Shan *et al.* [73] used DL approaches to redesign the complementarity-determining regions of human antibodies, enabling broad neutralization of SARS-COV-2 variants.

Future directions

Transient interactions

Experimentation suggests that many PPIs are transient and that the proteins may not form a stable complex that can be purified and crystallized [74]. A method to detect such interactions is *in vivo* crosslinking [75]. Ghadie *et al*. [76] correlated the importance of a PPI judged by the number of disease-causing mutations at the PPI interface with how transient the interaction is. They concluded that transient interactions are subject to a similar selection pressure as stable interactions and are thus no less significant [76]. Because coevolutionary signals and binding affinities for transient interactions tend to be lower [77], predicting and modeling such interactions is a future challenge to be addressed.

PPIs via intrinsically disordered regions

IDRs cover >25% of residues in human proteins based on our analysis of human proteins in the AlphaFold Database. Teilum *et al*. [78] considered various examples of IDRs and concluded that their interactions are not less specific than those of globular domains, but are more diverse both in terms of the number of interacting partners and the multivalence of the interactions. Structural flexibility and the repetitive nature of IDRs allow them to accommodate more interacting partners and form multivalent interactions. Alignment of positions in homologous globular domains does not apply to IDRs that function by amino acid composition rather than by positional conservation. Therefore, position-based covariation methods cannot be used to predict interactions between IDRs. ML methods show some success in this task [79], opening the field for further research.

Prospects ahead

Already not far from accurate modeling of stable protein complexes, the field will likely expand to other molecules such as nucleic acids and polysaccharides, and will eventually generate 3D models of living cells and organisms. Methodologically, this may take place through a multidisciplinary approach that marries AI-based methods with molecular dynamics and kinetic models. The latest achievement in this direction is a simulation of a minimal cell with ~500 genes [80]. Although mostly kinetic, this simulation is a step towards atomic details [81]. As we progress to elucidating the 3D interactome of humans and between humans and pathogens, more drugs targeting PPI interfaces might be developed. A common notion is that PPIs are notoriously difficult to design drugs for, but targeting PPIs may be the best solution for some diseases. Several drugs have been developed recently to target PPIs [82]. Alzyoud *et al*. [83] found that drugs targeting PPI interfaces violate established norms of hydrophobicity and size. Thus, specific criteria for PPI-targeting drugs could be developed, and research in this direction should not avoid non-standard compounds that may not appear to be drug-like.

Concluding remarks

Essential resources for the PPI studies discussed here are summarized in Table 1. Deciphering coevolutionary signals in protein sequence alignments and applying advanced DL networks have enabled a breakthrough in protein structure modeling, further transforming *in silico* studies of PPIs. The unprecedented accuracy of modern DL methods heralds a new era where

Outstanding questions

Current methods derived from AlphaFold can accurately model 50–70% of stable protein complexes. What is the reason that they fail in the remaining cases?

Weak and transient interactions dominate human PPIs, but existing computational methods do not perform well on such interactions. In addition, these transient interactions tend to evade experimental structural determination. What is the best way to characterize these interactions?

Interactions mediated by IDRs are poorly understood. Interactions are frequently mediated by amino acid composition (e.g., positive and negative charges) and they do not adopt rigid structures. Therefore, the methods discussed here are not suitable for this task, but how should this challenge be addressed in the future?

Post-translational modifications (PTMs) such as phosphorylation play important roles in regulating PPIs. We observed that AlphaFold could correctly model a fraction of PPIs that are mediated by phosphorylated residues without explicitly knowing the existence of such modifications. Will explicit consideration of PTMs in future methods improve the accuracy of PPI modeling?

PPIs are frequently dynamic, and proteins may change conformation upon binding to other proteins. There is some promise of revealing such dynamics using AlphaFold, but to what extent can AlphaFold be applied in this area?

Characterizing interactions between different species, especially between hosts and pathogens, is of biomedical significance. However, existing computational methods perform poorly on such PPIs because of the lack of coevolution signals. How should the field address this challenge?

How can we model the interactions between proteins and other molecules such as nucleic acids, polysaccharides, lipids, and drugs? How far away we are from being able to model all macromolecular interactions within cells and building 3D models of entire cells at atomic resolution?

Table 1. Major resources for PPI studies

Name	Website	Description
BioGRID [84]	http://thebiogrid.org/	A database of physical, genetic, and chemical interactions. It is regularly updated to include PPIs identified in experiments
STRING [85]	http://string-db.org/	A database of functional interactions. It is regularly updated to include protein pairs that are predicted or tested to function together
CASP	http://predictioncenter.org/	The website for 'critical assessment of protein structure prediction', a place to track progress in predicting protein 3D structure
CAPRI	http://www.ebi.ac.uk/msd-srv/capri/	The website for 'critical assessment of predictions of interactions', a place to track progress in modeling 3D structures of protein complexes
AlphaFold [10]	http://github.com/deepmind/alphafold	GitHub repository for AlphaFold maintained by DeepMind from where the source code of AlphaFold can be downloaded
ColabFold [86]	http://github.com/sokrypton/ColabFold	GitHub repository maintained by Dr Sergey Ovchinnikov who provides Google Colab notebooks to allow everyone to use AlphaFold
Robetta [11]	http://robetta.bakerlab.org	Robetta web server where tools (e.g., RoseTTAFold) developed in David Baker's laboratory for protein structure prediction can be accessed
ClusPro [87]	http://cluspro.bu.edu/login.php	A top-performing server for protein–protein docking in several rounds of CAPRI; the authors combine AlphaFold and ClusPro
AlphaFold Database [88]	http://alphafold.ebi.ac.uk	A web resource for the DeepMind team to deposit their predicted structures. As of November 2022 it contains models for ~200 million proteins
ModelArchive	http://www.modelarchive.org/	A database to deposit predicted structures by modern methods
Interactome INSIDER [43]	http://interactomeinsider.yulab.org/	A resource to explore human genetic variation mapped to determined or predicted interfaces of known PPIs
PrePPI [7]	http://honig.c2b2.columbia.edu/preppi/	A database of determined and predicted PPIs in the human proteome
Interactome3D [44]	http://interactome3d.irbbarcelona.org/	A database of determined and predicted PPIs by homology from an array of organisms. 3D structures of these PPIs are also provided

computation plays a vital role in PPI detection and modeling of protein complexes. Despite rapid development, computational PPI studies face challenges to be addressed in the future (see Outstanding questions). Given the accumulation of experimental PPI data and constant advancement of computational methods, the field should be able to accurately catalog and model a significant portion of the biologically essential PPIs, understand their cellular roles, and learn how to combat diseases caused by perturbations of these interactions in the near future.

Acknowledgments

Q.C. is a Southwestern Medical Foundation endowed scholar. J.Z. is supported by training grant RP210041 from the Cancer Prevention and Research Institute of Texas. This research is also supported by grant I-2095-20220331 to Q.C. from the Welch Foundation. The authors thank Dr Nick V. Grishin for helpful suggestions and discussions. Figure I was drawn with the help of BioRender.

Declaration of interests

The authors declare no conflicts of interest.

References

1. Berggard, T. *et al.* (2007) Methods for the detection and analysis of protein–protein interactions. *Proteomics* 7, 2833–2842
2. Cheng, F. *et al.* (2021) Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nat. Genet.* 53, 342–353
3. Kim, M. *et al.* (2021) A protein interaction landscape of breast cancer. *Science* 374, eabf3066
4. Thompson, T.B. *et al.* (2020) Protein–protein interactions in neurodegenerative diseases: a conspiracy theory. *PLoS Comput. Biol.* 16, e1008267

5. Brito, A.F. and Pinney, J.W. (2017) Protein–protein interactions in virus–host systems. *Front. Microbiol.* 8, 1557
6. Zheng, L.L. *et al.* (2014) The domain landscape of virus–host interactomes. *Biomed. Res. Int.* 2014, 867235
7. Zhang, Q.C. *et al.* (2012) Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 490, 556–560
8. Hopf, T.A. *et al.* (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3, e03430
9. Ovchinnikov, S. *et al.* (2014) Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* 3, e02030
10. Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589
11. Baek, M. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876
12. Gobel, U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18, 309–317
13. de Juan, D. *et al.* (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.* 14, 249–261
14. Marks, D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6, e28766
15. Morcos, F. *et al.* (2011) Direct-coupling analysis of residue co-evolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* 108, E1293–E1301
16. Kamisetty, H. *et al.* (2013) Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* 110, 15674–15679
17. Seemayer, S. *et al.* (2014) CCMpred – fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* 30, 3128–3130
18. Altschuh, D. *et al.* (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* 193, 693–707
19. Moult, J. *et al.* (2016) Critical assessment of methods of protein structure prediction: progress and new directions in round XI. *Proteins* 84, 4–14
20. Wang, S. *et al.* (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* 13, e1005324
21. Yang, J. *et al.* (2020) Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* 117, 1496–1503
22. Zheng, W. *et al.* (2021) Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins* 89, 1734–1751
23. AlQuraishi, M. (2019) AlphaFold at CASP13. *Bioinformatics* 35, 4862–4865
24. Senior, A.W. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710
25. Kryshtafovych, A. *et al.* (2019) Critical assessment of methods of protein structure prediction (CASP) – round XIII. *Proteins* 87, 1011–1020
26. Kryshtafovych, A. *et al.* (2021) Critical assessment of methods of protein structure prediction (CASP) – round XIV. *Proteins* 89, 1607–1617
27. Vakser, I.A. (1996) Low-resolution docking: prediction of complexes for underdetermined structures. *Biopolymers* 39, 455–464
28. Padhorny, D. *et al.* (2016) Protein–protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proc. Natl. Acad. Sci. U. S. A.* 113, E4286–E4293
29. Piersimoni, L. *et al.* (2022) Cross-linking mass spectrometry for investigating protein conformations and protein–protein interactions. A method for all seasons. *Chem. Rev.* 122, 7500–7531
30. Kundrotas, P.J. and Vakser, I.A. (2013) Global and local structural similarity in protein–protein complexes: implications for template-based docking. *Proteins* 81, 2137–2142
31. Yu, J. *et al.* (2016) InterEvDock: a docking server to predict the structure of protein–protein interactions using evolutionary information. *Nucleic Acids Res.* 44, W542–W549
32. Jones, G. *et al.* (2022) Elucidation of protein function using computational docking and hotspot analysis by ClusPro and FTMap. *Acta Crystallogr. D Struct. Biol.* 78, 690–697
33. Egbert, M. *et al.* (2022) FTMove: a web server for detection and analysis of cryptic and allosteric binding sites by mapping multiple protein structures. *J. Mol. Biol.* 434, 167587
34. Pozzati, G. *et al.* (2022) Scoring of protein–protein docking models utilizing predicted interface residues. *Proteins* 90, 1493–1505
35. Sanchez-Garcia, R. *et al.* (2019) BIPSPI: a method for the prediction of partner-specific protein–protein interfaces. *Bioinformatics* 35, 470–477
36. Lensink, M.F. *et al.* (2021) Prediction of protein assemblies, the next frontier: the CASP14-CAPRI experiment. *Proteins* 89, 1800–1823
37. Evans, R. *et al.* (2022) Protein complex prediction with AlphaFold-Multimer. *BioRxiv* Published online March 10, 2022. https://doi.org/10.1101/2021.10.04.463034
38. Gao, M. *et al.* (2022) AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* 13, 1744
39. Bryant, P. *et al.* (2022) Improved prediction of protein–protein interactions using AlphaFold2. *Nat. Commun.* 13, 1265
40. Hadarovich, A. *et al.* (2021) Structural motifs in protein cores and at protein–protein interfaces are different. *Protein Sci.* 30, 381–390
41. Yin, R. *et al.* (2022) Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci.* 31, e4379
42. Humphreys, I.R. *et al.* (2021) Computed structures of core eukaryotic protein complexes. *Science* 374, eabm4805
43. Meyer, M.J. *et al.* (2018) Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods* 15, 107–114
44. Mosca, R. *et al.* (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods* 10, 47–53
45. Rajagopala, S.V. *et al.* (2014) The binary protein–protein interaction landscape of Escherichia coli. *Nat. Biotechnol.* 32, 285–290
46. Cong, Q. *et al.* (2019) Protein interaction networks revealed by proteome coevolution. *Science* 365, 185–189
47. Green, A.G. *et al.* (2021) Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* 12, 1396
48. Burke, D.F. *et al.* (2023) Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* 30, 216–225
49. Zhang, J. *et al.* (2022) Computed cancer interactome explains the effects of somatic mutations in cancers. *Protein Sci.* 31, e4479
50. Pei, J. *et al.* (2022) Human mitochondrial protein complexes revealed by large-scale coevolution analysis and deep learning-based structure modeling. *Bioinformatics* 38, 4301–4311
51. Dehal, P. and Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3, e314
52. Marques, A.C. *et al.* (2008) Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol.* 9, R54
53. Zhang, S.-W. and Wei, Z.-G. (2015) Some remarks on prediction of protein–protein Interaction with machine learning. *Med. Chem.* 11, 254–264
54. Li, Y. *et al.* (2021) Robust and accurate prediction of protein–protein interactions by exploiting evolutionary information. *Sci. Rep.* 11, 16910
55. Sledzieski, S. *et al.* (2021) D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein–protein interactions. *Cell Syst.* 12, 969–982
56. Alvarez-Ponce, D. (2017) Recording negative results of protein–protein interaction assays: an easy way to deal with the biases and errors of interactomic data sets. *Brief. Bioinform.* 18, 1017–1020
57. Livesey, B.J. and Marsh, J.A. (2022) The properties of human disease mutations at protein interfaces. *PLoS Comput. Biol.* 18, e1009858
58. Jubb, H.C. *et al.* (2017) Mutations at protein–protein interfaces: small changes over big surfaces have large impacts on human health. *Prog. Biophys. Mol. Biol.* 128, 3–13
59. Schymkowitz, J. *et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388

60. Xiong, P. *et al.* (2017) BindProfX: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *J. Mol. Biol.* 429, 426–434
61. Jankauskaite, J. *et al.* (2019) SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 35, 462–469
62. Rodrigues, C.H.M. *et al.* (2021) mmCSM-PPI: predicting the effects of multiple point mutations on protein–protein interactions. *Nucleic Acids Res.* 49, W417–W424
63. Zhou, G. *et al.* (2020) Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR Genom. Bioinform.* 2, lqaa015
64. de Chassey, B. *et al.* (2008) Hepatitis C virus infection protein network. *Mol. Syst. Biol.* 4, 230
65. Mariano, R. and Wuchty, S. (2017) Structure-based prediction of host–pathogen protein interactions. *Curr. Opin. Struct. Biol.* 44, 119–124
66. Lasso, G. *et al.* (2019) A structure-informed atlas of human–virus interactions. *Cell* 178, 1526–1541
67. Guven-Maiorov, E. *et al.* (2017) Prediction of host–pathogen interactions for *Helicobacter pylori* by interface mimicry and implications to gastric cancer. *J. Mol. Biol.* 429, 3925–3941
68. Bell, E.W. *et al.* (2022) PEPPI: whole-proteome protein–protein interaction prediction through structure and sequence similarity, functional association, and machine learning. *J. Mol. Biol.* 434, 167530
69. Dou, J. *et al.* (2018) De novo design of a fluorescence-activating beta-barrel. *Nature* 561, 485–491
70. Cao, L. *et al.* (2022) Design of protein-binding proteins from the target structure alone. *Nature* 605, 551–560
71. Ding, W. *et al.* (2022) Protein design via deep learning. *Brief. Bioinform.* 23, bbac102
72. Wang, J. *et al.* (2022) Scaffolding protein functional sites using deep learning. *Science* 377, 387–394
73. Shan, S. *et al.* (2022) Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2122954119
74. Chichili, V.P.R. *et al.* (2013) A method to trap transient and weak interacting protein complexes for structural studies. *Intrinsically Disord Proteins* 1, e25464
75. Pertl-Obermeyer, H. and Obermeyer, G. (2020) In vivo cross-linking to analyze transient protein–protein interactions. *Methods Mol. Biol.* 2139, 273–287
76. Ghadie, M.A. and Xia, Y. (2022) Are transient protein–protein interactions more dispensable? *PLoS Comput. Biol.* 18, e1010013
77. Mintseris, J. and Weng, Z. (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* 102, 10930–10935
78. Teilum, K. *et al.* (2021) On the specificity of protein–protein interactions in the context of disorder. *Biochem. J.* 478, 2035–2050
79. Perovic, V. *et al.* (2018) IDPpi: protein–protein interaction analyses of human intrinsically disordered proteins. *Sci. Rep.* 8, 10563
80. Thornburg, Z.R. *et al.* (2022) Fundamental behaviors emerge from simulations of a living minimal cell. *Cell* 185, 345–360
81. Feig, M. and Sugita, Y. (2019) Whole-cell models and simulations in molecular detail. *Annu. Rev. Cell Dev. Biol.* 35, 191–211
82. Lu, H. *et al.* (2020) Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduct. Target. Ther.* 5, 213
83. Alzyoud, L. *et al.* (2022) Structure-based assessment and druggability classification of protein–protein interaction sites. *Sci. Rep.* 12, 7975
84. Oughtred, R. *et al.* (2021) The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* 30, 187–200
85. Szklarczyk, D. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613
86. Mirdita, M. *et al.* (2022) ColabFold: making protein folding accessible to all. *Nat. Methods* 19, 679–682
87. Kozakov, D. *et al.* (2017) The ClusPro web server for protein–protein docking. *Nat. Protoc.* 12, 255–278
88. Tunyasuvunakool, K. *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596
89. Bruckner, A. *et al.* (2009) Yeast two-hybrid, a powerful tool for systems biology. *Int. J. Mol. Sci.* 10, 2763–2788
90. Dunham, W.H. *et al.* (2012) Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* 12, 1576–1590
91. Nogales, E. and Scheres, S.H. (2015) Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Mol. Cell* 58, 677–689
92. von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403
93. Huang, H. *et al.* (2007) Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* 3, e214
94. Ochoa, D. and Pazos, F. (2010) Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics* 26, 1370–1371
95. Xue, L.C. *et al.* (2011) HomPPI: a class of sequence homology based protein–protein interface prediction methods. *BMC Bioinforma.* 12, 244
96. Vakser, I.A. (2014) Protein–protein docking: from interaction to interactome. *Biophys. J.* 107, 1785–1793
97. Aggarwal, S. and Yadav, A.K. (2016) False discovery rate estimation in proteomics. *Methods Mol. Biol.* 1362, 119–128
98. Yan, Y. and Huang, S.Y. (2020) Modeling protein–protein or protein–DNA/RNA complexes using the HDOCK webserver. *Methods Mol. Biol.* 2165, 217–229
99. Vreven, T. *et al.* (2015) Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.* 427, 3031–3041
100. Kotthoff, I. *et al.* (2022) Dockground scoring benchmarks for protein docking. *Proteins* 90, 1259–1266