



Computational tools to predict context-specific protein complexes

Attila Csikász-Nagy^{1,2}, Erzsébet Fichó¹, Santiago Noto^{1,3} and István Reguly^{1,2}

Abstract

Interactions between thousands of proteins define cells' protein–protein interaction (PPI) network. Some of these interactions lead to the formation of protein complexes. It is challenging to identify a protein complex in a haystack of protein–protein interactions, and it is even more difficult to predict all protein complexes of the complexome. Simulations and machine learning approaches try to crack these problems by looking at the PPI network or predicted protein structures. Clustering of PPI networks led to the first protein complex predictions, while most recently, atomistic models of protein complexes and deep-learning-based structure prediction methods have also emerged. The simulation of PPI level interactions even enables the quantitative prediction of protein complexes. These methods, the required data sources, and their potential future developments are discussed in this review.

Addresses

¹ Cytocast Hungary Kft, Budapest, Hungary

² Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary

³ Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, Brazil

Corresponding author: Csikász-Nagy, Attila (csikasz-nagy.attila@itk.ppke.hu)

Current Opinion in Structural Biology 2024, 88:102883

This review comes from a themed issue on **Biophysical Methods (2024)**

Edited by **Gabriela Chiosis** and **Elizabeth A. Komives**

For complete overview of the section, please refer the article collection - [Biophysical Methods \(2024\)](#)

Available online 9 July 2024

<https://doi.org/10.1016/j.sbi.2024.102883>

0959-440X/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Introduction

Proteins rarely have catalytic activities on their own. They form complexes to perform their biological functions [1,2]. Protein–protein interactions can lead to stable protein complexes, while enzymatic interactions

can last for short times, thus making them more challenging to detect. Several experimental methods have been developed to identify protein–protein interactions (PPIs) and constituting proteins of protein complexes [3]. Protein complexes can be identified by MS Tap-tag, but that gives info only on the proteins that appear together in a larger complex, it is unclear which protein binds to which. Pairwise interactions can be detected by yeast two-hybrid [4], surface plasmon resonance, bio-film/bilayer interference, microcalorimetry, pull-down assays, and structure analysis methods like NMR spectroscopy. These methods require purified proteins, which limit their capabilities. In-vivo interaction detections can be performed under a microscope by duolink proximity ligation assay, immunofluorescence detection, or FRET. Combining these essays has dramatically increased our knowledge and filled several databases. Although we can produce more and more data on protein interactions, we still need to rely on computational tools to get a more thorough picture of protein complexes [5].

Interactions between proteins with known structures can be predicted by molecular docking simulations, and with the appearance of deep learning algorithms to predict 3D structures from amino acid sequences, a new wave of techniques emerged to predict which proteins are present in a protein complex [6]. Machine learning algorithms are also used to find dense regions of protein–protein interaction networks, originally built based on experimental data described above [7]. Various clustering methods can identify protein complexes in these networks, and more sophisticated versions can even distinguish between core complexes and modules that can move between complexes [8]. Large-scale rule-based simulations of protein–protein interactions with minimal structural information can also predict protein complex compositions and, in this case, the abundance and alternative forms of complexes [9]. The final goal of all these prediction methods is to understand how the biological functions of cells are achieved through the interaction of proteins. Once we gain this knowledge, the next step is to reveal how these interactions are perturbed upon disease conditions (including mutations) and how drugs affect protein complex functions by perturbing their compositions and abundances.

Here, we review the computational methods used to predict compositions of protein complexes. Some of these methods go even further and through proteome-wide approaches predict the complexome (all protein complexes in a cell). In this review we give an outlook on how these tools will serve us in disease detection and treatments. We discuss methods that start from protein structural data or predictions of structures, and we also touch upon complex prediction methods that neglect the structure of proteins and rely on large-scale PPI data. Another way to group these computational approaches is to look at the mathematical methods used for predictions. Simulations and machine learning methods are used for structure- and PPI-based approaches. Before going through the four combinations of these, we review the data sources which serve these computational tools.

Data sources

Protein–protein databases frequently collect data through curation efforts or incorporating experimentally validated PPIs from various other data sources. Additionally, some resources extend their datasets using computational prediction (e.g., STRING). Moreover, there are several organism-specific databases available (e.g., *E. coli* – SGD [10], *Drosophila* – FlyBase [11]). Most protein complex prediction algorithms require either only a PPI network or both PPI networks and some biological information as input. The most often used initial PPI data sources for protein complex prediction and curated “gold standard” databases of protein complexes are collected in Table 1.

Predicting protein complexes

Protein complexes are assembled through interactions between binding domains of proteins. To understand the key processes of a cell, it is vital to reveal the detailed interactions in each protein complexes. We can use machine learning algorithms to predict the internal details of complexes [22] or simulate the emergence of complexes using various mathematical methods [23,24]. Both approaches can be started from initial data on protein structures or PPI information. Below, we review methods and tools from all four groups: ML or simulation-based methods relying on structure or PPI data (Table 2). We describe how these methods are available; whether they can be downloaded (all associated data and scripts), available on a webserver (dedicated or Google Colab), and if they require registration or are open to anyone.

Simulations of protein complex formation based on structural information

Docking simulations have traditionally been the go-to tools for modelling protein–protein interactions and recently advanced tools following this concept emerged to predict how protein complexes are assembled. These approaches consider the underlying physics and, therefore, are very expensive computationally – but before the AI revolution, these were the most widespread tools. Some of the most prominent models include ZDOCK [40,41], HADDOCK [26], LightDock [42], and ClusPro [43] that simulate all-atom-level models for proteins and their complexes [44]. These

Table 1			
A list of the most commonly used protein–protein interaction databases as input for predictions and curated databases of protein complexes used to validate predictions.			
Datasource	URL	Description	Number of species
DIP [12]	https://dip.doe-mbi.ucla.edu/dip/Main.cgi	Experimentally determined interactions between proteins	10 model organisms (including <i>H. sapiens</i> , <i>E. coli</i>)
BioGRID [13]	https://thebiogrid.org/	Curated protein–protein and genetic interactions	24 main model organisms
HINT [14]	http://hint.yulab.org/	Manually curated interactome dataset	12 main model organisms
HIPPIE [15]	https://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/	Functionally annotated human interactome	<i>H. sapiens</i>
IntAct [16]	https://www.ebi.ac.uk/intact/home	Database of molecular interactions	All model organisms
MINT [17]	https://mint.bio.uniroma2.it/	Annotated PPI dataset	672 organisms
STRING [18]	https://string-db.org/	Database of known and predicted PPIs (with confidence scores)	All model organisms
ComplexPortal [19]	https://www.ebi.ac.uk/complexportal/home	Manually curated, encyclopaedic resource of macromolecular complexes	28 main model organisms
CORUM [20]	https://mips.helmholtz-muenchen.de/corum/	Manually annotated protein complexes from mammalian organisms	Mammalian model organisms (<i>H. sapiens</i> , <i>M. musculus</i> etc.)
hu.MAP [21]	http://humap2.proteincomplexes.org/	Comprehensive database of protein complexes	<i>H. sapiens</i>
The last three rows are marked with grey as these databases are considered “gold standards” for comparing protein complex predictions.			

Table 2

List of available computation tools to predict protein complex compositions.

Tool name	URL	Type	Feature
Complexes++ [25]	https://github.com/bio-phys/complexespp	Simulation/Structure Downloadable	Study large-scale protein complexes and their solutions
HADDOCK [26]	https://wenmr.science.uu.nl/haddock2.4/	Simulation/Structure Downloadable and webserver (registration required)	Information-driven docking for biomolecular complexes
AlphaFold Multimer [6]	https://github.com/google-deepmind/alphafold	ML/Structure Downloadable and webserver (open, Google Colab notebook)	Gives structural interaction details
AlphaFold 3 [27]	https://alphafoldserver.com/	ML/Structure Webserver (registration required)	The largest scale structural interaction predicting tool
ColabFold [28]	https://github.com/sokrypton/ColabFold	ML/Structure Downloadable and webserver (open)	Protein folding and structures
CombFold [29]	https://codeocean.com/capsule/8791899/tree/v1	ML/Structure Downloadable and webserver (registration required)	Predicting large protein and heteromeric complexes
MoLPC [30]	https://github.com/patrickbryant1/MoLPC	ML/Structure Downloadable and webserver (open)	Predicting large protein complexes
deepAssembly [31–33]	http://zhanglab-bioinf.com/DeepAssembly/	ML/Structure Webserver (open)	Deep learning approach for protein complex prediction
ClusterOne [7]	https://paccanarolab.org/clusterone/	ML/PPI Downloadable	Detecting overlapping complexes ^{Come}
HiSCF [34]	https://github.com/allenv5/HiSCF	ML/PPI Downloadable	Clustering of motifs on the PPI ^{Come}
MCL [35]	https://www.micans.org/mcl/	ML/PPI Downloadable	Clustering on a PPI ^{Come}
PcGAN [36]	https://github.com/yul-pan/PCGAN	ML/PPI Downloadable	Generative model for protein complexes
PCprophet [37]	https://github.com/anfoss/PCprophet	ML/PPI Downloadable	Protein complexes and characterize their changes across experimental conditions
Super.Complex [38]	https://sites.google.com/view/supercomplex/super-complex-v3-0	ML/PPI Downloadable	AutoML for protein complex prediction. Human - SARS-CoV-19 complexes
TADW-SC [39]	https://github.com/kamalberahmand/TADW-SC	ML/PPI Downloadable	Integration of protein features
Cytocast	https://www.cytocast.com/	Simulation/PPI Webserver (registration required)	Overlapping complexes, quantitative predictions ^{Come}

^{Come} – the tool can predict all complexes of the complexome.

approaches use three-dimensional structural information for modeling and scoring, and some further integrate evolutionary information and even experimental data [44], leading to successful CAPRI models [45]. CASP-CAPRI challenges have been an excellent indicator of how the community approaches the 3D structural modeling of proteins and multimers. The use of AI has become prevalent in the field, with CASP14-CAPRI [46] already including numerous such solutions, and CASP15-CAPRI [47] overwhelmingly relying on deep learning methods that were taught how to map amino acid sequences to 3D structural information, and then how to integrate available data on protein docking.

Predicting protein complexes based on structural information

The success of deep learning approaches, like AlphaFold [6] or RoseTTAFold [48], in predicting the 3D structure of monomer proteins from amino-acid sequences in the CASP14 propelled research in a new direction for predicting protein complexes. In some cases, protein complex structures could be obtained by concatenating multiple protein sequences by a gap character and using it as an input of AlphaFold [49,50] which led to the development of the protein complex predicting tool, AlphaFold-Multimer (AFM) [6]. AFM takes AlphaFold structure prediction capabilities a step forward by modifying the architecture and loss functions to

consider multimers to predict the structure of protein complexes. AFM uses multimer chains of known stoichiometry as input, along with multiple pairwise sequence alignments, rather than the simple concatenation of sequences. With these modifications, they were able to predict the structure of heteromeric and homomeric interfaces more accurately.

While AFM achieves state-of-the-art performance predicting protein complexes, it has high memory requirements and takes a long time to run. Considerable effort has been dedicated to make structure prediction more accessible for further research. For example, an accessible Python notebook approach to use AFM was later built by ColabFold [28]. ColabFold makes use of MMSeq2 to speed up search and prediction of complex structures [51]. AF2Complex [52] proposes an approach that adopts AlphaFold and AFM while avoiding pairwise MSA, obtaining highly accurate predictions. AFM accuracy also degrades with the number of chains (>9). To address this, MoLPC [30] proposes to model protein complexes as combinations of subcomponents. It uses AFM or FoldDock [49] to produce dimeric and trimeric subcomponents and uses a Monte Carlo tree search to assemble complexes. This approach works well for homomeric assemblies. CombFold proposes a similar approach with improved performance for heteromeric assemblies. In CombFold, the subcomponents are merged using a combinatorial and hierarchical method instead of the Monte Carlo tree search. Deep learning methods can sometimes introduce artifacts into the generated predictions. Recently, Hou et al. [53] showed that AFM produces complexes with topological links that are unlikely to occur in nature and the authors also propose [28] a way to detect these links.

Other deep learning approaches have been proposed to predict protein complex structures, such as SADA [54] or DeepAssembly, with a performance comparable to AFM. DeepAssembly uses inter-domain interactions predicted by deep learning methods to infer complex structures. Interchain contacts can also be used to predict protein complex structure [55–57]. DeepHomo 2.0 is able to predict both interchain contacts and complexes, even outperforming AFM when considering larger lists of predicted complexes [58]. Protein language models are widely used to learn representations for protein sequences, and have also been successfully used for protein complex prediction, outperforming AFM in certain tasks. For example, ESMPair [59] is able to outperform AFM, more noticeably for structures predicted with low confidence and for complexes in eukaryotes [60,61]. Deep learning has also been used for protein–protein docking. For example, GeoDock provides a fast and accessible framework for predicting complex structure using only the sequence and structure of the docking partners, and was shown to outperform several deep learning

approaches, although falls behind AFM. These are only a few of many different deep learning approaches developed in recent years. For a more comprehensive review on deep learning approaches we refer to a review by Lin et al. [62].

Identification of protein complexes based on PPI networks

Protein complexes are assembled by binding reactions of multiple proteins through a few interaction sites. Large-scale experimental methods (discussed in the introduction) can identify which molecules can interact with each other in a complex, and this information can be used to predict protein complexes, even without considering the actual structure of molecules and atomistic details of protein interactions.

A large set of protein–protein binding reactions are often modeled using a PPI Network. In this graph, the nodes represent proteins, and a link exists between a pair of proteins if they are known to interact. These links can be weighted or unweighted, depending on the available information about the interactions. In these networks, proteins belonging to the same complex tend to be highly interconnected and well separated from the rest, allowing complexes to be detected with clustering methods. The most popular clustering methods for protein complex prediction are MCL and ClusterOne. Such clustering of a PPI network gives a prediction for all the proteins of a cell, so these methods predict the whole complexome.

MCL, or Markov Chain Linking, finds dense, separated subgraphs by performing a random walk in the PPI Network. However, it is not able to distinguish overlapping clusters, failing to assign a protein to more than one complex at a time. ClusterOne was proposed to detect these overlapping clusters in a PPI Network. ClusterOne starts with a ‘seed’ protein and greedily expands the cluster by sampling neighbor proteins from the PPI Network until certain criteria are satisfied and a complex is formed. Other types of clustering have also been used [63], but both MCL and ClusterOne are still preferred by recent protein complex prediction methods [34,64].

Approaches different than clustering have also been used to leverage the PPI Network and predict protein complexes. For example, matrix factorization variations have been shown to be predictive of protein complexes from the PPI Network [65]. SuperComplex [38] used an AutoML approach to predict protein complexes from the PPI Network. SuperComplex compared several combinations of classical pre-processing tools, feature selection tools, and machine learning models such as Decision Trees, Naive Bayes, XGBoost, K-nearest Neighbours, SVM or logistic regression for protein complex prediction. SuperComplex was also used to

predict complexes generated in the interaction between human proteins and the virus SARS-CoV-19 proteins.

PcGAN [36] is a Generative Adversarial Network (GAN) proposed for protein complex prediction. PCGAN learns two models: a generator, which learns to generate complexes, and a discriminator, which learns to distinguish the generated complexes from real complexes. These are trained together until the discriminator is not able to distinguish real from generated complexes. This framing allows the generation of novel, previously unseen protein complexes with possible biological significance that could increase the knowledge about protein complexes.

Protein complexes can also be predicted directly from proteomics data, bypassing the building of a network. For example, PCProphet [37] uses cofractionation MS and complex abundance data to directly infer protein complexes. Expanding on this idea, O'Reilly *et al.* [66] predicts complexes using crosslinking MS in combination with PCProphet to obtain PPI candidates that are fed to AFM.

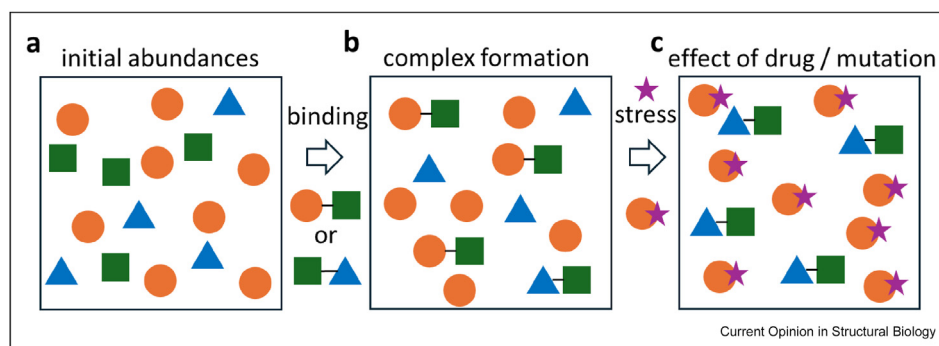
Simulations of protein complex formation based on PPI networks

Simulations of regulatory network dynamics enable researchers to study how enzymatic and other regulatory interactions drive cellular decision-making [67]. Similar approaches can be used to simulate protein complex formation on the PPI network scale. Classical mathematical models with ordinary differential equations can capture only the dynamics with proper rate constants of binding and unbinding events, and even in this case, it could give a steady state distribution of molecules between complexes. The main challenge is to model each molecular interaction, and follow each individual

protein, as they dynamically form protein complexes. The rule-based modeling framework [58] is inspired by process algebra and can be used to simulate the dynamic interactions of individual proteins [68,69]. This concept was used by Rizzetto *et al.* [9] to predict how individual proteins can assemble into protein complexes. This method takes into consideration each protein, together with its binding domains, and simulates how domain–domain interactions [61] between multiple proteins lead to the formation of protein complexes (Fig. 1 a,b). Since the abundance of each protein is considered in simulations, the method enables both qualitative and quantitative predictions of protein complexes [9]. Later, they have shown that the method can be used to predict how protein complex abundances vary between different human tissue types where protein expression levels differ [70]. This methodology enables the simulation of the effects of a drug or mutation that interferes with the binding capabilities of a given protein. The changes caused by such perturbations on a given protein could affect the abundance of protein complexes that do not contain this protein, but are a potential binding partner (Fig. 1 b,c). Such systemic effects of local perturbations can lead to a global change in the complexome, which can have major physiological outcomes.

The challenge with this method is that it requires proteome-wide precise interaction data, information on binding domains, and high computing capacity to simulate millions of interactions between molecules [24]. On a lower scale, the method was also used to predict the compositions of protein complexes at the post-synaptic density of neurons from a limited set of available proteins [71]. This work was later followed by an analysis of how mutations affect these complexes [72].

Figure 1



Simulations of protein complex formation on the PPI level. An initial set of proteins (a) following some PPI database-defined binding rules can form protein complexes (b), where an abundance of complexes can also be detected. Stress in the form of a drug or a mutation (c), which interferes with the binding capabilities of one of the proteins, can lead to a systemic effect on protein complexes and change the abundances of non-directly affected complexes.

Conclusions

The ultimate goal of predicting protein complexes is to understand how they regulate the physiological behavior of cells and how we can control it to avoid or treat diseases. There is high potential in designing drugs that directly target PPIs and protein complex formation [73–75]. As a next step, we need to move away from static pictures of protein complexes and study how proteins dynamically move between complexes [76]. This is particularly important for proteins, which can participate in multiple complexes and take part in many interactions. Binding partners of such proteins compete for a limited pool of proteins; thus, changes in protein abundances can have a major effect on the complexome. Structure-based prediction tools are great for predicting the specific interactions between proteins in a complex, while PPI-based coarse-grained approaches can better scale-up to the complexome level. Yet, from the currently existing protein complex prediction tools, only simulations of PPI data can capture the dynamical abundance changes in complexes [9]. With the growth in computing capacities, scale-up of machine learning-based structural predictions of protein complexes, and large-scale particle-based simulations [69,70] soon we might be able to predict how a single mutation in the binding site of a single protein might affect not only the complexes of this given protein, but the whole complexome. A recently emerging aspect increasing the difficulty of the problem is to account for the phenomenon of protein phase separation, for which specific prediction approaches exist [77] but has yet to be incorporated into the tools discussed here.

Protein–protein interactions can be controlled by posttranslational modifications as well [78]. Various intra- and extracellular signals can induce conformational changes in proteins, which can either enable or disable their binding capabilities. Prediction tools also exist for phosphorylation site identification [79], but we still need a tool to connect these modifications to the binding capabilities of proteins systematically. With the improvements in deep-learning-based structure predictions [76], we are not too far from reaching a stage when we can predict how a signaling pathway affects downstream protein structures, their PPIs and eventually the whole complexome. On smaller scales, we can already see examples moving in this direction [80].

Advanced experimental techniques are also giving us new data on protein complexes. The bloom of cryo-EM [81], proximity ligation assay [82], and other newer approaches give us more details to feed into prediction tools. The challenge is to integrate these data sources and reliably predict complexes and their dynamic changes upon perturbations. Machine learning-based clustering methods improved a lot since their appearance in the field [83]. They are widely spread and easy

to use, but they predict only static protein complexes — with potential overlaps between them [7]. Atomistic-level models are growing from single molecule predictions to protein complex predictions, but we are far from reaching the whole complexome. The combination of deep-learning-based structure predictions with PPI-level simulations might give a new boost to the field when dynamical assemblies and disassemblies of protein complexes and their interacting drugs, lipids, and metabolites can be precisely captured. A further challenge will be to develop experimental tools to validate the predictions of such detailed models.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Attila Csikasz-Nagy reports a relationship with Cytocast Hungary Kft. that includes: board membership, employment, and equity or stocks. Istvan Reguly reports a relationship with Cytocast Hungary Kft. that includes: employment and equity or stocks. Erzsebet Ficho reports a relationship with Cytocast Hungary Kft. that includes: employment and equity or stocks. Santiago Noto reports a relationship with Cytocast Hungary Kft. that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

We are grateful to Alberto Paccanaro and Zoltán Gáspári for critical reading of the manuscript.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- * of special interest
- ** of outstanding interest

1. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, *et al.*: **A census of human soluble protein complexes**. *Cell* 2012, **150**:1068–1081.
2. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, *et al.*: **Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae***. *Nature* 2006, **440**:637–643.
3. Rao VS, Srinivas K, Sujini GN, Kumar GNS: **Protein-protein interaction detection: methods and analysis**. *Int J Proteomics* 2014, **2014**:1–12.
4. Yu H, Braun P, Yöldöröm MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, *et al.*: **High-Quality binary protein interaction map of the yeast interactome network**. *Science* 2008, **322**:104–110.
5. Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Sepasi Tehrani H, Mirzaie M, Fakheri BA, Mohammad-Noori M:

- Protein complex prediction: a survey.** *Genomics* 2020, **112**: 174–183.
6. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Židek A, Bates R, Blackwell S, Yim J, *et al.*: **Protein complex prediction with AlphaFold-Multimer.** *bioRxiv* 2022, <https://doi.org/10.1101/2021.10.04.463034>.
- AlphaFold-Multimer is an extension of AlphaFold2 designed to predict protein complex structures.
7. Nepusz T, Yu H, Paccanaro A: **Detecting overlapping protein complexes in protein-protein interaction networks.** *Nat Methods* 2012, **9**:471–472.
 8. Gavin A-C, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M, *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415**:141–147.
 9. Rizzetto S, Priami C, Csikász-Nagy A: **Qualitative and quantitative protein complex prediction through proteome-wide simulations.** *PLoS Comput Biol* 2015, **11**, e1004424.
 10. Wong ED, Miyasato SR, Aleksander S, Karra K, Nash RS, Skrzypek MS, Weng S, Engel SR, Cherry JM: **Saccharomyces genome database update: server architecture, pan-genome nomenclature, and external resources.** *Genetics* 2023, **224**.
 11. Öztürk-Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, Matthews BB, Millburn G, dos Santos G, Tabone CJ, *et al.*: **FlyBase: updates to the *Drosophila* genes and genomes database.** *Genetics* 2024, <https://doi.org/10.1093/genetics/iyad211>.
 12. Salwinski L: **The database of interacting proteins: 2004 update.** *Nucleic Acids Res* 2004, **32**:449D–451.
 13. Oughtred R, Rust J, Chang C, Breitkreutz B, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, *et al.*: **The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions.** *Protein Sci* 2021, **30**:187–200.
 14. Das J, Yu H: **HINT: high-quality protein interactomes and their applications in understanding human disease.** *BMC Syst Biol* 2012, **6**:92.
 15. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH: **HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks.** *Nucleic Acids Res* 2017, **45**:D408–D414.
 16. del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E, Perfetto L, How K, Ratan P, Shirodkar G, *et al.*: **The IntAct database: efficient access to fine-grained molecular interaction data.** *Nucleic Acids Res* 2022, **50**: D648–D653.
 17. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardoza AP, Santonico E, *et al.*: **MINT, the molecular interaction database: 2012 update.** *Nucleic Acids Res* 2012, **40**:D857–D861.
 18. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, *et al.*: **The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.** *Nucleic Acids Res* 2023, **51**: D638–D646.
- The new update of STRING database incorporates functional enrichment analyses with experimentally validated and predicted protein-protein interactions.
19. Meldal BHM, Perfetto L, Combe C, Lubiana T, Ferreira Cavalcante JV, Bye-A-Jee H, Waagmeester A, del-Toro N, Shrivastava A, Barrera E, *et al.*: **Complex Portal 2022: new curation frontiers.** *Nucleic Acids Res* 2022, **50**:D578–D586.
- The Complex Portal is a manually curated database of macromolecular complexes with known functions from various model organisms. It summarizes complex composition, topology, function, and cross-references to other resources.
20. Tsitsiridis G, Steinkamp R, Giurgiu M, Brauner B, Fobo G, Frishman G, Montrone C, Ruepp A: **CORUM: the comprehensive resource of mammalian protein complexes—2022.** *Nucleic Acids Res* 2023, **51**:D539–D545.
- This is an updated version of the CORUM database, which is one of the largest, manually curated, experimentally validated database of mammalian complexes. It is used as a training set by several complex prediction methods.
21. Drew K, Wallingford JB, Marcotte EM: **hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies.** *Mol Syst Biol* 2021, **17**.
 22. Casadio R, Martelli PL, Savojardo C: **Machine learning solutions for predicting protein-protein interactions.** *WIREs Computational Molecular Science* 2022, **12**.
 23. Shukla R, Tripathi T: **Molecular dynamics simulation of protein and protein-ligand complexes.** In *Computer-aided drug design*. Singapore: Springer; 2020:133–161.
 24. Rizzetto S, Csikász-Nagy A: **Toward large-scale computational prediction of protein complexes.** 2018:271–295.
 25. Linke M, Quoika PK, Bramas B, Köfinger J, Hummer G: **Complexes++: efficient and versatile coarse-grained simulations of protein complexes and their dense solutions.** *J Chem Phys* 2022, **157**.
- Coarse-grained simulations to predict protein complexes.
26. Honorato RV, Koukos PI, Jiménez-García B, Tsaregorodtsev A, Verlati M, Giachetti A, Rosato A, Bonvin AMJJ: **Structural biology in the clouds: the WeNMR-EOSC ecosystem.** *Front Mol Biosci* 2021, **8**.
 27. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, *et al.*: **Accurate structure prediction of biomolecular interactions with AlphaFold 3.** *Nature* 2024, <https://doi.org/10.1038/s41586-024-07487-w>.
- The latest version of AlphaFold (AlphaFold3) is extended with an updated diffusion-based architecture and is capable of predicting the joint structure of complexes, including proteins, nucleic acids, small molecules, ions, and modified residues.
28. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M: **ColabFold: making protein folding accessible to all.** *Nat Methods* 2022, **19**:679–682.
 29. Shor B, Schneidman-Duhovny D: **CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2.** *Nat Methods* 2024, <https://doi.org/10.1038/s41592-024-02174-0>.
 30. Bryant P, Pozzati G, Zhu W, Shenoy A, Kundrotas P, Elofsson A: **Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search.** *Nat Commun* 2022, **13**.
 31. Xia Y, Zhao K, Liu D, Zhou X, Zhang G: **Multi-domain and complex protein structure prediction using inter-domain interactions from deep learning.** *Commun Biol* 2023, **6**:1221.
 32. Zhao K, Xia Y, Zhang F, Zhou X, Li SZ, Zhang G: **Protein structure and folding pathway prediction based on remote homologs recognition using PAtreader.** *Commun Biol* 2023, **6**:243.
 33. Liu D, Zhang B, Liu J, Li H, Song L, Zhang G: **Assessing protein model quality based on deep graph coupled networks using protein language model.** *Briefings Bioinf* 2023, **25**.
 34. Hu L, Zhang J, Pan X, Yan H, You Z-H: **HiSCF: leveraging higher-order structures for clustering analysis in biological networks.** *Bioinformatics* 2021, **37**:542–550.
 35. Van Dongen S: **Graph clustering via a discrete uncoupling process.** *SIAM J Matrix Anal Appl* 2008, **30**:121–141.
 36. Pan Y, Wang Y, Guan J, Zhou S: **PCGAN: a generative approach for protein complex identification from protein interaction networks.** *Bioinformatics* 2023, **39**, btad473.
- A new method using a generative adversarial network to predict associations between proteins.
37. Fossati A, Li C, Uliana F, Wendt F, Frommelt F, Sykacek P, Heusel M, Hallal M, Bludau I, Capraz T, *et al.*: **PCprophet: a framework for protein complex prediction and differential analysis using proteomic data.** *Nat Methods* 2021, **18**: 520–527.

38. Palukuri MV, Marcotte EM: **SuperComplex: a supervised machine learning pipeline for molecular complex detection in protein-interaction networks.** *PLoS One* 2021, **16**, e0262056.
39. Berahmand K, Nasiri E, Pir mohammadiani R, Li Y: **Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding.** *Comput Biol Med* 2021, **138**, 104933.
40. Pierce BG, Hourai Y, Weng Z: **Accelerating protein docking in ZDOCK using an advanced 3D convolution library.** *PLoS One* 2011, **6**, e24657.
41. Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z: **ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers.** *Bioinformatics* 2014, **30**:1771–1773.
42. Jiménez-García B, Roel-Touris J, Barradas-Bautista D: **The LightDock server: artificial intelligence-powered modeling of macromolecular interactions.** *Nucleic Acids Res* 2023, **51**: W298–W304.
43. Desta IT, Porter KA, Xia B, Kozakov D, Vajda S: **Performance and its limits in rigid body protein-protein docking.** *Structure* 2020, **28**:1071–1081.e3.
44. * van Noort CW, Honorato RV: **Bonvin AMJJ: information-driven modeling of biomolecular complexes.** *Curr Opin Struct Biol* 2021, **70**:70–77.

This review summarizes recent software developments and focuses on applications for modeling antibody–antigen complexes and membrane protein complexes.

45. Lensink MF, Nadzirin N, Velankar S, Wodak SJ: **Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition.** *Proteins: Struct, Funct, Bioinf* 2020, **88**:916–938.
46. Lensink MF, Brysbaert G, Mauri T, Nadzirin N, Velankar S, Chaleil RAG, Clarence T, Bates PA, Kong R, Liu B, *et al.*: **Prediction of protein assemblies, the next frontier: the CASP14-CAPRI experiment.** *Proteins: Struct, Funct, Bioinf* 2021, **89**: 1800–1823.
47. ** Lensink MF, Brysbaert G, Raouraoua N, Bates PA, Giulini M, Honorato RV, van Noort C, Teixeira JMC, Bonvin AMJJ, Kong R, *et al.*: **Impact of AlphaFold on structure prediction of protein complexes: the CASP15-CAPRI experiment.** *Proteins: Struct, Funct, Bioinf* 2023, **91**:1658–1683.

The latest review of prediction performances highlights the usage and advantages of AlphaFold and AlphaFold2-Multimer in different prediction methods.

48. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, *et al.*: **Accurate prediction of protein structures and interactions using a three-track neural network.** *Science* 2021, **373**:871–876.
49. ** Bryant P, Pozzati G, Elofsson A: **Improved prediction of protein-protein interactions using AlphaFold2.** *Nat Commun* 2022, **13**:1265.

AlphaFold2 can improve the accuracy of protein–protein interaction predictions, and it can be used to model dimers with increased efficiency.

50. ** Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR, *et al.*: **Computed structures of core eukaryotic protein complexes.** *Science* 2021, **374**, eabm4805.

This study used a combination of RoseTTAFold and AlphaFold to screen yeast proteins through paired multiple sequence alignments. It highlights how a monomeric protein can scale up to protein complexes using deep learning methods.

51. Steinegger M, Söding J: **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.** *Nat Biotechnol* 2017, **35**:1026–1028.
52. Gao M, Nakajima An D, Parks JM, Skolnick J: **AF2Complex predicts direct physical interactions in multimeric proteins with deep learning.** *Nat Commun* 2022, **13**:1744.
53. * Hou Y, Xie T, He L, Tao L, Huang J: **Topological links in predicted protein complex structures reveal limitations of AlphaFold.** *Commun Biol* 2023, **6**:1098.

AlphaFold-Multimer has limitations in certain structure types; exceptionally, their confidence score tends to be lower in topological linker regions. Understanding these features to evaluate prediction methods from a different perspective is also possible.

54. Peng C-X, Zhou X-G, Xia Y-H, Liu J, Hou M-H, Zhang G-J: **Structural analogue-based protein structure domain assembly assisted by deep learning.** *Bioinformatics* 2022, **38**: 4513–4521.
55. Yan Y, Huang S-Y: **Accurate prediction of inter-protein residue–residue contacts for homo-oligomeric protein complexes.** *Briefings Bioinf* 2021, **22**.
56. Baek M, Anishchenko I, Park H, Humphreys IR, Baker D: **Protein oligomer modeling guided by predicted interchain contacts in CASP14.** *Proteins: Struct, Funct, Bioinf* 2021, **89**:1824–1833.
57. Xie Z, Xu J: **Deep graph learning of inter-protein contacts.** *Bioinformatics* 2022, **38**:947–953.
58. Lin P, Yan Y, Huang S-Y: **DeepHomo2.0: improved protein–protein contact prediction of homodimers by transformer-enhanced deep learning.** *Briefings Bioinf* 2023, **24**.
59. Chen B, Xie Z, Qiu J, Ye Z, Xu J, Tang J: **Improved the heterodimer protein complex prediction with protein language models.** *Briefings Bioinf* 2023, **24**.
60. Lupo U, Sgarbossa D, Bitbol A-F: **Protein language models trained on multiple sequence alignments learn phylogenetic relationships.** *Nat Commun* 2022, **13**:6298.
61. Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, Rochereau C, Ahdritz G, Zhang J, Church GM, *et al.*: **Single-sequence protein structure prediction using a language model and deep learning.** *Nat Biotechnol* 2022, **40**:1617–1623.
62. Lin P, Li H, Huang S-Y: **Deep learning in modeling protein complex structures: from contact prediction to end-to-end approaches.** *Curr Opin Struct Biol* 2024, **85**, 102789.
63. Berahmand K, Nasiri E, Pir mohammadiani R, Li Y: **Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding.** *Comput Biol Med* 2021, **138**, 104933.
64. Hu LZ, Goebels F, Tan JH, Wolf E, Kuzmanov U, Wan C, Phanse S, Xu C, Schertzberg M, Fraser AG, *et al.*: **EPIC: software toolkit for elution profile-based inference of protein complexes.** *Nat Methods* 2019, **16**:737–742.
65. Liu G, Liu B, Li A, Wang X, Yu J, Zhou X: **Identifying protein complexes with clear module structure using pairwise constraints in protein interaction networks.** *Front Genet* 2021, **12**.
66. * O'Reilly FJ, Graziadei A, Forbrig C, Bremenkamp R, Charles K, Lenz S, Elfmann C, Fischer L, Stülke J, Rappsilber J: **Protein complexes in cells by AI-assisted structural proteomics.** *Mol Syst Biol* 2023, **19**, e11544.

This approach reveals PPIs inside intact cells, provides structural insight into their interaction interfaces, identifies biologically relevant interactions, and applies to genetically intractable organisms, including pathogenic bacteria.

67. Tyson JJ, Csikasz-Nagy A, Gonze D, Kim JK, Santos S, Wolf J: **Time-keeping and decision-making in living cells: Part I.** *Interface Focus* 2022, **12**.
68. Priami C, Regev A, Shapiro E, Silverman W: **Application of a stochastic name-passing calculus to representation and simulation of molecular processes.** *Inf Process Lett* 2001, **80**: 25–31.
69. Regev A, Silverman W, Shapiro E: **Representation and simulation of biochemical processes using the π -calculus process algebra.** In *Biocomputing 2001*. World Scientific; 2000: 459–470.
70. Rizzetto S, Moyseos P, Baldacci B, Priami C, Csikász-Nagy A: **Context-dependent prediction of protein complexes by SiComPre.** *NPJ Syst Biol Appl* 2018, **4**:37.
71. * Miski M, Keömley-Horváth BM, Rákóczi Megyeriné D, Csikász-Nagy A, Gáspári Z: **Diversity of synaptic protein complexes as a function of the abundance of their constituent proteins: a modeling approach.** *PLoS Comput Biol* 2022, **18**, e1009758.

This work aims to characterize protein complexes in the postsynaptic density using a large-scale simulation by the Gillespie algorithm.

72. Miski M, Weber Á, Fekete-Molnár K, Keömley-Horváth BM, Csikász-Nagy A, Gáspári Z: **Simulated complexes formed from a set of postsynaptic proteins suggest a localised effect of a hypomorphic Shank mutation.** *bioRxiv* 2023, <https://doi.org/10.1101/2023.10.16.562557>.
73. Murakami Y, Tripathi LP, Prathipati P, Mizuguchi K: **Network analysis and in silico prediction of protein–protein interactions with applications in drug discovery.** *Curr Opin Struct Biol* 2017, **44**:134–142.
74. Cao L, Coventry B, Goreshnik I, Huang B, Sheffler W, Park JS, Jude KM, Marković I, Kadam RU, Verschuereen KHG, *et al.*: **Design of protein-binding proteins from the target structure alone.** *Nature* 2022, **605**:551–560.
75. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, *et al.*: **De novo design of protein structure and function with RFdiffusion.** *Nature* 2023, **620**:1089–1100.
76. Przytycka TM, Singh M, Slonim DK: **Toward the dynamic interactome: it's about time.** *Briefings Bioinf* 2010, **11**:15–29.
77. Vernon RM, Forman-Kay JD: **First-generation predictors of biological protein phase separation.** *Curr Opin Struct Biol* 2019, **58**:88–96.
78. de Lichtenberg U, Jensen LJ, Brunak S, Bork P: **Dynamic complex formation during the yeast cell cycle.** *Science* 2005, **307**:724–727.
79. Chen M, Zhang W, Gou Y, Xu D, Wei Y, Liu D, Han C, Huang X, Li C, Ning W, *et al.*: **Gps 6.0: an updated server for prediction of kinase-specific phosphorylation sites in proteins.** *Nucleic Acids Res* 2023, **51**:W243–W250.
80. Hashimoto Y, Sheng X, Murray-Nerger LA, Cristea IM: **Temporal dynamics of protein complex formation and dissociation during human cytomegalovirus infection.** *Nat Commun* 2020, **11**:806.
81. Earl LA, Falconieri V, Milne JL, Subramaniam S: **Cryo-EM: beyond the microscope.** *Curr Opin Struct Biol* 2017, **46**:71–78.
82. Hegazy M, Cohen-Barak E, Koetsier JL, Najor NA, Arvanitis C, Sprecher E, Green KJ, Godsel LM: **Proximity ligation assay for detecting protein-protein interactions and protein modifications in cells and tissues in situ.** *Curr Protoc Cell Biol* 2020, **89**.
83. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinf* 2003, **4**:2.