CM2D3: Furnishing the Human Interactome with Structural Models of Protein Complexes Derived by Comparative Modeling and Docking

Patricia Mirela Bota [1], Altair C. Hernandez [2], Joan Segura [3], Oriol Gallego [2], Baldo Oliva [1*] and Narcis Fernandez-Fuentes [4*]

*1 - **Structural Bioinformatics Lab (GRIB-IMIM),** Universitat Pompeu Fabra, 08950 Barcelona, Catalonia, Spain*

*2 - **Live-cell Structural Biology,** Department of Medicine and Life Sciences, University Pompeu Fabra, Barcelona 08005, Catalonia, Spain*

*3 - **Research Collaboratory for Structural Bioinformatics Protein Data Bank,** San Diego Supercomputer Center, University of California, La Jolla, CA 92093, USA*

*4 - **Institute of Biological,** Environmental and Rural Sciences, Aberystwyth University, SY233EE Aberystwyth, United Kingdom*

***Correspondence to Baldo Oliva and Narcis Fernandez-Fuentes:** baldo.oliva@upf.edu (B. Oliva), narcis@bioinsilico.org (N. Fernandez-Fuentes) @PatriciaMBota (P. Mirela Bota), @pantair (A.C. Hernandez), @GallegoLab (O. Gallego)*

https://doi.org/10.1016/j.jmb.2023.168055

Edited by Rita Casadio

Abstract

The human interactome is composed of around half a million interactions according to recent estimations and it is only for a small fraction of those that three-dimensional structural information is available. Indeed, the structural coverage of the human interactome is very low and given the complexity and time-consuming requirements of solving protein structures this problem will remain for the foreseeable future. Structural models, or predictions, of protein complexes can provide valuable information when the experimentally determined 3D structures are not available. Here we present CM2D3, a relational database containing structural models of the whole human interactome derived both from comparative modeling and data-driven docking. Starting from a consensus interactome derived from integrating several interactomics databases, a strategy was devised to derive structural models by computational means. Currently, CM2D3 includes 33338 structural models of which 5121 derived from comparative modeling and the remaining from docking. Of the latter, the structures of 14554 complexes were derived from monomers modeled by M4T while the rest were modeled with structures as predicted by AlphaFold2. Lastly, CM2D3 complements existing resources by focusing on models derived from both free-docking, as opposed to template-based docking, and hence expanding the available structural information on protein complexes to the scientific community. **Database URL:** http://www.bioinsilico.org/CM2D3

Background

Past and present large scale, high-throughput, initiatives seek to describe all protein–protein interactions (PPIs) –the *interactome*- that occur in cells in a number of model organisms including human.[1] The justification to such pharaonic endeavors lies in the fact, as comprehensively described in a range of publications, that PPIs underpin all cellular processes both in health and disease.[2–3] Among those initiatives is the HuRi project,[1] which has allowed the charting and identification of a large number of highly reliable binary PPIs. However important these initiatives are, realizing the full

Please cite tts article as: P. Mirela Bota, A.C. Hernandez, J. Segura, et al., CM2D3: Furnishing the Human Interactome with Structural Models of Protein Complexes Derived by Comparative Modeling and Docking Journal of Molecular Biology, https://doi.org/10.1016/j.jmb.2023.168055

potential of interactomes requires their three-dimensional (3D) structural, atomistic, details.[4] Contrary to the revolution experienced in recent years on DNA or RNA sequencing, the structure determination of protein complexes lags far behind the experimental determination of PPIs, resulting in an ever-expanding gap. Indeed, the structural coverage of the human interactome is very low when comparing the number of solved structures of protein complexes and the predicted number of interactions.[5]

Computational protein structural prediction is now a decades old problem, that nonetheless has witnessed a new revival with Artificial Intelligence (AI)-powered methods such AlphaFold and subsequent improved version AlphaFold2 (AF2)[6] or RoseTTAFold[7] having achieved unprecedented accuracy. While such accuracy of such methods has been well established in single proteins, its application to large-scale prediction of protein complexes remains limited,[8,9] although a very recent work using a AF2 inspired methodology has been successfully used to model a larger set of protein complexes.[10] However, comparative modeling and protein docking are still more scalable and can be employed to derive the structure of protein complexes. The goal of these tools is to improve the structural coverage of the human interactome, which in turn will provide useful information to fully understand the function of protein complexes. In the context of databases devoted to computationally predicted structure of PPIs, there are a number of resources available such PrePPI,[11] Interactome3D,[12] GWIDD,[13] INstruct.[14] While these resources contain an important number of interactions, they are based on one or the other approach to derive structural models, i.e. derived from comparative modeling or docking but not both.

To fill this niche, we have developed the CM2D3 database: Comparative Modeling and Data-Driven Docking. CM2D3 compiles structural models of protein complexes derived both by comparative modeling and data-docking using our original methods MODPIN[15] and V^{2}DOCK[16] (that combined VORFFIP[17] and PatchDock[18]). By creating a combined modeling pipeline, CM2D3 can improve the structural coverage of the human interactome. The modeling pipeline is reasonably fast and thus it will keep abreast with ever-increasing interactomics data. It will complement existing resources and thus provide a valuable resource to the scientific community. Finally, CM2D3 features a user-friendly interface and includes a number of tools to facilitate the search and analyses of protein complexes.

Database content, implementation, and use

Sources of experimentally determined PPIs

The starting point of CM2D3 was the integration of several interactomics datasets (Intact,[19] BioGRID[20], HIPPIE,[21] HitPredict[22], DIP[23] and InBioMap[24]) using BIANA.[25] Upon integrating these datasets into a consensus interactome, over 5.0 10^6 interactions were subjected to a highly stringent filtering process devised to identify direct, i.e. physical, interactions between proteins. Firstly, we considered only interactions identified by the following experimental methods: cross-linking, two hybrid array, two hybrid prey pooling, two hybrid, biochemical assay, validated two hybrid, proximity labeling, enzymatic study, two hybrid pooling, protein array or pull-down. Secondly, only interaction pairs identified by at least three of these methods were accepted. This step was included to ensure only direct interactions were considered as, for instance, pull-down experiment identify both direct and indirect interactions. The resulting interactome was composed of 37419 interactions from 11490 proteins that were submitted to the modeling pipeline (see Figure 1).

Modeling pipeline

The modeling pipeline was devised as follows. On a first approach the structure prediction was attempted by comparative modeling using ModPIN.[15] If the modeling was not successful, then data driven docking was used. If not available at the PDB databank,[26] the structure of the individual proteins was derived by comparative modeling using M4T[27] or downloaded from the AlphaFold protein structure database.[28] If the latter, a quality control step was included to eliminate regions of low confidence using the Phenix package.[29] Of the original 37419 interactions, 5121 were solved by comparative modeling and the remaining by data-driven docking. Of the latter, 14554 were derived using M4T to model the monomers and the remaining 13663 were acquired from the AlphaFold protein structure database (see Figure 2). For 4081 pairs it was not possible to obtain a model by neither comparative modeling nor docking and therefore were modelled using AF2. AF2 predictions were performed using the multimer implementation of AF2 version v.2.2.0[9] using as input a FastA file with the sequence of the pair of proteins and the flag *--model_preset=multimer*.

As discussed, CM2D3 contains both models derived by three different approaches: comparative modeling, protein docking and AF2-multimer resulting in 37419 modeled complexes. Comparing to existing resources, Interactome3D[12] contains over 9000 complexes of with around 2000 are complete models; INstruct[14] archives of 6500 and GWIDD[13] a collection of around 2500 models derived from docking. A higher number of proteins complexes was attempted by Elofsson et col[10]: 65,484 from which over 3000 were considered as high-confidence models.

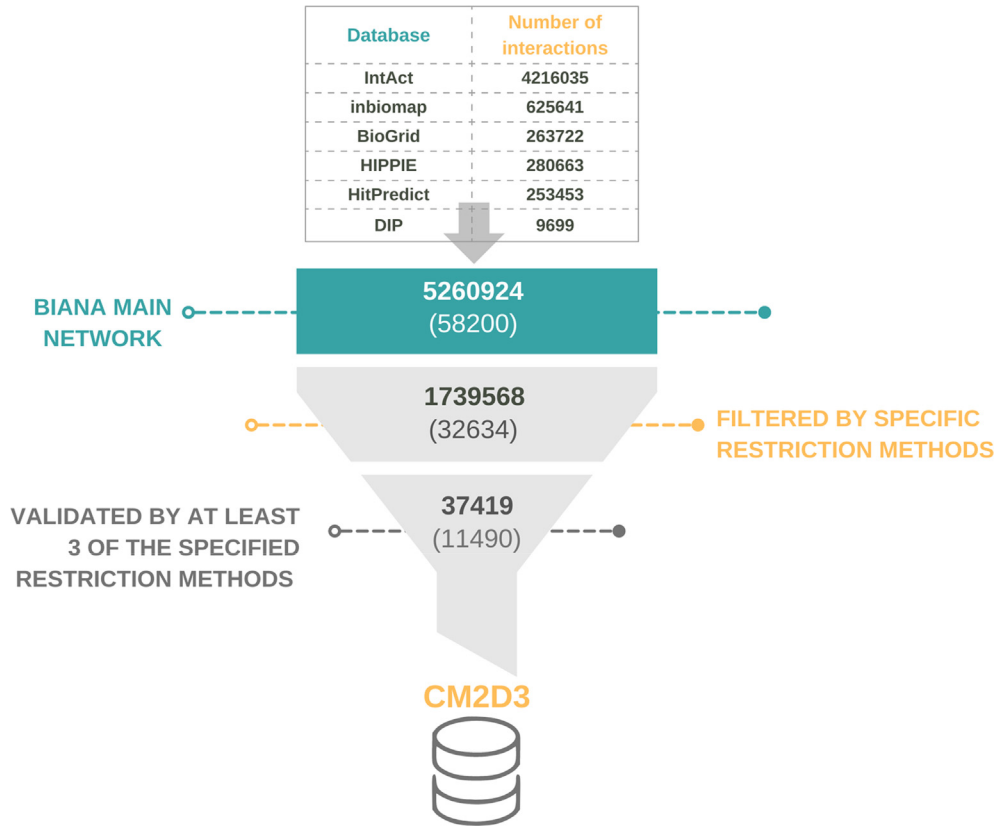

Figure 1. Overview of the integration of the interactomics datasets and filtering process. The number of interactions are shown in each step, with the number of proteins in brackets.

Comparing structural models to state-of-the art predictions

The structural pipeline devised in this work relies on fully-benchmarked methods published before by us, i.e. MODPIN[15] and V^{2}DOCK.[16] Nonetheless and in order to assess the level of which the models we obtain are similar to those obtained by more advanced methodology we performed a limited comparison. A random selection of 50 protein complexes present in CM2D3, 25 derived by comparative modeling (MODPIN) and 25 by docking (V^{2}DOCK), were selected modelled using AF2[6] in 'multimer' mode. The resulting structures of the protein complexes were compared using MM-align[30] and the structure similarity quantified using TM-scores.

As shown in Figure 1 on the supplementary material, the distribution of the best TM-scores is well-above 0.5 in the case of comparisons to models derived by comparative modeling ranging to above 0.9. In the case of docking models, while the TM-score is slightly worse, still the structure similarity between docking- and AF2-derived models is high. It is worth mentioning that while in the case of comparative modeling all models were used in the comparison, in the case of docking only top 50 (among potentially hundreds) were used. It is also important to stress that this comparison was simply to show that the models obtained by 'traditional' approaches, i.e. comparative modeling and docking, resemble to those obtained by more advanced techniques based on AI, i.e. AF2.

Database implementation and interfacing

CM2D3 comprises two major components: a relational database management system for data storage and management and a web application to interface the database. Data is stored in a relational MySQL database whose design was optimized to provide a fast and optimal access to the information. It makes extensive use of master and internal keys and cross-references between tables. The MySQL server runs in a dedicated computer that also mirrors all external databases that are cross-referenced in CM2D3.

The web application runs on an Apache web-server hosted on a CentOS Linux operating system. Python and JavaScript are used to interface and access the database. Web pages resulting from queries are generated dynamically thus ensuring up-to-date information is available to users. As a general rule, table headers, icons, and

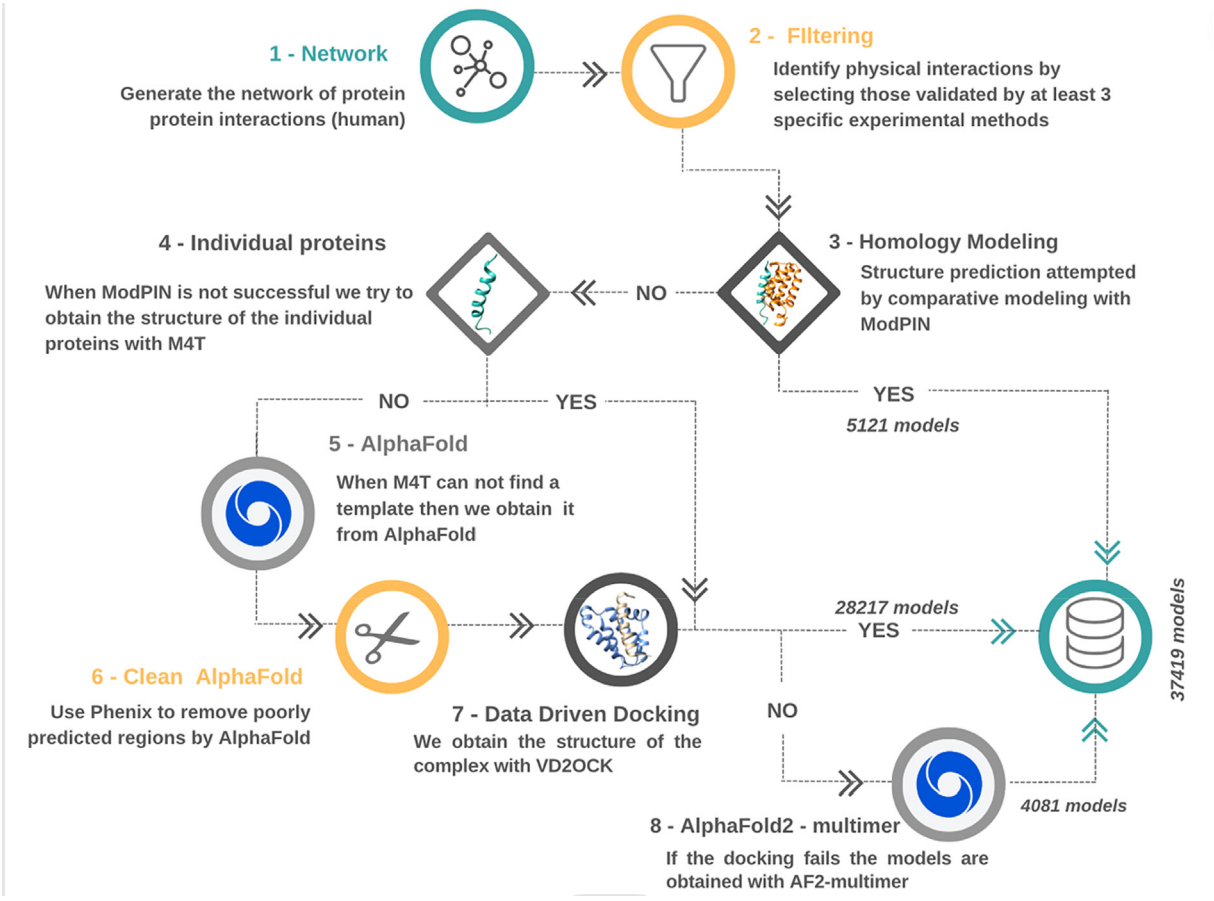

Figure 2. Overview of the modeling pipeline. The flowchart shows the integration and modeling of complexes, including the number of models obtained by each method (comparative modeling, data-driven docking, and AF2-multimer).

other elements shown in web pages are active, i.e. hovering over them will reveal a short help and/or perform a task (e.g. sort a table). Besides, there is documentation and explanations about contents, annotation process, and the use of CM2D3 in the help pages.

Database interface: Retrieving information

The structural models of protein complexes compiled in CM2D3 are interfaced by a user-friendly and intuitive website accessible at https://bioinsilico.org/CM2D3 Users can retrieve the information of protein complexes querying by a list of protein pairs, using the Uniprot entry gene ID or Uniprot primary accession e.g. ARHGB_HUMAN RHOA_HUMAN; O15085 P61586; ARHGEF11 or by a list of proteins, e.g. ASPP1_HUMAN; Q96KQ4. Upon querying either by pairs or single proteins, the server returns a list of structural models that users can browse, including the visualization of the structure of the complex on a Mol* applet.[31] The Mol* applet includes a wide-range set of functionalities that allows an in-depth analysis of the structure as well as the tailoring of visualization (see Figure 3). Finally, users can also download the coordinates of the complexes.

Future perspective and long-term sustainability

As described, the structural information of protein complexes opens a number of avenues in the fields of biology and biomedicine where this information is needed. We envisage that the information compiled in CM2D3 would be of interest to a range of scientists working in the area of PPI for biotechnology (design, functionalization, etc.) as well as biomedicine (therapeutic, impact of mutations, etc.). It is for this reason that future perspectives to this work will include the integration of mutational data that can pinpoint the importance of residues for individual proteins or protein complexes.

The structural modeling pipeline is fully automated, the modeling of protein complexes is done in a high-performance computing cluster, therefore the entire process is relatively fast and

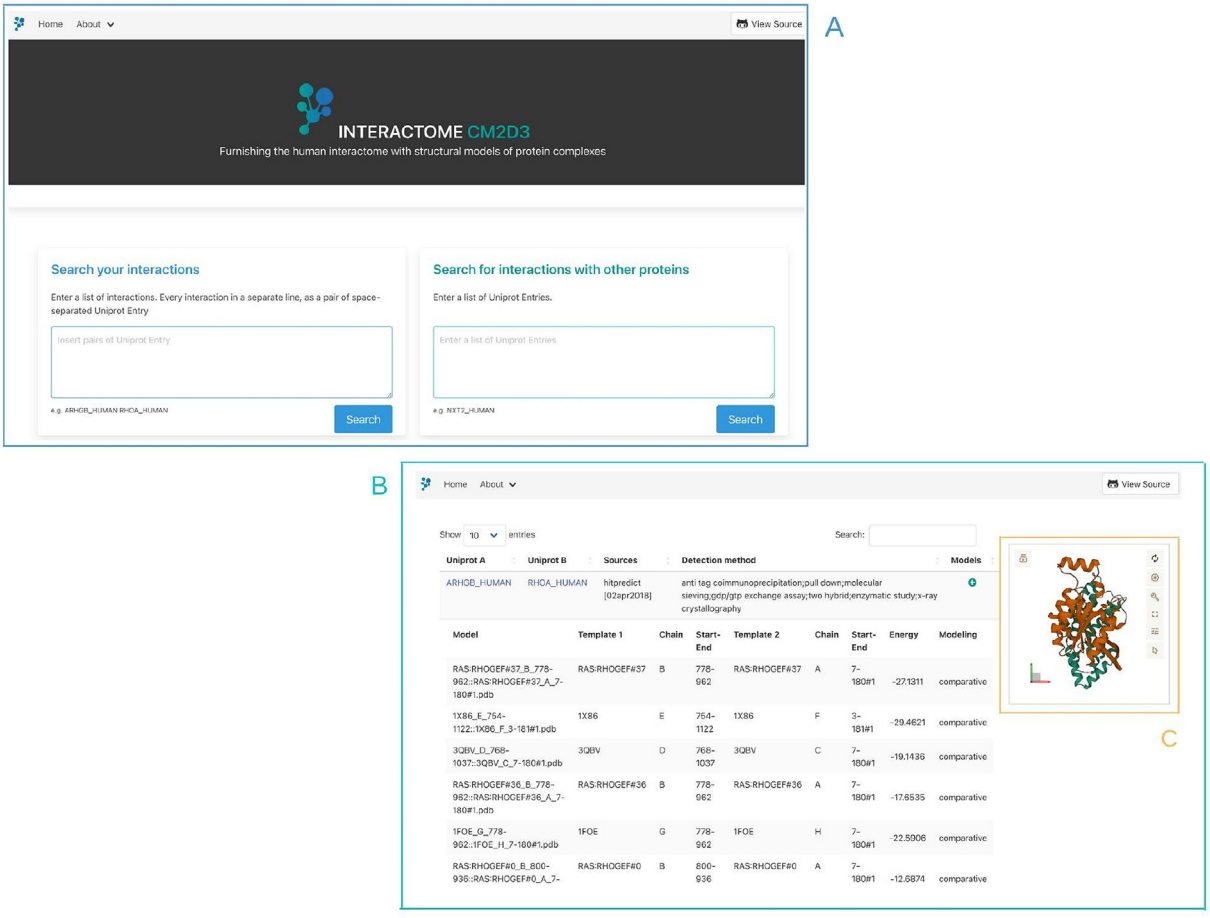

Figure 3. Snapshots of the CM2D3 database. (A) Snapshot of the front page with the two querying options: search by protein-pairs or search by individual proteins. (B) Example of a table of results for a query of a single protein (AKTIP_HUMAN). The table shows the interacting partners (HOOK1 and HOOK2 from human), the provenance of the interactomics datasets and the experimental methods that detected the interactions. (C) Snapshot of the Mol* applet showing the structure of one of the complexes.

requires minimum supervision, thus ensuring long-term sustainability. Upon release of novel interactomics data, this will be downloaded to the local repository and checked against existing data in CM2D3. Novel interactions unavailable in and within the remit of CM2D3 will be submitted to the modeling pipeline, novel structural models will then be uploaded to CM2D3. Up-to-date information about database contents and date of last update is presented in the home page. Besides, the GitHub repository with the scripts and programs used to construct CM2D3 is available (https://github.com/abotlp/InteractomeCM2D3).

Conclusion

A database of structural models of protein complexes: CM2D3, is presented. Starting from several interactomics datasets, the BIANA framework[25] was used to integrate and select direct physical interactions validated by at least three experimental methods. Structural models of complexes of this network were derived by both comparative modeling (using MODPIN) or data driven docking (using V^2DOCK). This wealth of information in CM2D3 is interfaced by a clear and intuitive web interface with a number of functionalities allowing for an easy and convenient access to data. The integration and modeling pipeline is fully automated and can also be downloaded from a GitHub repository to facilitate world wide access and update. CM2D3 complements many other resources available to the scientific community with a number of applications in biology and biomedicine.

Funding

Authors acknowledge support from the Biotechnology and Biological Science Research Council (BBS/E/W/0012843D), PID2020-113203RB-I00 and “Unidad de Excelencia María de Maeztu” (ref: CEX2018-000792-M), funded by

the MCIN and the AEI (https://doi.org/10.13039/501100011033), Spanish funding agency (PID2021-127773NB-I00/MICIN/AEI/https://doi.org/10.13039/501100011033/FEDER, UE), and HFSP (ref: RGP0017/2020). ACH acknowledges support from MINECO (ref: PRE2019-088514) and the Boehringer Ingelheim Fonds.

CRediT authorship contribution statement

Patricia Mirela Bota: Methodology, Investigation, Software, Formal analysis, Data curation, Writing – review & editing. **Altair C. Hernandez:** Data curation, Software, Writing – review & editing. **Joan Segura:** Conceptualization, Writing – review & editing. **Oriol Gallego:** Funding acquisition, Writing – review & editing. **Baldo Oliva:** Conceptualization, Supervision, Formal analysis, Funding acquisition, Project administration, Writing – review & editing. **Narcis Fernandez-Fuentes:** Conceptualization, Supervision, Investigation, Formal analysis, Funding acquisition, Writing – original draft.

DATA AVAILABILITY

All the data is available from webserver or GitHub

Acknowledgements

Authors wish to thank Alfons Gonzalez-Pauner from the GRIB IT support unit for invaluable technical support.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2023.168055.

Received 31 October 2022;
Accepted 16 March 2023;
Available online xxxx

Keywords:
comparative modeling;
protein docking;
human interactome;
database;
structural models;
protein complexes

References

1. Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., et al., (2020). A reference map of the human binary protein interactome. *Nature* **580**, 402–408.
2. Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., et al., (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509.
3. Vidal, M., Cusick, M.E., Barabasi, A.L., (2011). Interactome networks and human disease. *Cell* **144**, 986–998.
4. Mirela-Bota, P., Aguirre-Plans, J., Meseguer, A., Galletti, C., Segura, J., Planas-Iglesias, J., et al., (2021). Galaxy InteractoMIX: An Integrated Computational Platform for the Study of Protein-Protein Interaction Data. *J. Mol. Biol.* **433**, 166656
5. Stumpf, M.P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., et al., (2008). Estimating the size of the human interactome. *PNAS* **105**, 6959–6964.
6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
7. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., et al., (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876.
8. Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., et al., (2021). Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805.
9. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv. 2022:2021.10.04.463034.
10. Burke, D.F., Bryant, P., Barrio-Hernandez, I., Memon, D., Pozzati, G., Shenoy, A., et al., (2023). Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* **30**, 216–225.
11. Zhang, Q.C., Petrey, D., Garzon, J.I., Deng, L., Honig, B., (2013). PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.* **41**, D828–D833.
12. Mosca, R., Ceol, A., Aloy, P., (2013). Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53.
13. Kundrotas, P.J., Zhu, Z., Vakser, I.A., (2010). GWIDD: Genome-wide protein docking database. *Nucleic Acids Res.* **38**, D513–D517.
14. Meyer, M.J., Das, J., Wang, X., Yu, H., (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* **29**, 1577–1579.
15. Meseguer, A., Dominguez, L., Bota, P.M., Aguirre-Plans, J., Bonet, J., Fernandez-Fuentes, N., et al., (2020). Using collections of structural models to predict changes of binding affinity caused by mutations in protein-protein interactions. *Protein Sci.* **29**, 2112–2130.
16. Segura, J., Marin-Lopez, M.A., Jones, P.F., Oliva, B., Fernandez-Fuentes, N., (2015). VORFFIP-driven dock: V-D2OCK, a fast and accurate protein docking strategy. *PLoS One* **10**, e0118107.
17. Segura, J., Jones, P.F., Fernandez-Fuentes, N., (2011). Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinf.* **12**, 352.

18. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J., (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* **33**, W363–W367.
19. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al., (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363.
20. Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., et al., (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* **39**, D698–D704.
21. Alanis-Lobato, G., Andrade-Navarro, M.A., Schaefer, M.H., (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* **45**, D408–D414.
22. Lopez, Y., Nakai, K., Patil, A., (2015). HitPredict version 4: comprehensive reliability scoring of physical protein-protein interactions from more than 100 species. *Database (Oxford)* **2015**
23. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S. M., Eisenberg, D., (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305.
24. Li, T., Wernersson, R., Hansen, R.B., Horn, H., Mercer, J., Slodkowicz, G., et al., (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64.
25. Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J., Oliva, B., (2010). Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinf.* **11**, 56.
26. Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., et al., (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.* **39**, D392–D401.
27. Fernandez-Fuentes, N., Madrid-Aliste, C.J., Rai, B.K., Fajardo, J.E., Fiser, A., (2007). M4T: a comparative protein structure modeling server. *Nucleic Acids Res.* **35**, W363–W368.
28. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al., (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444.
29. Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., et al., (2002). PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1948–1954.
30. Mukherjee, S., Zhang, Y., (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.* **37**, e83.
31. Sehnal, D., Bittrich, S., Deshpande, M., Svobodova, R., Berka, K., Bazgier, V., et al., (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* **49**, W431–W437.