frontiers in Molecular Biosciences

METHODS
published: 23 August 2021
doi: 10.3389/fmolb.2021.716973

DeepComplex: A Web Server of Predicting Protein Complex Structures by Deep Learning Inter-chain Contact Prediction and Distance-Based Modelling

*Farhan Quadir†, Raj S. Roy†, Elham Soltanikazemi† and Jianlin Cheng**

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, United States

Proteins interact to form complexes. Predicting the quaternary structure of protein complexes is useful for protein function analysis, protein engineering, and drug design. However, few user-friendly tools leveraging the latest deep learning technology for inter-chain contact prediction and the distance-based modelling to predict protein quaternary structures are available. To address this gap, we develop DeepComplex, a web server for predicting structures of dimeric protein complexes. It uses deep learning to predict inter-chain contacts in a homodimer or heterodimer. The predicted contacts are then used to construct a quaternary structure of the dimer by the distance-based modelling, which can be interactively viewed and analysed. The web server is freely accessible and requires no registration. It can be easily used by providing a job name and an email address along with the tertiary structure for one chain of a homodimer or two chains of a heterodimer. The output webpage provides the multiple sequence alignment, predicted inter-chain residue-residue contact map, and predicted quaternary structure of the dimer. DeepComplex web server is freely available at http://tulip.rnet.missouri.edu/deepcomplex/web_index.html

Keywords: protein quaternary structure prediction, protein complex structure prediction, protein interaction, deep learning, inter-chain contact prediction, distance-based modeling

OPEN ACCESS

Edited by:
Masahito Ohue,
Tokyo Institute of Technology, Japan

Reviewed by:
Ilpo Vattulainen,
University of Helsinki, Finland
Marc F. Lensink,
Centre National de la Recherche Scientifique (CNRS), France

****Correspondence:***
Jianlin Cheng
chengji@missouri.edu

†These authors share first authorship

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *29 May 2021*
Accepted: *12 August 2021*
Published: *23 August 2021*

Citation:
Quadir F, Roy RS, Soltanikazemi E and Cheng J (2021) DeepComplex: A Web Server of Predicting Protein Complex Structures by Deep Learning Inter-chain Contact Prediction and Distance-Based Modelling. Front. Mol. Biosci. 8:716973. doi: 10.3389/fmolb.2021.716973

INTRODUCTION

Proteins interact to form complexes to perform biological functions like gene regulation, signal transduction and enzymatic catalysis (Szklarczyk et al., 2015; Quadir et al., 2021). High-throughput experimental approaches (e.g., yeast two-hybridization) can figure out whether two proteins form a permanent or transient complex; however, these techniques cannot accurately determine the 3D shape of the complex. Biophysical experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM) can determine where and how the proteins interact. These approaches, however, are expensive and time-consuming, and hence can be only applied to a small number of proteins. Therefore, developing accurate computational approaches to reconstruct the quaternary structures of protein complexes has been an important long-standing challenge (Biasini et al., 2014).

During the last 2 decades, many computational techniques, which are fast and inexpensive, have been developed to generate the quaternary structural models of dimers using the tertiary structures of interacting proteins as input (Smith and Sternberg, 2002; Chen et al., 2003; Gray et al., 2003; Comeau

et al., 2004; Tovchigrechko and Vakser, 2006; Lyskov and Gray, 2008; de Vries et al., 2010; Hwang et al., 2010). Although the classic *ab initio* docking methods have achieved some success for some protein complexes, according to the last several rounds of Critical Assessments of Predictions of Interactions (CAPRI) (Janin, 2005; Janin, 2002), the general accuracy of these approaches is still low (Lensink et al., 2019). Template-based modelling approaches can predict the quaternary structure of some dimers accurately if good structural templates are available (Biasini et al., 2014; Lensink et al., 2019). However, these methods can only be applied to a small portion of proteins, for which the experimental complex structures of interacting homologues (interlogs) are available (Quadir et al., 2021).

Ab initio docking tools perform the energy optimization-based scoring to rank structural models. RosettaDock (Lyskov and Gray, 2008) uses the Monte Carlo approach to dock proteins based on the energy optimization. Some tools like ClusPro (Comeau et al., 2004) and ZDOCK (Pierce et al., 2014) are built upon the Fast Fourier Transformation (FFT) approach to search for the geometric complementarity. Also based on FFT, HDOCK (Yan et al., 2017) can perform both template-based modelling and template free docking between interacting proteins as well as between the interaction of proteins and nucleic acids. A few tools have been able to leverage some inter-chain co-evolutionary information in their protocol. InterEvDock (Yu et al., 2016) performs docking by incorporating co-evolutionary information obtained from the paired multiple sequence alignments (MSAs) of the interlogs, and then selects the top models based on FRODOCK (Garzon et al., 2009) score, SOAP-PP (Dong et al., 2013) score and InterEvScore (Andreani et al., 2013). GREMLIN (Ovchinnikov et al., 2014) and EVcomplex (Hopf et al., 2014; Schelling et al., 2018) can perform the co-evolution based interchain contact prediction using the statistical/mathematical direct coupling analysis (DCA) on the MSA of interlogs. However, due to the limited prediction capability of the statistical/mathematical methods, they can only be applied to a portion of protein complexes with deep MSAs.

Although deep learning methods like convolution neural networks, graph neural networks, residual networks, and transformers, has been used in the computational modelling of protein structures, most focus was put on the development of deep learning methods for intra-chain contact prediction, intra-chain residue-residue distance prediction, quality assessment, and tertiary structure prediction (Zeng et al., 2018; Alquraishi and Valencia, 2019; Baek et al., 2021; Jumper et al., 2021; Quadir et al., 2021; Yan and Huang, 2021). The use of very deep and complex deep learning networks coupled with multiple sequence alignments (e.g., Google DeepMind's AlphaFold), has significantly advanced tertiary structure prediction. Additionally, recent advances in techniques such as cryo-EM in the form of better electron guns, energy filters, cameras, etc. has led to improvements in the atomic resolution of tertiary and quaternary structures of proteins available in the protein data bank (PDB), and hence, has increased the quantity and quality of data available for training, testing and validation of computational prediction of structures of proteins (Nakane et al., 2020), particularly protein complexes. But, the use of deep learning for prediction of quaternary structures of protein complexes has not been well explored, especially when it comes to heteromeric protein complexes.

Recently, ComplexContact (Zeng et al., 2018) web server was developed for inter-chain contact prediction of heterodimers. It used a deep learning network that was pretrained for intra-chain contact prediction to predict inter-chain contacts of heterodimers from the features obtained using a homology-based, genome-based and phylogeny-based multiple sequence alignment. Another deep learning method, DeepHomo (Yan and Huang, 2021), was developed for inter-chain contact prediction of C2 symmetry homodimers. Also, DNCON2_Inter (Quadir et al., 2021) predicted inter-chain contacts of homodimers by removing intra-chain contacts, with some degree of flexibility, from the contacts predicted from the multiple sequence alignment of a monomer. Inspired from the successful performance of AlphaFold2 (Jumper et al., 2021) in CASP14, recently RoseTTAFold (Baek et al., 2021) was developed which performs end-to-end direct prediction of tertiary structure (atomic coordinates) of proteins directly from multiple sequence alignments using three-track attention-based neural networks. Based on the test on a few complexes, the work also shows the deep learning's potential of predicted the quaternary structures of dimers and trimers provided the paired multiple sequence alignment of sufficient depth is available. However, despite the recent exploration, few general tools are available to use the inter-chain contact prediction directly to generate the final quaternary structure of the complexes. Therefore, it is necessary to develop a user-friendly, robust pipeline leveraging the cutting-edge deep learning technology to predict inter-chain contacts and use them with the distance-based modelling to generate high-quality quaternary structures of protein complexes.

Here we introduce DeepComplex, an automated web server for ab initio prediction of protein complex structures. DeepComplex employs deep learning techniques to predict inter-chain residue-residue contacts from protein sequences first (Quadir et al., 2021; Zeng et al., 2018; Quadir et al., 2020). It then utilizes a gradient descent-based optimization method to use the predicted contacts together with physicochemical and geometrical information as restraints to model the quaternary structures of interacting proteins rather accurately (Soltanikazemi et al., 2021). DeepComplex provides an easy and convenient way for users to quickly obtain predicted quaternary structures of both water-soluble and membrane-associated protein dimers of any organisms.

DESIGN, USE AND PERFORMANCE OF DEEPCOMPLEX WEB SERVER

Server Input

The DeepComplex web server prompts a user to provide a handful of required inputs to start the prediction process illustrated in **Figure 1**. Basic inputs like email address and job name are used to identify prediction tasks and send results back to users. Two radio buttons are used for users to choose a prediction type: homodimer or heterodimer. If it is a homodimer, the tertiary structure information of only a single chain in the PDB format needs to be copied into the text box or uploaded as a file. Otherwise, the tertiary structure information of both chains in a heterodimer needs to be provided. Once the job is

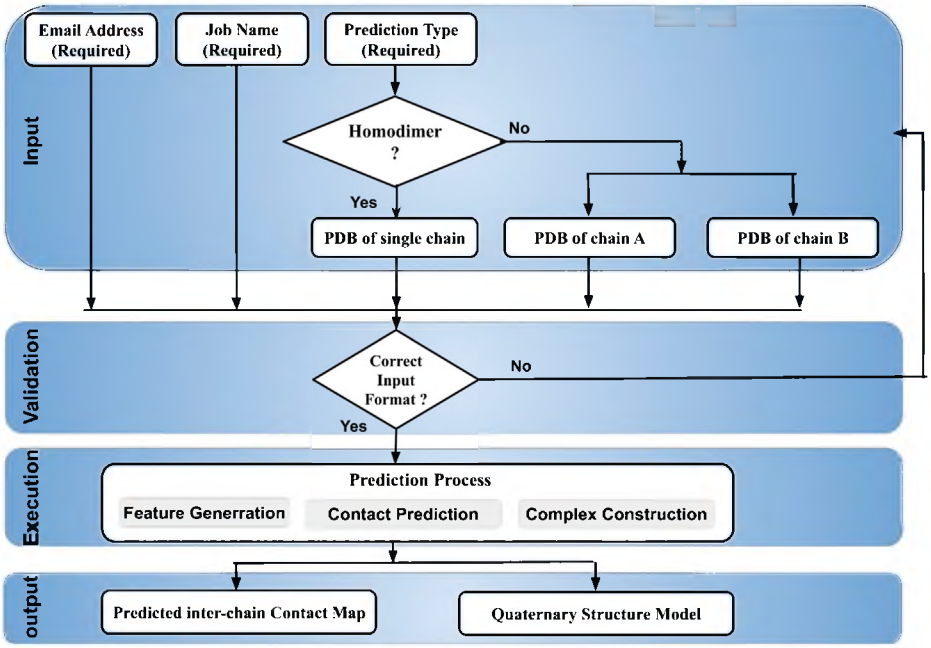

FIGURE 1 | The workflow of DeepComplex illustrating the input, validation, prediction and output.

finished, an email is sent to the email address containing a link to the webpage of the prediction results. The results can be viewed and downloaded at the webpage.

Server Processing

When a user submits a job, the provided information is first validated and the job is then queued for execution. Once the job is scheduled to run, the protein sequence is extracted from the input structure file and is used to generate the multiple sequence alignment from which the residue-residue co-evolution features as well as other features such as secondary structure and solvent accessibility are generated. They are used for the deep learning-based inter-chain contact prediction (Quadir et al., 2020). These inter-chain contacts are then used to generate the distance restraints for the gradient descent optimization method (Soltanikazemi et al., 2021) to predict quaternary structures of dimers.

Server Output

On the successful completion of the job, the user is notified via an email containing a link to the output webpage. DeepComplex outputs a comprehensive set of results such as the sequence of the individual chains, the multiple sequence alignment, the predicted inter-chain contact map, and the reconstructed quaternary structure of the dimer. The predicted structure is shown in JSmol (Hanson et al., 2013), which provides a web-based interactive visualization of the complex structure. The visualization of the multiple sequence alignment (MSA) in a separate window is also possible through a click on the "Alignment File" link. All the information in the output page is downloadable individually or as a zipped file (deepcomplex_results.tar.gz). **Figure 2** shows the continuous screen shots from the input to the final output for both homodimer and heterodimer cases.

Server Implementation

DeepComplex is hosted by the tulip.rnet.missouri.edu server. The operating system of the server is CentOS Linux September 7, 2009, which runs on 64-bit 3 GHz AMD Opteron CPU with 16 cores and 64 GB RAM. The front end of the web server is implemented with HTTP, HTML, cgi-bin, and JavaScript. The backend of the server is implemented in Linux shell script, C, C++, JavaScript, Perl, R, PHP and Python. DeepComplex is freely available and does not require any registration.

Server Performance

The performance of the prediction pipeline of DeepComplex was mainly tested on 115 homodimers from the Homo_Std dataset (Quadir et al., 2021; Soltanikazemi et al., 2021) with predicted inter-chain contacts. The average time taken for a full prediction to be completed for a protein of length of around 500 residues is approximately 477 min with the bulk of the time being for input feature generation. The average TM-score (Zhang and Skolnick, 2005) of the homodimer complex structures generated by the method is 0.76 (average interface RMSD (Lensink et al., 2016) is 7.04 Å; average ligand RMSD (Lensink et al., 2016) is 17.85 Å; and, average percentage of native contacts in predicted model or f_nat (Lensink et al., 2016) score is 30.54%), and for >40% of these homodimers, high-quality structural models with TM-score $\geq$ 0.9 are obtained (Soltanikazemi et al., 2021). The final quality of the complex structures heavily depends on the precision of the predicted contacts and if precision of the contact prediction is over 20%, in most cases good quality quaternary structures can be built by the system. The distance-based complex structure modelling method was also tested on a dataset of 73 heterodimers with true inter-chain contacts, which achieved an average TM-score of 0.92, average

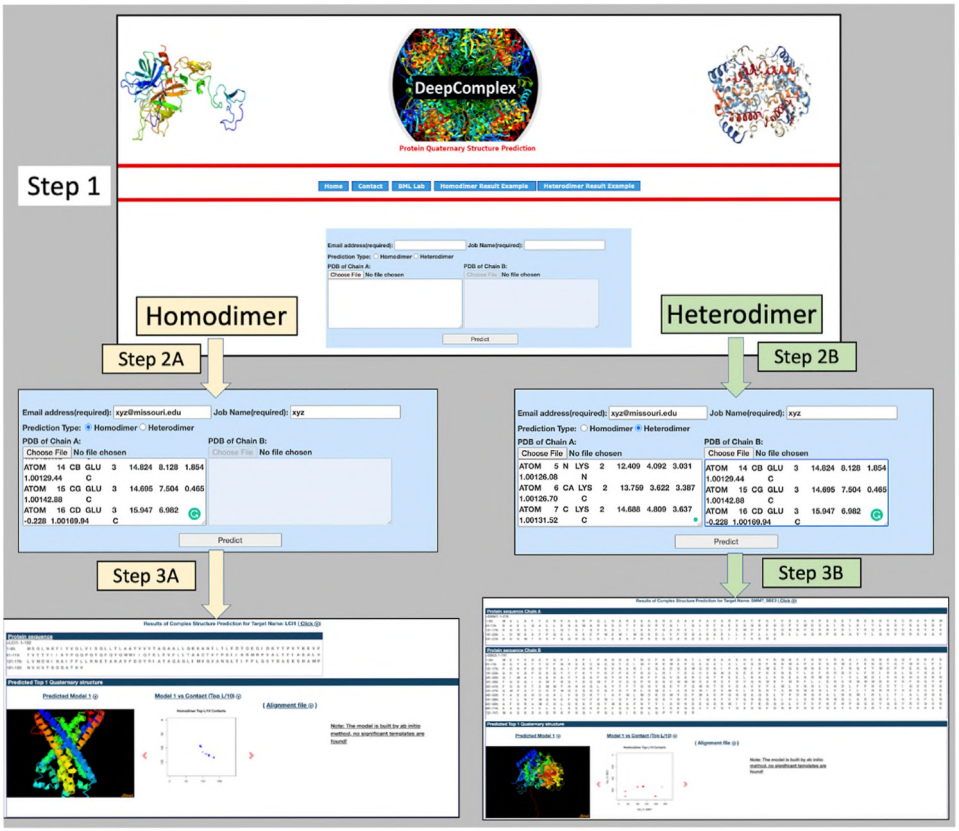

FIGURE 2 | The input and output interface of DeepComplex. Step 1 is the landing page; Steps 2A and 3A are input/output pages for homodimers; and Steps 2B and 3B are input/output pages for heterodimers. In Step 2A, homodimer is selected and the tertiary structure information for one chain is provided; and the output webpage is displayed on Step 3A. Similarly for heterodimer in Step 2B, heterodimer is selected and tertiary structures of the two interacting proteins are provided; and the output is displayed on Step 3B.

interface RMSD of 0.72 Å, average ligand RMSD is 3.75 Å, and average f_nat score of 90.31% for the reconstructed complex structures. The performance on the heterodimers with predicted inter-chain contacts will be evaluated in the near future. The source code of the deep learning inter-chain contact prediction and the gradient descent optimization used by this web server is available at https://github.com/jianlin-cheng/DeepComplex.

CONCLUSION

The DeepComplex web server is a convenient, effective, and user-friendly tool for predicting the structure of homo and heterodimeric complexes. It applies deep learning to predict inter-chain residue-residue contacts of a dimer and then uses them to derive distance restraints for a gradient descent-based optimization method to reconstruct the final quaternary structure. DeepComplex is one of the first web servers that can predict both inter-chain contacts via the deep learning as well as quaternary structures of dimeric complexes via the distance-based modelling. It provides a unique tool for *ab initio* protein quaternary structure prediction which is very different from traditional docking methods.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: https://github.com/jianlin-cheng/DeepComplex.

AUTHORS CONTRIBUTION

JC conceived the project. FQ, RR, ES, and JC designed the server. FQ, RR, and ES implements the server. FQ, RR, ES, and JC wrote the article.

FUNDING

Research reported in this publication was supported in part by three Department of Energy grants (DE-AR0001213, DE-SC0021303 and DE-SC0020400), two NSF grants (DBI 1759934 and IIS1763246) to JC, and an NIH grant (R01GM093123) to JC.

REFERENCES

Alquraishi, M., and Valencia, A. (2019). AlphaFold at CASP13. *Bioinformatics* 35, 4862–4865. doi:10.1093/bioinformatics/btz422

Andreani, J., Faure, G., and Guerois, R. (2013). InterEvScore: a Novel Coarse-Grained Interface Scoring Function Using a Multi-Body Statistical Potential Coupled to Evolution. *Bioinformatics* 29, 1742–1749. doi:10.1093/bioinformatics/btt260

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Rie Lee, R., et al. (2021). Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science.* 10, 1–8. doi:10.1126/science.abj8754

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: Modelling Protein Tertiary and Quaternary Structure Using Evolutionary Information. *Nucleic Acids Res.* 42, W252. doi:10.1093/nar/gku340

Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: An Initial-Stage Protein-Docking Algorithm. *Proteins* 52, 80–87. doi:10.1002/prot.10389

Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004). ClusPro: an Automated Docking and Discrimination Method for the Prediction of Protein Complexes. *Bioinformatics* 20, 45–50. doi:10.1093/bioinformatics/btg371

de Vries, S. J., van Dijk, M., and Bonvin, A. M. (2010). The HADDOCK Web Server for Data-Driven Biomolecular Docking. *Nat. Protoc.* 5, 883–897. doi:10.1038/nprot.2010.32

Dong, G. Q., Fan, H., Schneidman-Duhovny, D., Webb, B., and Sali, A. (2013). Optimized Atomic Statistical Potentials: Assessment of Protein Interfaces and Loops. *Bioinformatics* 29, 3158–3166. doi:10.1093/bioinformatics/btt560

Janin, J. (2002). Welcome to CAPRI: A Critical Assessment of PRedicted Interactions.. *Proteins Structure, Function, and Genetics* 47, 257. doi:10.1002/prot.1011

Garzon, J. I., Lopéz-Blanco, J. R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J., et al. (2009). FRODOCK: a New Approach for Fast Rotational Protein-Protein Docking. *Bioinformatics* 25, 2544–2551. doi:10.1093/bioinformatics/btp447

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., et al. (2003). Protein-Protein Docking With Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* 331, 281–299. doi:10.1016/s0022-2836(03)00670-3

Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T., and Sussman, J. L. (2013). JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied toProteopedia. *Isr. J. Chem.* 53, 207–216. doi:10.1002/ijch.201300024

Hopf, T. A., Schärfe, C. P., Rodrigues, J. P., Green, A. G., Kohlbacher, O., Sander, C., et al. (2014). Sequence Co-Evolution Gives 3D Contacts and Structures of Protein Complexes. *eLife* 3, e03430. doi:10.7554/eLife.03430

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein-Protein Docking Benchmark Version 4.0. *Proteins* 78, 3111–3114. doi:10.1002/prot.22830

Janin, J. (2005). Assessing Predictions of Protein-Protein Interaction: The CAPRI experiment. *Protein Sci.* 14, 278–283. doi:10.1110/ps.041081905

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction With AlphaFold. *Nature.* doi:10.1038/s41586-021-03819-2

Lensink, M. F., Brysbaert, G., Nadzirin, N., Velankar, S., Chaleil, R. A. G., Gerguri, T., et al. (2019). Blind Prediction of Homo- and Hetero-Protein Complexes: The CASP13-CAPRI experiment. *Proteins* 87, 1200–1221. doi:10.1002/prot.25838

Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S. Y., Schneidman-Duhovny, D., Sali, A., et al. (2016). Prediction of Homoprotein and Heteroprotein Complexes by Protein Docking and Template-Based Modeling: A CASP-CAPRI experiment. *Proteins* 84 (Suppl. 1), 323–348. doi:10.1002/prot.25007

Lyskov, S., and Gray, J. J. (2008). The RosettaDock Server for Local Protein-Protein Docking. *Nucleic Acids Res.* 36, W233–W238. doi:10.1093/nar/gkn216

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., and Brown, P. M. G. E. (2020). Single-particle Cryo-EM at Atomic Resolution. *Nature* 587, 152. doi:10.1038/s41586-020-2829-0

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and Accurate Prediction of Residue-Residue Interactions across Protein Interfaces Using Evolutionary Information. *eLife* 3, e02030. doi:10.7554/eLife.02030

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B.-H., Vreven, T., and Weng, Z. (2014). ZDOCK Server: Interactive Docking Prediction of Protein-Protein Complexes and Symmetric Multimers. *Bioinformatics* 30, 1771–1773. doi:10.1093/bioinformatics/btu097

Quadir, F., Raj, R., Soltanikazemi, E., and Cheng, J.(2020). Deepcomplex. Available at: https://github.com/jianlin-cheng/DeepComplex

Quadir, F., Roy, R. S., Halfmann, R., and Cheng, J. (2021). DNCON2_Inter: Predicting Interchain Contacts for Homodimeric and Homomultimeric Protein Complexes Using Multiple Sequence Alignments of Monomers and Deep Learning. *Sci. Rep.* 11, 12295. doi:10.1038/s41598-021-91827-7

Schelling, M., Hopf, T. A., and Rost, B. (2018). Evolutionary Couplings and Sequence Variation Effect Predict Protein Binding Sites. *Proteins* 86, 1064–1074. doi:10.1002/prot.25585

Smith, G. R., and Sternberg, M. J. (2002). Prediction of Protein-Protein Interactions by Docking Methods. *Curr. Opin. Struct. Biol.* 12, 28–35. doi:10.1016/s0959-440x(02)00285-3

Soltanikazemi, E., Quadir, F., Roy, R. S., and Cheng, J. (2021). Distance-based Reconstruction of Protein Quaternary Structures from Inter-chain Contacts. *bioRxiv.* 05, 445503. doi:10.1101/2021.05.24.445503

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* 43, D447–D452. doi:10.1093/nar/gku1003

Tovchigrechko, A., and Vakser, I. A. (2006). GRAMM-X Public Web Server for Protein-Protein Docking. *Nucleic Acids Res.* 34, W310–W314. doi:10.1093/nar/gkl206

Yan, Y., Zhang, D., Zhou, P., Li, B., and Huang, S. Y. (2017). HDOCK: a Web Server for Protein-Protein and Protein-DNA/RNA Docking Based on a Hybrid Strategy. *Nucleic Acids Res.* 45, W365–W373. doi:10.1093/nar/gkx407

Yan, Y., and Huang, S.-Y. (2021). Accurate Prediction of Inter-protein Residue–Residue Contacts for Homo-Oligomeric Protein Complexes. *Brief. Bioinform.* 00, 1–13. doi:10.1093/bib/bbab038

Yu, J., Vavrusa, M., Andreani, J., Rey, J., Tufféry, P., Guerois, R., et al. (2016). InterEvDock: a Docking Server to Predict the Structure of Protein-Protein Interactions Using Evolutionary Information. *Nucleic Acids Res.* 44, W542. doi:10.1093/nar/gkw340

Zeng, H., Wang, S., Zhou, T., Zhao, F., Li, X., Wu, Q., et al. (2018). ComplexContact: A Web Server for Inter-Protein Contact Prediction Using Deep Learning. *Nucleic Acids Res.* 46, W432–W437. doi:10.1093/nar/gky420

Zhang, Y., and Skolnick, J. T. M-align. (2005). TM-align: a Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* 33, 2302–2309. doi:10.1093/nar/gki524

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Quadir, Roy, Soltanikazemi and Cheng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.