Structural biology: the Transformational Era

Shoshana Wodak[1]

[1]VIB-VUB Center for Structural Biology

July 26, 2023

Structural biology: the Transformational Era

by

Shoshana J. Wodak, VIB-VUB Center for Structural Biology, Plailaan 2, 1050 Brussels Belgium. Email: Shoshana.wodak@gmail.com, ORDIC: http://orcid.org/0000-0002-0701-6545

Structural biology has been undergoing an unprecedented transformation recently thanks to major breakthroughs in experimental methods such as cryogenic electron microscopy (cryo-EM) and ground-breaking computational approaches for predicting the 3D structure of proteins based on cutting edge deep-learning methods.

Owing to spectacular advances in detector technology and software algorithms, cryo-EM has revolutionized biology by enabling the determination of complex biomolecular structures at near-atomic resolution[1]. Over less than a decade, the number of near-atomics-resolution structures solved using cryo-EM has grown exponentially [2]. Foregoing the need for crystal formation, it has enabled to elucidate the structures of important receptors and membrane proteins, historically refractory to crystallographic studies [3]. Furthermore, increasingly sophisticated computational and experimental cryo-EM methods are making it possible to unveil different conformational and/or compositional states of the systems under study [4-6], thereby providing valuable information on the dynamic properties of these systems underpinning their biological function.

In parallel, the progressive introduction of new generation methods in deep learning - a subfield of machine learning- to a maturing protein modelling field has recently culminated with the phenomenal success of AlphaFold2 (AF2), the deep-leaning engine developed by Google DeepMind, in predicting the 3D structure of single chain proteins to an accuracy rivaling with that of experimentally determined structures [7, 8]. This achievement has been a game changer with immense repercussions across the fields of computational and experimental structural biology [9, 10]. The software of these algorithms was made freely available to the public [11] [https://github. com/deepmind/alphafold] setting the stage for rapid further developments [12]. Additionally, DeepMind has partnered with the European Bioinformatics Institute (EBI) to create AlphaFold-DB [13], offering open access to over 200 million protein structures predicted by AlaFold, providing broad coverage of UniProt [14].

The vast increase in high accuracy coverage of protein structure space is already having a major impact in many areas of scientific research, including elucidating aspects of evolutionary relationships and protein function [15], identifying potential drug targets[16] and greatly aiding experimental structure determination[17]. However, AF2 as designed, and hence also AlphaFold-DB, provide no information on the dynamic properties of proteins nor on the alternative conformations that proteins sample to carry out their function [18]. Information is also lacking on functionally important bound small molecule ligands, and on the oligomeric structure of native proteins, where two or more proteins (subunits) form higher order complexes[19]. Of these essential areas the prediction of protein complexes, has received special attention in the last two

years. Viewed as the next frontier for deep learning–based structure prediction methods, the community devised ways of extending the power of AF2 to the prediction of protein complexes. Creative uses of AF2 and AlphaFold2-Multimer, the inference engine of AlphaFold directly trained on protein complexes from the PDB[20], which include aggressive sampling of candidate solutions combined with effective scoring and ranking models, helped yielding high-quality models for 40% of the assembly targets in the CASP-CAPRI (*Critical Assessment of Structure Prediction -Critical Assessment of PRedicted Interactions [21]*) blind prediction challenge of 2022 compared to the mere 8% produced in previous challenges [Lensink et al. (under review)]. These are very encouraging results, suggesting nevertheless that significant room remains for improvement [21].

Free access to the code of AF2 and similar deep-learning based software like RoseTTAfold [22], offered by various community-based resources such as ColabFold [12] played a key role in these advances. Access to these resources is also having a resounding impact on the experimental determination of protein structures. In several instances, hard-to- solve X-ray and cryo-EM structures have been elucidated by using AlphaFold predicted structures in molecular replacement protocols [23, 24]. AlphaFold and RoseTTAFold models have been used successfully to fit residual electron density in cryo-EM maps, most notably in a recent assembly of the human nuclear pore complex [25].

This special issue of Proteomics features seven contributions showcasing how the new wave of deep-learning tools and generated data are being leveraged and integrated into cutting edge research in the life sciences and how the frontier between experimental and computational approaches is increasingly blurred. Contributions to this issue also underscore the importance of free access to the data generated by both experimental and computational approaches. These data are inherently complex and noisy, hence the crucial role of tools for extracting useful information from these data, a key step in generating new knowledge.

Varadi and Velankar, the team at the PDBe (Protein Databank Europe), developing and managing the AlpfaFold-DB, in close collaboration with Google DeepMind, describe the specifics of the database, the key meta-information it includes and the impact it is having across the fields of life-sciences research and development. They discuss the challenges of organizing analyzing and providing meaningful user access to 214 million unique protein structures, compared to around 200,000 PDB structures corresponding to 60,000 unique protein sequences. Our attention is attracted to the specifics of the new body of data, including the confidences scores associated with the predicted models, the new insights they provide and some important limitations. Also highlighted is the important role public data providers play in integrating the new structural information with other key biological data and disseminating it across other key resources such as UniProt and more specialized databases such as and InterPro [26] and Pfam [27]among others.

Tüting *et al.* , describe how AlphaFold predicted structures enables the interpretation of cryo-EM maps from native cell extracts. Combining data on crosslinking mass spectrometry[28] with other proteomics techniques and systematic fitting of predicted structures of single chain proteins from AlphaFold-DB into medium-resolution cryo-EM maps of yeast native cell extracts, enabled the team to derived models of the large multi-component heterogenous and plastic protein assembly of the 2.6 MDa complex of yeast fatty acid synthase, the closest one can come today to characterising such assemblies in-situ using cryo-EM.

The study of **Pei *et al*** . al, is another edifying example of how AlphaFold predicted structures are being used to generate new knowledge on cellular processes, in this case providing insights into the critical regulatory roles played by PARylation (the posttranslational modification of proteins by linear or branched chains of ADP-ribose units). To this end the study gathered data on sites modified by PARylation on acidic residues (Asp (Asp (D)/Glu (E)) in more than 300 human proteins. Following the example of an earlier study[29], the joint multiple sequence alignments generated for these proteins were fed to the AlphaFold2 inference engine to predict a set of 260 confident interaction interfaces. Mapping the PARylation sites of interest into these interfaces revealed these sites to occur preferentially in coil and disordered regions and that interaction interfaces featuring these sites involve short linear sequence motifs[30] in both disordered and globular domains. More specifically, D/E-PARylation sites were found in the interfaces of key components of the RNA transcription and export complex, suggesting that systematic PARylation-based regulation intervenes

in multiple RNA-related processes.

Deep Learning methods are also making headway in other areas of structural and systems biology. **Cohen and Schneidman-Duhovny**report a new deep learning model for improving the information on the spatial proximity of residues in multi-subunit complexes derived from crosslinking mass spectrometry (XLMS), which the cryo-EM study of Tüting*et al* . in this issue critically relied on to model the large yeast fatty acid synthase complex from cryo-EM data. Chemical crosslinking followed by mass spectrometry [28] is increasingly used to derive distance constraints or restraints in integrative modeling techniques used to build models of large multi-component protein assemblies. One of the challenges in interpreting crosslinking data is designing a scoring function capable of quantifying how well a candidate model fits the data. Most available approaches set an upper limit on the distance between a cross-linked residue pair and compute the fraction of satisfied crosslinks, neglecting the crucial influence of the spatial neighbourhood on the distance spanned by the crosslinker. This shortcoming is addressed by the deep learning model XlinkNet, trained to predict the optimal distance range -instead of only an upper limit- for a crosslinked residue pair based on their spatial environment of the predicted structure. The model trained and validated using many thousands protein structures from the wwPDB and AlphaFoldDB, and XLMS data on tens of thousands of crosslinks, was shown to accurately classify the distances ranges of most of the tested crosslinks and provide valuable insights into the associated structural determinants. The authors also stress the pressing need for better curation and seamless links to publicly available structural information for*in-vitro* crosslinking data (mainly deposited in the PRIDE database [31]).

Accounting for the dynamic properties of proteins or modeling the alternative conformations that proteins sample to carry out their function, is a long-standing challenge that main-stream protein modeling techniques have been struggling with and deep-learning methods still do not master. **Christoffer and Kihara** propose an approach for modeling conformational changes often associated with the formation of protein complexes, which they apply to protein-nucleic acid complexes. These are very challenging complexes to model because their formation is associated with a large flexibility of the components (see for example ref [32]). The proposed approach focuses on modeling this type of motion for the protein components alone, starting from the unbound version of the corresponding structures and considers systems where this motion involves the reorientation and displacement of relatively rigid domains linked by flexible segments. A customized protein docking algorithm designed to handle this type of motion [33] is used to predict the most likely collective binding modes of all individual domains to the nucleic acid component. Next, an anisotropic network model (ANM) [34] is employed to deform the full protein structures to match the docked domains, and further refine the resulting models to optimize interactions with the nucleic acid component(s). Benchmarking this approach on a limited set of protein-nucleic acid complexes where such large-scale collective motions take place, and illustrating representative examples, suggest that it represents a promising strategy for tackling this difficult modeling problem.

Reliably scoring and ranking candidate models of protein complexes and assigning the oligomeric state of proteins are other important challenges unmet by current modeling algorithms, including deep learning-based methods such as AlphaFold. The latter rely primarily on various confidence scores to rank models whose relation to the physical properties of the protein remains uncertain [35]. **Schweke*et al.*** report a community-wide efforts to tackle these problems. This effort exploits QS-Align [36] and ProtCID [37], two noteworthy specialized resources that characterize protein complexes and their interfaces. Using these resources the study produces a carefully crafted benchmark dataset of ˜1700 homodimer protein crystal structures, which includes both physiological and non-physiological complexes. This dataset is used to evaluate the performance of protein interface scoring functions in discriminating between both types of complexes. The unique features of the dataset stems from its size, accuracy, and the fact that it contained particularly challenging complexes to segregate correctly. Evaluating 252 scoring functions developed by 13 expert groups, this study demonstrates the complementarity of these scoring function and shows that the combined power of these functions outperforms individual scores, paving the way for further optimizing such functions. This has important implications for the development of improved methods for the prediction of protein-protein interactions. The benchmark dataset and its analysis should serve as a valuable resource for such future work.

The last 2 decades have seen an explosive growth of protein-protein interaction (PPI) data derived from both small-scale and proteome-scale interrogations in organisms from bacteria to human [38] as well from various computational methods [39] including AlphaFold [40]. Data from these studies have been used to construct PPI networks, and various properties of these networks have been scrutinized to gain biological insights. With the PPI data being inherently noisy, extracting meaningful information from these networks requires cross referencing and integrating the PPI data with many other types of data, such as protein and gene sequences, gene and protein expression levels, as well as structural data [41]. The availability of tools and resources that facilitate such integration and ensuing analyses is therefore crucial, and particularly relevant to the main topic of this Journal. LEVELNET the resource presented by **Behbahani *et al.*** is such facilitator. Focusing on proteins whose 3D structures are available in the PDB, LEVELNET integrates and explores PPI networks from multiple sources of evidence. It builds a grid of networks for each source representing different views of the associated interactions. It allows to cluster interactions made by groups of related proteins based on sequence identity and to infer interactions through homology transfer. Examples of potential applications include the investigation of the structural evidence supporting PPIs associated with specific biological processes, comparing the PPI networks obtained through computational inference versus homology transfer, and creating PPI benchmark datasets with desired properties.

This transformational era is propelling structural biology to the mainstream of research in the life sciences and beyond. This momentum will benefit from ensuring free access to data and tools, and from enhancing the synergy between multidisciplinary research, data providers, and community-wide initiatives that critically benchmark and evaluate progress in the field.

References

[1] Cheng, Y., Grigorieff, N., Penczek, P. A., Walz, T., A primer to single-particle cryo-electron microscopy. *Cell* 2015, *161* , 438-449.

[2] Afonine, P. V., Klaholz, B. P., Moriarty, N. W., Poon, B. K.*, et al.* , New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr D Struct Biol* 2018,*74* , 814-840.

[3] de Oliveira, T. M., van Beek, L., Shilliday, F., Debreczeni, J. E., Phillips, C., Cryo-EM: The Resolution Revolution and Drug Discovery.*SLAS Discov* 2021, *26* , 17-31.

[4] Baretic, D., Pollard, H. K., Fisher, D. I., Johnson, C. M.*, et al.* , Structures of closed and open conformations of dimeric human ATM. *Sci Adv* 2017, *3* , e1700933.

[5] Zhong, E. D., Bepler, T., Berger, B., Davis, J. H., CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat Methods* 2021, *18* , 176-185.

[6] Kinman, L. F., Powell, B. M., Zhong, E. D., Berger, B., Davis, J. H., Uncovering structural ensembles from single-particle cryo-EM data using cryoDRGN. *Nat Protoc* 2023, *18* , 319-339.

[7] Jumper, J., Evans, R., Pritzel, A., Green, T.*, et al.* , Highly accurate protein structure prediction with AlphaFold.*Nature* 2021.

[8] Jumper, J., Evans, R., Pritzel, A., Green, T.*, et al.* , Applying and improving AlphaFold at CASP14. *Proteins* 2021,*89* , 1711-1721.

[9] Callaway, E., 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* 2020,*588* , 203-204.

[10] Akdel, M., Pires, D. E. V., Pardo, E. P., Janes, J.*, et al.* , A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* 2022, *29* , 1056-1067.

[11] Jumper, J., Hassabis, D., Protein structure predictions to atomic accuracy with AlphaFold. *Nature Methods* 2022, *19* , 11-12.

[12] Mirdita, M., Ovchinnikov, S., Steinegger, M., ColabFold - Making protein folding accessible to all. *bioRxiv* 2021, 2021.2008.2015.456425.

[13] Varadi, M., Anyango, S., Deshpande, M., Nair, S.*, et al.* , AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.*Nucleic acids research* 2022, *50* , D439-D444.

[14] UniProt, C., UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021, *49* , D480-D489.

[15] Bordin, N., Dallago, C., Heinzinger, M., Kim, S.*, et al.* , Novel machine learning approaches revolutionize protein knowledge.*Trends Biochem Sci* 2023, *48* , 345-359.

[16] Ren, F., Ding, X., Zheng, M., Korzinkin, M.*, et al.* , AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor.*Chem Sci* 2023, *14* , 1443-1452.

[17] Varadi, M., Velankar, S., The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*2022, e2200128.

[18] Fleishman, S. J., Horovitz, A., Extending the New Generation of Structure Predictors to Account for Dynamics and Allostery. *J Mol Biol* 2021, *433* , 167007.

[19] Perrakis, A., Sixma, T. K., AI revolutions in biology: The joys and perils of AlphaFold. *EMBO Rep* 2021, *22* , e54046.

[20] Evans, R., O'Neill, M., Pritzel, A., Antropova, N.*, et al.* , Protein complex prediction with AlphaFold-Multimer. *BioRxiv*2021.

[21] Wodak, S. J., Vajda, S., Lensink, M. F., Kozakov, D., Bates, P. A., Critical Assessment of Methods for Predicting the 3D Structure of Proteins and Protein Complexes. *Annu Rev Biophys* 2023, *52* , 183-206.

[22] Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J.*, et al.* , Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021, *373* , 871-876.

[23] Kryshtafovych, A., Moult, J., Albrecht, R., Chang, G. A.*, et al.* , Computational models in the service of X-ray and cryo-electron microscopy structure determination. *Proteins* 2021, *89* , 1633-1646.

[24] McCoy, A. J., Sammito, M. D., Read, R. J., Implications of AlphaFold2 for crystallographic phasing by molecular replacement.*Acta Crystallogr D Struct Biol* 2022, *78* , 1-13.

[25] Mosalaganti, S., Obarska-Kosinska, A., Siggel, M., Turonova, B.*, et al.* , Artificial intelligence reveals nuclear pore complexity. *bioRxiv* 2021.

[26] Blum, M., Chang, H. Y., Chuguransky, S., Grego, T.*, et al.* , The InterPro protein families and domains database: 20 years on.*Nucleic Acids Res* 2021, *49* , D344-D354.

[27] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M.*, et al.* , Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021, *49* , D412-D419.

[28] Iacobucci, C., Gotze, M., Sinz, A., Cross-linking/mass spectrometry to get a closer view on protein interaction networks.*Curr Opin Biotechnol* 2020, *63* , 48-53.

[29] Bryant, P., Pozzati, G., Elofsson, A., Improved prediction of protein-protein interactions using Alpha-Fold2. *Nat Commun* 2022,*13* , 1265.

[30] Tompa, P., Davey, N. E., Gibson, T. J., Babu, M. M., A million peptide motifs for the molecular biologist. *Mol Cell* 2014,*55* , 161-169.

[31] Perez-Riverol, Y., Bai, J., Bandla, C., Garcia-Seisdedos, D.*, et al.* , The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022,*50* , D543-D552.

[32] Dimitrova-Paternoga, L., Jagtap, P. K. A., Chen, P. C., Hennig, J., Integrative Structural Biology of Protein-RNA Complexes.*Structure* 2020, *28* , 6-28.

[33] Christoffer, C., Kihara, D., Domain-Based Protein Docking with Extremely Large Conformational Changes. *J Mol Biol* 2022,*434* , 167820.

[34] Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C.*, et al.* , Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* 2001, *80* , 505-515.

[35] Roney, J. P., Ovchinnikov, S., State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Phys Rev Lett* 2022,*129* , 238101.

[36] Dey, S., Prilusky, J., Levy, E. D., QSalignWeb: A Server to Predict and Analyze Protein Quaternary Structure. *Front Mol Biosci* 2021, *8* , 787510.

[37] Xu, Q., Dunbrack, R. L., Jr., ProtCID: a data resource for structural information on protein interactions. *Nat Commun* 2020,*11* , 711.

[38] Wodak, S. J., Vlasblom, J., Turinsky, A. L., Pu, S., Protein-protein interaction networks: the puzzling riches. *Curr Opin Struct Biol* 2013, *23* , 941-953.

[39] Singh, R., Park, D., Xu, J., Hosur, R., Berger, B., Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic Acids Res* 2010, *38* , W508-515.

[40] Petrey, D., Zhao, H., Trudeau, S. J., Murray, D., Honig, B., PrePPI: A Structure Informed Proteome-wide Database of Protein-Protein Interactions. *J Mol Biol* 2023, *435* , 168052.

[41] Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D.*, et al.* , The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021, *49* , D605-D612.