protein function and eventually to tame it to the benefit of humankind.

David T. Jones and Janet M. Thornton

¹Department of Computer Science and Department of Structural and Molecular Biology, University College London, Gower Street, London, UK.

²European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, UK.

✉e-mail: d.t.jones@ucl.ac.uk; thornton@ebi.ac.uk

Published online: 11 January 2022

<https://doi.org/10.1038/s41592-021-01365-3>

References

- Pereira, J. et al. *Proteins* **89**, 1687–1699 (2021).
- Jumper, J. et al. *Nature* **596**, 583–589 (2021).
- Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
- Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. *Bioinformatics* **23**, 1282–1288 (2007).
- Vaswani, A. et al. in *Advances in Neural Information Processing Systems 31 (NIPS 2017) Proceedings* (ed. Guyon, I. et al.) 5998–6008 (Association for Computing Machinery, 2017).
- Fuchs, F. B., Worrall, D. E., Fischer, V. & Welling, M. in *Advances in Neural Information Processing Systems 1970–1981* (Curran Associates, 2020).
- Taylor, W. R. & Orengo, C. A. *J. Mol. Biol.* **208**, 1–22 (1989).
- Mariani, V., Biasini, M., Barbato, A. & Schwede, T. *Bioinformatics* **29**, 2722–2728 (2013).
- Senior, A. W. et al. *Nature* **577**, 706–710 (2020).
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
- van der Lee, R. et al. *Chem. Rev.* **114**, 6589–6631 (2014).
- Outeiral, C., Nissley, D. A. & Deane, C. M. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.09.20.461137> (2021).
- Akdel, M. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.09.26.461876> (2021).
- Diwan, G. D., Gonzalez-Sanchez, J. C., Apic, G. & Russell, R. B. *J. Mol. Biol.* **433**, 167180 (2021).
- Pak, M. A. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.09.19.460937> (2021).
- Moffat, L., Greener, J. G. & Jones, D. T. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.08.24.457549> (2021).
- Jendrusch, M., Korb, J. O. & Sadiq, S. K. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.10.11.463937> (2021).
- Bryant, P., Pozzati, G. & Elofsson, A. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.09.15.460468> (2021).
- Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.10.23.465575> (2021).
- Evans, R. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.10.04.463034> (2021).
- McCoy, A. J., Sammiti, M. D. & Read, R. J. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.05.18.444614> (2021).
- Beckham, K. S. H. et al. *Sci. Adv.* <https://doi.org/10.1126/sciadv.abg9923> (2021).
- Leman, J. K. et al. *Nat. Methods* **17**, 665–680 (2020).
- Thornton, J. M., Laskowski, R. A. & Borkakoti, N. *Nat. Medicine* **27**, 1666–1669 (2021).
- Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. *Nat. Rev. Mol. Cell Biol.* <https://doi.org/10.1038/s41580-021-00407-0> (2021).
- Orengo, C. A. et al. *Structure* **5**, 1093–1108 (1997).
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. *J. Mol. Biol.* **247**, 536–540 (1995).
- Cheng, H. et al. *PLOS Comput. Biol.* **10**, e1003926 (2014).
- Baek, M. et al. *Science* **373**, 871–876 (2021).

Acknowledgements

The authors acknowledge help from R. Laskowski, who generated Fig. 3. J.M.T. acknowledges funding from EMBL.

Competing interests

The authors declare no competing interests.

A paradigm shift in structural biology

The release of protein structure predictions from AlphaFold will increase the number of protein structural models by almost three orders of magnitude. Structural biology and bioinformatics will never be the same, and the need for incisive experimental approaches will be greater than ever. Combining these advances in structure prediction with recent advances in cryo-electron microscopy suggests a new paradigm for structural biology.

Sriram Subramaniam and Gerard J. Kleywegt

It has long been a goal of computational biology to bypass the need to determine structures experimentally by predicting accurate 3D structures directly from amino acid sequences. These efforts have steadily gathered momentum over the years, fueled by the biannual CASP community challenge¹ to correctly predict protein folds and structures with accuracy measured against structures independently determined by X-ray crystallography, NMR spectroscopy or cryo-electron microscopy (cryo-EM) methods. The astonishing accuracy with which protein folds can now be predicted by the programs AlphaFold (developed at DeepMind², a subsidiary of Alphabet Inc.) and RoseTTAFold (developed at the University of Washington in Seattle³) represents a dramatic advance in structural biology. DeepMind's collaboration with the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) to make structure predictions available at proteome scale has

resulted in the AlphaFold Protein Structure Database resource (AlphaFold DB; <https://alphafold.ebi.ac.uk>), which in its initial release contained a set of 365,000 predicted structures for unique UniProt entries, covering most of the human proteome and those of 20 other model organisms, and which now contains 800,000 predicted structures, covering most entries in SwissProt^{4,5}. This database is expected to grow to >100 million models in 2022.

What is the scale of this addition? As of 10 November 2021, there were 183,954 released entries for structures of biological macromolecules in the Protein Data Bank (PDB)⁶, collected over the past five decades. Of these, 161,086 were at least in part determined from X-ray crystallography, 13,448 from solution NMR and 8,971 from cryo-EM, with the remaining 837 coming from miscellaneous methods such as solid-state NMR and neutron and electron crystallography. The vast majority of these entries (~75%) report structures that contain

a single kind of polypeptide, including multimers, with the rest corresponding to heteromeric complexes of varying kinds or to RNA or DNA structures. Collectively, these structures represent 56,194 unique proteins (UniProt entries). By sometime in 2022, the AlphaFold DB will thus feature a >2000-fold increase in structural coverage of the known protein sequences and a >700-fold increase in the number of structures.

The arrival of AlphaFold and RoseTTAFold heralds a paradigm shift⁷ in structural biology. Structure determination itself is only the first step in the science of structural biology, which seeks to use structure as a basis to derive insights and hypotheses regarding biological functions and mechanisms; but the barriers to entry have now been lowered. An important point to appreciate is that AlphaFold and RoseTTAFold predict structures of only the polypeptide components. These predictions are thus not yet complete structural models

Model confidence

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

Fig. 1 | Three-dimensional structure predicted by AlphaFold for human BRD4. BRD4 is a 1,400-amino-acid chromatin-binding protein with kinase and acetyltransferase activities that links chromatin structure and transcription. The model is color-coded by the confidence of the prediction (on a scale from 0 to 100). The three well-folded compact regions are domains of BRD4 that have previously yielded to analysis by X-ray crystallography (PDB entries 3MUL, 2DWW and 2JNS), while the long helix is a prediction by AlphaFold for a part of the sequence for which no experimental structural information has been reported yet. Data and graphics available from AlphaFold DB under CC-BY 4.0 license.

in the classical sense of an atomic model obtained by X-ray crystallography: that is, they do not include description of atomic positions for cofactors, metal ions, water molecules and any other ordered, bound ligands. Nevertheless, the availability of the predictions means that across all biological disciplines, studies involving proteins can begin with a structural model, with the focus of experiments being the testing of a series of predictions that can validate, repudiate or refine the model and the structural hypothesis. Iteration of prediction and experimental validation will now become the process that defines the discipline of structural biology. AlphaFold predictions have already demonstrated sustained accuracy over a range of targets, with insights into where the predictions can be improved^{2–4,8}.

AlphaFold's structure prediction method also independently generates a quantitative estimate of reliability for every residue of a predicted structure, as well as of the reliability of the relative

position and orientation of different parts of it. Models can now be generated for proteins that have been intractable to current methods for experimental structure determination. Although there have been methods to predict structure for a long time, the jump in accuracy achieved by AlphaFold makes model generation from sequence a more credible proposition than before, opening the door to the use of protein structural models (instead of just sequences) by biologists, geneticists, medicinal chemists and physiologists and thus redefining boundaries between disciplines in biology. AlphaFold models can function as 'hypothesis generators' for all functional biologists, providing ideas for the design of experiments that can test the effects of mutations or changes in binding partners. Needless to say, a solid knowledge of structural biology and protein structure will be an essential prerequisite for users to correctly interpret the predicted models.

It is useful to explore where the current gaps lie in this new landscape. The developers of AlphaFold noted that their dataset of human proteome structures "covers 58% of residues with a confident prediction, of which a subset (36% of all residues) [has] very high confidence"⁴. The 42% that AlphaFold cannot predict with a high degree of accuracy is likely to include proteins and regions of proteins that are intrinsically disordered, including regions that may be engaged in interactions with other proteins, RNA, DNA or other molecular partners. The human bromodomain protein BRD4, which plays a critical role in transcription and in cell cycle regulation, is a good example of this category of proteins: three of the four regions of BRD4 that are predicted with high confidence by AlphaFold have been previously amenable to analysis by X-ray crystallography, but for a majority of the polypeptide, the predictions are at low confidence (Fig. 1). An unexpected feature of AlphaFold predictions appears to be that the regions where the structure predictions are of low confidence also have a high probability of being disordered in isolation^{9,10}. For such regions, structural insights will need to be derived through extensive biochemical analysis and through structural exploration of complexes formed with other partner proteins or nucleic acids that stabilize the interfaces.

Another important limitation is that the predicted structural models do not provide insight into conformational dynamics. For membrane proteins such as transporters and ion channels, AlphaFold predictions typically generate one of the states of the

protein, but understanding the mechanism of action will require knowledge of the broader conformational landscape. This is also true for multidomain and allosteric proteins, for which function is intimately connected to changes in tertiary and quaternary structure spanning multiple conformational states. The development of computational methods to address this gap can yield great dividends, but for the foreseeable future, we will need experimental approaches to discover and validate the trajectory of protein conformational changes.

Discovery of more effective drugs is an area where the use of machine-learning methods will undoubtedly have great impact, and accurate prediction of drug-binding modes can be critically useful as starting points for experiments¹¹. Similarly, predictions that account for the effects on structure of post-translational protein modifications, incorporation of cofactors, metal ions and nonstandard amino acids, and ligand binding will also be extremely valuable. At present, AlphaFold predictions do not cover this level of detail, nor does the program provide clues to the locations of water molecules, which may have critical functional roles in reaction mechanisms. Nevertheless, the availability of the new AlphaFold models significantly increases the number of targets that can be explored for drug design and will catalyze many more efforts at structure-guided drug discovery.

Unraveling cell biological mechanisms requires an understanding of the diverse set of protein–protein interactions that are at the heart of cell function. One notable gap evident in the initial reports of AlphaFold and RoseTTAFold was that the predictions covered only single-chain proteins. Since the original publications in July 2021, there have been exciting advances reported on this front^{12,13}, with spectacular success in predicting certain classes of protein–protein interfaces. Despite these developments, it is worth noting that the scale of the problem is immense because of the massive network of interactions that take place in and outside the milieu of the cell. Here again, BRD4, a key cellular protein whose function involves interactions with numerous other proteins and nucleic acids, and that is also a target for cancer drug discovery, serves as a good example. An interaction map of BRD4 from the String database¹⁴ illustrates the enormous complexity of the protein's interactions in the cell (Fig. 2). Some of these interactions are mediated by the folded regions, but many more are likely to involve the unstructured regions indicated

Fig. 2 | A protein interaction network for BRD4 generated by the String database¹⁴. The model shows the set of proteins that likely interact with BRD4 (see Fig. 1). Each node represents all of the proteins produced by a single protein-encoding gene locus but does not include information on the multimerization status of the designated proteins. For a detailed annotation of the information contained in the network map, please visit <https://string-db.org/> and run the search using BRD4 as an input. Data and graphics available from String under CC-BY 4.0 license.

in Fig. 1. The use of effective computational and experimental tools to derive and validate these interaction maps is certain to remain an active research area for a long time to come.

Fortunately, several of the gaps in the predictions of these powerful new methods correspond almost exactly to the areas where cryo-EM can provide useful information. Cryo-EM methods based on ‘single-particle’ imaging enable the determination of 3D structures of macromolecular assemblies that are frozen by rapid vitrification in cryogens such as liquid ethane. The rapid freezing process allows the preservation of protein structures under near-native conditions. The rate of deposition of cryo-EM structures in the Electron Microscopy Data Bank (EMDB) has been increasing each year (Fig. 3), and remarkable insights into the structures of both small and large molecular assemblies are being generated. Regions located at protein surfaces that are only partially ordered or can be ordered with the binding of stabilizing antibodies have now been

resolved in many instances. Visualizing multiple conformational and functional states that are populated in the samples by 3D classification is a unique strength of the cryo-EM toolbox.

Because cryo-EM structure determination does not require crystallization, heterogeneous post-translational modifications such as glycosylation (for example, in the SARS-CoV-2 spike glycoprotein^{15,16}) can be studied without needing to be engineered away by mutagenesis, enzymatic degradation or chemical modification. The success of cryo-EM methods in resolving bound ligands and delineating structured water molecules now makes it possible to obtain detailed information relevant to molecular mechanisms at a level that until recently was attainable only with X-ray crystallography. Nevertheless, hurdles are often encountered on the road to successful cryo-EM structural analysis, especially at the stage of specimen preparation. Further, because cryo-EM methods rely on averaging tens or hundreds of thousands of images containing objects

with similar molecular composition and conformation, there are still major challenges in addressing structures of intrinsically disordered proteins and transient interactions. However, these challenges also represent opportunities to capitalize on conditions under which disordered regions become ordered by interaction with other proteins, nucleic acids or ligands.

Whereas single-particle cryo-EM is used to obtain 3D structures from proteins and related complexes after biochemical purification, cryo-electron tomography (cryo-ET) adds another powerful dimension to the use of cryo-EM in understanding how proteins and other macromolecules assemble and interact in the context of an intact cell^{17,18}. Structural information in cryo-ET is obtained by collecting a ‘tilt series’ of images of the same region of a sample at different orientations relative to the electron beam. This set of images can be converted computationally into volumes (also referred to as ‘tomograms’) using principles similar to those used in computed axial tomography (CAT). When this approach is used to image whole cells, meaningful information is only recovered from the cell peripheries, where the sample is thin enough for imaging by transmission electron microscopy. To image the cell interior, thin sections can be prepared using focused ion beam (FIB) milling, followed by lift-out methods that allow the extraction of specific regions of cells by sectioning with a gallium ion beam¹⁹. Tomograms obtained from these thin sections are beginning to provide unprecedented insights into subcellular organization²⁰. Yet another exciting area on the horizon is the use of focused ion beams combined with scanning electron microscopy (FIB-SEM) to rapidly image whole cells and tissue segments in 3D²¹. Although FIB-SEM imaging generates images at lower resolution than cryo-ET, application of artificial-intelligence-driven annotation of the resulting volumes can provide unique insights into subcellular organization and macromolecular localization²².

The intersection of protein-structure prediction and cryo-ET may well turn out to be one of the most fertile areas at the interface between structural and cell biology in the coming decade. Although membrane organization and cytoskeletal elements, as well as large complexes such as ribosomes, can be visualized directly in tomograms, the structural maps obtained by cryo-ET of complexes and subcellular structures are generally at much lower resolution than what is typically achieved with cryo-EM of the corresponding purified complexes.

Fig. 3 | The number of entries at resolutions better than 6 Å released by the Electron Microscopy Data Bank per year from 2012 to 2021. Data from https://www.ebi.ac.uk/emdb/statistics/emdb_resolution_trends_2.

Averaging the superimposed subvolumes of thousands of copies of specific complexes can improve the resolution, but only rarely do such averages approach atomic resolution. However, by combining the information obtained by tomography with AlphaFold and RoseTTAFold predictions, it should be possible to obtain near-atomic-resolution models of progressively larger complexes in their physiological contexts inside the cell. Mass spectrometric and proteomics methods that help characterize the identity of interacting proteins also provide invaluable information to understand the organization of multiprotein subcellular assemblies. Hybrid approaches of this kind that integrate different structural and biochemical modalities are harbingers of a great future for structural cell biology²³.

In the decades leading up to the current era of machine-learning approaches to structure prediction, NMR spectroscopy, X-ray crystallography and cryo-EM have each demonstrated their strengths and limitations in structural biology. Solution NMR studies have helped resolve structures

of small proteins and provided insights into ligand binding and dynamics. X-ray crystallographic methods have helped resolve the structures of a wide range of biomolecules and their complexes ranging from peptides to viruses, provided that they could be coaxed into forming well-ordered 3D crystals. As noted above, cryo-EM methods are emerging as powerful tools to further expand this repertoire to large, dynamic and heterogeneous assemblies. With the arrival of AlphaFold, the toolbox of structural biology has been augmented with a new and formidable set of computational tools. There is little doubt that the growth of structural biology in the next decade will rely profoundly on the synergy between experiment and prediction, fueled by the remarkable complementarity between machine-learning-based predictions and especially cryo-EM technology, which arguably represent the yin and the yang of the new alphabet of structural biology.

Note added in proof: On 9 December 2021, version 2 of the AlphaFold DB was released which now contains 800,000 predicted

structural models and covers most of the entries in SwissProt.

Sriram Subramaniam^{1,2} and Gerard J. Kleywegt³

¹University of British Columbia, Vancouver, BC, Canada. ²Gandeeva Therapeutics Inc, Vancouver, BC, Canada. ³European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. [✉]e-mail: sriram.subramaniam@ubc.ca; gerard@ebi.ac.uk

Published online: 11 January 2022
<https://doi.org/10.1038/s41592-021-01361-7>

References

- Kryshtafovich, A., Schwede, T., Topf, M., Fidelis, K. & Moutl, J. *Proteins* **89**, 1607–1617 (2021).
- Jumper, J. et al. *Nature* **596**, 583–589 (2021).
- Baek, M. et al. *Science* **373**, 871–876 (2021).
- Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
- Varadi, M. et al. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkab1061> (2021).
- wwPDB consortium. *Nucleic Acids Res.* **47**, D520–D528 (2019).
- Kuhn, T. S. *The Structure of Scientific Revolutions* (Univ. of Chicago Press, 1962).
- Akdel, M. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.09.26.461876> (2021).
- Ruff, K. M. & Pappu, R. V. *J. Mol. Biol.* **433**, 167208 (2021).
- Burke, D. F. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.11.08.467664> (2021).
- Vamathevan, J. et al. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
- Evans, R. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.10.04.463034> (2021).
- Humphreys, I. R. et al. *Science* <https://doi.org/10.1126/science.abm4805> (2021).
- Szklarczyk, D. et al. *Nucleic Acids Res.* **49**(D1), D605–D612 (2021).
- Walls, A. C. et al. *Cell* **181**, 281–292.e286 (2020).
- Wrapp, D. et al. *Science* **367**, 1260–1263 (2020).
- Mahamid, J. et al. *Science* **351**, 969–972 (2016).
- Marx, V. *Nat. Methods* **15**, 575–578 (2018).
- Wagner, F. R. et al. *Nat. Protoc.* **15**, 2041–2070 (2020).
- Schaffer, M. et al. *Nat. Methods* **16**, 757–762 (2019).
- Narayan, K. & Subramaniam, S. *Nat. Methods* **12**, 1021–1031 (2015).
- Swedlow, J. R. & Collinson, L. *Nature* **599**, 39–40 (2021).
- O'Reilly, F. J. et al. *Science* **369**, 554–557 (2020).

Acknowledgements

S.S. is supported by awards from a Canada Excellence Research Chair Award, the VGH Foundation, Genome BC and the Tai Hung Fai Charitable Foundation. G.J.K. is supported by EMBL-EBI, with funding from the EMBL member states, as well as by the Wellcome Trust. We thank Z. Wang (EMBL-EBI) for producing Fig. 3 and D. Singer and A. Singh (National Cancer Institute, US National Institutes of Health) for comments.

Competing interests

S.S. is founder and CEO of Gandeeva Therapeutics Inc., a drug discovery company based in Vancouver, BC, Canada.