



# The opportunities and challenges posed by the new generation of deep learning-based protein structure predictors

Mihaly Varadi<sup>1</sup>, Nicola Bordin<sup>2</sup>, Christine Orengo<sup>2</sup> and Sameer Velankar<sup>1</sup>

## Abstract


The function of proteins can often be inferred from their three-dimensional structures. Experimental structural biologists spent decades studying these structures, but the accelerated pace of protein sequencing continuously increases the gaps between sequences and structures. The early 2020s saw the advent of a new generation of deep learning-based protein structure prediction tools that offer the potential to predict structures based on any number of protein sequences. In this review, we give an overview of the impact of this new generation of structure prediction tools, with examples of the impacted field in the life sciences. We discuss the novel opportunities and new scientific and technical challenges these tools present to the broader scientific community. Finally, we highlight some potential directions for the future of computational protein structure prediction.


## Addresses

<sup>1</sup> Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

<sup>2</sup> Institute of Structural and Molecular Biology, University College, London, London, WC1E 6BT, UK

Corresponding author: Varadi, Mihaly ([mvaradi@ebi.ac.uk](mailto:mvaradi@ebi.ac.uk))

 (Bordin N.)

 (Varadi M.)

**Current Opinion in Structural Biology** 2023, **79**:102543

This review comes from a themed issue on **Artificial Intelligence (AI) Methodology in Structural Biology**

Edited by **Andreas Bender**, **Chris de Graaf** and **Noel O'Boyle**

For a complete overview see the [Issue](#) and the [Editorial](#)

Available online xxx

<https://doi.org/10.1016/j.sbi.2023.102543>

0959-440X/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## Keywords

Protein Structure Predictions, Deep learning, Structural biology, Structural Bioinformatics.

## Introduction

One of the fundamental approaches to understanding the function of proteins is to investigate their three-

dimensional structures [1]. The most reliable approach to determining the structure of proteins involves time-consuming and expensive experimental techniques. Based on structures archived in the Protein Data Bank (PDB), the most prevalent experimental method is X-ray crystallography, followed by electron microscopy (EM) [2]. While these methods can yield high resolution and high-quality protein structures, the associated effort and costs make it impossible for structural biologists to keep up with the ever-increasing number of known protein sequences. Indeed, the gap between protein sequences and their structures has grown by orders of magnitude over the past decade [3].

Protein structures are determined solely by their amino acid sequences, challenging researchers and scientific software developers to design algorithms to accurately predict protein structures from sequence data [4]. Computational protein structure prediction tools have been around for decades, but the early 2020s saw a huge step forward in terms of the accuracy of models [5]. The unprecedented quality of models predicted by AlphaFold 2.0 during the 14th Critical Assessment of Structure Prediction competition, followed by the release of an improved version of RoseTTAFold, made accurate protein structure prediction tools available to the broad scientific community [6–8]. These models can now match the accuracy of experimentally determined structures and sometimes even surpass them, as observed in an extensive comparison of NMR-based protein structures and predicted models [9,10]. It is important to note, that while these tools do not require template structures, they do rely on sufficiently deep multiple sequence alignments (MSA). Shallow MSAs lead to poor model quality, reflected in low local confidence scores. These algorithms provide confidence metrics such as the pLDDT score, which corresponds to the model's prediction of its score on the local Distance Difference Test, and the predicted aligned error (PAE), which gives information on the confidence of the relative position of residue pairs in the model. Since the release of these algorithms, research groups have performed rigorous validation and assessment of predicted coordinates and pLDDT confidence scores against various classes of proteins, such as

transmembrane proteins, centrosomal and centriolar proteins, and whole proteomes, with only the validation of PAE data yet to be comprehensive [11–13].

In 2022, over 214 million predicted protein structures became available in the AlphaFold Protein Structure Database (AlphaFold DB), covering most of the sequences in the UniProt database [14]. Access to predicted protein structures on this scale made structural data available to a broader audience than ever before. Researchers with no prior experience in protein modelling can now use these models to tackle challenging biological problems, noting that familiarity with model confidence metrics is still essential to making robust interpretations.

In this review, we give an overview of how the massive amount of predicted protein structures and the underlying open-source algorithms impact the life sciences. We discuss new opportunities and new challenges posed by these significant developments. Finally, we speculate about the directions protein structure prediction might move towards next.

### The impact of high-accuracy protein structure models

The new generation of protein prediction tools required data from the public protein sequence and protein structure resources to train their algorithms. Predicted models now benefit structure determination efforts, structure-based drug design and structural bioinformatics analysis on a scale that was impossible before (Figure 1) [14].

Over the past decades, structural biologists solved over 190,000 macromolecular structures and made them publicly available through the PDB archive [2]. Now, tools such as AlphaFold help scientists predict protein structures that proved too elusive in the past. Predicted protein structures are now routinely used to assist in crystallographic phasing by molecular replacement [15,16] and to fit predictions against electron-microscopy maps [17]. Similar synergistic approaches that combine experimental data and predictions have helped determine the structure of challenging molecular machines, such as the nuclear pore complex [18,19].

Predicted models are not replacing experimentally determined protein structures, especially structures of large macromolecular assemblies, but they have affected specific software and data processing pipelines. For example, the Diamond Light Source, the UK's national synchrotron, has AlphaFold configured on-site to help researchers combine predicted models with the X-ray diffraction data they obtain as part of the downstream data processing pipeline. Indeed, synchrotrons, EM facilities and bioinformatics facilities now frequently host

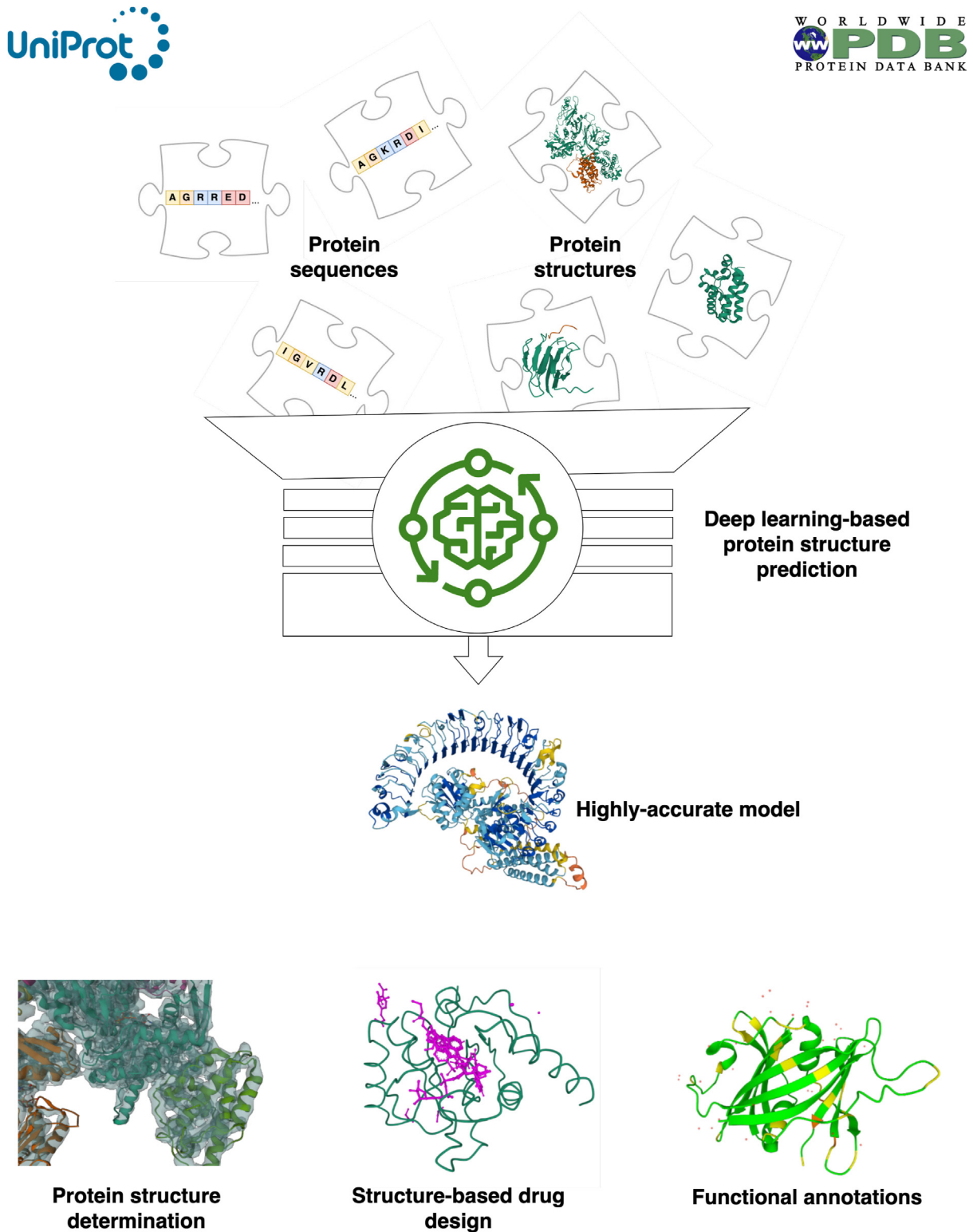
an instance of AlphaFold on-site. Another emerging practice is using predictions to design crystallisation constructs by identifying and excluding longer flexible segments, improving the chance of successful crystallisation [20]. In another application, predicted models help identify potentially interesting post-translational modification (PTM) sites. PTM sites are generally found on accessible, flexible regions of proteins, and AlphaFold models can help locate such regions. This approach effectively filters the number of potential PTM sites and helps researchers focus on the experimental evaluation of more likely candidate sites [21].

The large-scale application of deep learning-based protein structure modelling highlighted the prevalence of intrinsically disordered regions in every modelled proteome [22]. While initial reports suggested that the confidence measure of AlphaFold, the pLDDT score [12], strongly correlates with intrinsic disorder propensity, recent studies show a more complex relationship [23]. Based on large-scale analyses, pLDDT scores lower than 50 are caused either by genuinely poor predictions due to shallow multiple sequence alignments or negatively correlated with intrinsic disorder propensities. However, this correlation seems to hold only for so-called entropic chains and flexible linkers [24]. In the case of disordered regions that adopt stable tertiary structures through binding-induced folding, AlphaFold generally predicts the bound forms with high pLDDT scores [25].

The availability of experimental structures in the PDB led to the birth and steady growth of protein domain classification efforts such as CATH, SCOP, ECOD, SCOPe and SCOP2 [26–29], where protein domains are identified and classified according to their evolutionary history. The growth of these resources always depended on the growth of the PDB, and initiatives such as Gene3D and Pfam [30,31] aimed to increase the structural coverage of the sequence space by obtaining domain assignments using Hidden Markov Models (HMM) matches created from protein domains. AlphaFold dramatically changed the protein domain landscape, as millions of domain sequences became potentially well-modelled domain structures. While a boon for many scientists studying proteins without available structures in the PDB, the sheer size of the data and the potential consequences of basing further research on less-than-optimal models require careful vetting. For instance, 700,000 putative CATH domains were identified in the initial AlphaFold DB release of 21 model organisms, but filtering based on model quality and disordered regions reduced this number by 49% [32].

In addition to model quality considerations, predicted structures generally lack contextual molecules, which may cause inaccurate interpretations. For example, tools

Figure 1



**An overview of deep learning-based protein structure prediction workflows**

The new generation of protein prediction tools used protein sequence data from the UniProt database and protein structures from the PDB to train their models. These tools can provide predicted structures for virtually any protein sequence. This benefits protein structure determination efforts by fitting against experimental data and provides input to structure-based drug design pipelines and structure-based functional annotation software.

such as P2Rank rely on physicochemical attributes to identify potential binding sites. The absence of cofactors, ions and other small molecules in AlphaFold models can influence its behaviour [33]. Data resources such as AlphaFill help address this limitation by expanding AlphaFold models with cofactors, ions and ligands [34].

### New challenges posed by the scale of the available predicted structures

The dataset of 214 million predictions in the AlphaFold DB [14] immensely increased the coverage of the protein sequence space with protein structures and posed new challenges and opportunities in the fields of structural biology and structural bioinformatics. Analytical software optimised to process hundreds of thousands of protein structures may struggle to run efficiently on a much larger dataset. Even data retrieval of a custom data set to start larger scale analysis is not trivial.

While AlphaFold DB provides archive files (TAR files) for 48 proteomes and the Swiss-Prot data set, these are subsets of the data and have their own limitations. Specifically, the archive files only contain the atomic coordinates in compressed PDB and mmCIF files, but the equally important predicted aligned error (PAE) data is missing. The PAE data contains information about the confidence in the relative position of residue pairs. Without this information, it is impossible to determine if the position of two seemingly adjacent regions or domains in a predicted AlphaFold structure can be considered reliable.

While the PDB, mmCIF and PAE data can all be downloaded from the AlphaFold DB prediction pages, it is not a very efficient approach when collating a custom, large data set. To help address this, the complete dataset is made available on the Google Cloud Public Datasets platform. This allows users to retrieve the complete dataset (~23 terabytes, ~1 million TAR files) and to query the database for assembling and downloading large sets of predictions.

The availability of millions of predicted structures raises challenges ranging from data storage to identifying remote homologs and ways to traverse these new large swaths of structure space quickly. State-of-the-art methods for homology annotation of uncharacterized domains before the early 2020s relied on local alignment tools such as BLAST [35], searches against HMM libraries such as HMMER3 [36] and HHsuite [37] if only the sequences were available, or using accurate but very slow structural aligners like DALI, SSAP, TMalign, CE [38–41]. While still valuable, HMMs struggle to detect remote homologs, and using structure alignments is unfeasible with the sheer amount of structures available. Fortunately, the release of AlphaFold DB coincided with

new language models that were successfully trained on proteins, and multiple new predictors based on embeddings from these protein language models were created, tested and validated in various scenarios and were found to outperform established tools for homology detection (including HMMs) [42,43], disorder prediction [44] and ligand-binding prediction [33,45,46]. In the case of AlphaFold-derived protein domain models, they identified in the first release of AlphaFold a correct CATH homologous superfamily for 8% of domains that were elusive to Hidden Markov Models.

Predicted domain assignments to homologous superfamilies require validation by structural comparisons against known homology domains. Using current structural aligners based on double dynamic programming such as SSAP or DALI isn't a feasible solution due to the amount of AlphaFold-derived models. Almost concurrently with the first release of AlphaFold Database, Foldseek - a new, ultra-fast structural aligner by van Kempen and colleagues, was released with comparable accuracy to TMalign while being over 20,000x faster [47].

Combining embeddings-based predictions for homologous superfamily assignments and their validation using Foldseek opened the gates to large-scale annotations of protein domains across structure space. In addition to being applicable without requiring MSAs, embedding-based approaches have the added benefit of working on unlabelled data which could ease its application for tasks such as ligand binding prediction.

### Conclusion and future perspectives

Having unrestricted access to millions of predicted protein structures enables new and innovative research. While posing new challenges to existing scientific software, the amount of new structural data opens up many opportunities in several fields of the life sciences.

Predicting whole assemblies is perhaps the new frontier since these are the functional units in many biological processes. Indeed, accurate models for the whole human interactome may soon be within reach [48,49]. Shortly after the release of AlphaFold and RoseTTAFold, researchers experimented with adopting these algorithms to predict assemblies with some success. Concurrently, a team at DeepMind created a specialised version of AlphaFold, AlphaFold-Multimer, which achieved relatively good accuracy [50]. One could expect that the emphasis will shift to modelling assemblies, and the state-of-the-art algorithms will compete in the Critical Assessment of Prediction of Interactions (CAPRI) and CASP [51].

It would be similarly impactful to develop more accurate tools for modelling interactions between proteins and small molecules [52]. The availability of a reliable

molecular docking algorithm based on advanced AI technologies could revolutionise the field of structure-based drug discovery and accelerate medical research [53].

Creating AI tools that can provide a window into the dynamic nature of proteins is another potential direction [54]. While AlphaFold already demonstrated the prevalence of structurally flexible regions in many proteomes, these models only provide single snapshots from all the possible conformations [55]. Modelling biologically relevant conformational ensembles would open up new opportunities in understanding the biological function of many proteins and could allow drug discovery projects to target intrinsically disordered regions, which is notoriously challenging [56].

Other applications of AI-based structure prediction algorithms could include modelling the structural effects of post-translational modifications, the conformational consequences of mutations and variants, and applications in the field of protein design, but it is important to note that the current versions of popular tools like AlphaFold cannot predict the structural consequences of mutations [57–60].

The arrival of the new generation of accurate protein structure prediction tools is a transformative time for structural biology, structural bioinformatics, drug discovery and many other fields of the life sciences. While these tools apparently excel at their tasks, within certain limitations, the Research and innovation have been accelerated, and a new era of discovery through the application of advanced AI technologies has started.

### 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.

### Data availability

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

### Acknowledgements

The authors would like to acknowledge funding from Wellcome Trust via grants PDBc [218303/Z/19/Z], PDBc-KB [223739/Z/21/Z] and DeepMind (SV, MV) and Wellcome Trust grant [221327/Z/20/Z] (CO, NB).

### References

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

- \* of special interest
- \*\* of outstanding interest

1. PDBc-KB consortium: **PDBc-KB: Collaboratively defining the biological context of structural data**. *Nucleic Acids Res* 2022, **50**:D534–D542.

2. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S: **Protein Data Bank (PDB): the single global macromolecular structure archive**. *Methods Mol Biol Clifton NJ* 2017, **1607**:627–641.

3. UniProt Consortium: **UniProt: the universal protein knowledgebase in 2021**. *Nucleic Acids Res* 2021, **49**:D480–D489.

4. Anfinsen CB: **Principles that govern the folding of protein chains**. *Science* 1973, **181**:223–230.

5. Masrati G, Landau M, Ben-Tal N, Lupas A, Kosloff M, Kosinski J: **Integrative structural biology in the Era of accurate structure prediction**. *J Mol Biol* 2021, <https://doi.org/10.1016/j.jmb.2021.167127>.

This review provides compelling examples of the approaches to integrate computationally predicted protein structures with experimentally determined structures based on various types of experimental methods.

6. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, <https://doi.org/10.1038/s41586-021-03819-2>.

The primary publication of AlphaFold 2.0, which achieved unprecedented accuracy at the 14th CASP competition. This tool was used to predict a massive number of protein structures on the scale of hundreds of millions.

7. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, *et al.*: **Applying and improving AlphaFold at CASP14**. *Proteins* 2021, **89**:1711–1721.

8. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, *et al.*: **Accurate prediction of protein structures and interactions using a three-track neural network**. *Science* 2021, **373**:871–876.

The primary publication of the RoseTTAFold algorithm, which achieved performance on the level of AlphaFold 2.0 in protein structure prediction.

9. Fowler NJ, Williamson MP: **The accuracy of protein structures in solution determined by AlphaFold and NMR**. *Struct Lond Engl* 1993 2022, **30**:925–933.e2.

A comparative assessment of protein structures in the PDB determined by nuclear magnetic resonance spectroscopy and AlphaFold models. It found that the AlphaFold predictions are generally more accurate than NMR-based structures, except in the case of highly flexible regions.

10. Huang YJ, Zhang N, Bersch B, Fidelis K, Inouye M, Ishida Y, Kryshchuk A, Kobayashi N, Kuroda Y, Liu G, *et al.*: **Assessment of prediction methods for protein structures determined by NMR in CASP14 : impact of AlphaFold2**. *Proteins: Struct, Funct, Bioinf* 2021, **89**:1959–1976.

11. van Breugel M, Rosa E Silva I, Andreeva A: **Structural validation and assessment of AlphaFold2 predictions for centrosomal and centriolar proteins and their complexes**. *Commun Biol* 2022, **5**:312.

12. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Židek A, Bridgland A, Cowie A, Meyer C, Laydon A, *et al.*: **Highly accurate protein structure prediction for the human proteome**. *Nature* 2021, **596**:590–596.

13. Hegedűs T, Geisler M, Lukács GL, Farkas B: **Ins and outs of AlphaFold2 transmembrane protein structure predictions**. *Cell Mol Life Sci CMLS* 2022, **79**:73.

14. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, *et al.*: **AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models**. *Nucleic Acids Res* 2022, **50**:D439–D444.

The AlphaFold database provides access to over 200 million predicted protein structures, which is three orders of magnitude larger than the amount of experimentally determined protein structures in the PDB.

15. McCoy AJ, Sammito MD, Read RJ: **Implications of AlphaFold 2 for crystallographic phasing by molecular replacement**. *Acta Crystallogr Sect Struct Biol* 2022, **78**:1–13.

16. Chai L, Zhu P, Chai J, Pang C, Andi B, McSweeney S, Shanklin J, Liu Q: **AlphaFold protein structure database for sequence-independent molecular replacement**. *Crystals* 2021, **11**:1227.
17. Terwilliger TC, Poon BK, Afonine PV, Schlicksup CJ, Croll TI, Millán C, Richardson JaneS, Read RJ, Adams PD: Improved AlphaFold modeling with implicit experimental information. bioRxiv; <https://doi.org/10.1101/2022.01.07.475350>.
18. Fontana P, Dong Y, Pi X, Tong AB, Hecksel CW, Wang L, Fu T-M, Bustamante C, Wu H: **Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold**. *Science* 2022, **376**, eabm9326.
- An impressive demonstration of the power of combining high-accuracy deep learning-based protein structure predictions and cryo-EM data to model large and complex molecular machinery.
19. Mosalaganti S, Obarska-Kosinska A, Siggel M, Turonova B, Zimmerli CE, Buczak K, Schmidt FH, Margiotta E, Mackmull M-T, Hagen W, et al.: Artificial intelligence reveals nuclear pore complexity. bioRxiv; <https://doi.org/10.1101/2021.10.26.465776>.
20. Flower TG, Hurley JH: **Crystallographic molecular replacement using an in silico-generated search model of SARS-CoV-2 ORF8**. *Protein Sci Publ Protein Soc* 2021, **30**:728–734.
21. Bludau I, Willems S, Zeng W-F, Strauss MT, Hansen FM, Tanzer MC, Karayel O, Schulman BA, Mann M: **The structural context of posttranslational modifications at a proteome-wide scale**. *PLoS Biol* 2022, **20**, e3001636.
22. Binder JL, Berendzen J, Stevens AO, He Y, Wang J, Dokholyan NV, Oprea TI: **AlphaFold illuminates half of the dark human proteins**. *Curr Opin Struct Biol* 2022, **74**, 102372.
- This review highlights and discusses the way AlphaFold, through making massive numbers of predictions available in AlphaFold DB, made the high prevalence of intrinsically disordered regions in the proteomes of almost every organism much clearer to the broad scientific community.
23. Alderson TR, Pritišanac I, Moses AM, Forman-Kay JD: **Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2**. *bioRxiv* 2022, <https://doi.org/10.1101/2022.02.18.481080>.
24. Pajkos M, Dosztányi Z: **Functions of intrinsically disordered proteins through evolutionary lenses**. *Prog Mol Biol Transl Sci* 2021, **183**:45–74.
25. Piovesan D, Monzon AM, Tosatto SCE: Intrinsic Protein Disorder, Conditional Folding and AlphaFold2. bioRxiv; <https://doi.org/10.1101/2022.03.03.482768>.
26. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CSM, Woodridge L, Rauer C, Sen N, et al.: **CATH: increased structural coverage of functional space**. *Nucleic Acids Res* 2020, **49**:D266–D273.
27. Chandonia J-M, Guan L, Lin S, Yu C, Fox NK, Brenner SE: **SCoPE: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning**. *Nucleic Acids Res* 2022, **50**:D553–D559.
28. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536–540.
29. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim B-H, Grishin NV: **ECOD: an evolutionary classification of protein domains**. *PLoS Comput Biol* 2014, **10**, e1003926.
30. Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, Orengo C, Lees J: **Gene3D: extensive prediction of globular domains in proteins**. *Nucleic Acids Res* 2018, **46**:D435–D439.
31. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al.: **Pfam: the protein families database in 2021**. *Nucleic Acids Res* 2021, **49**:D412–D419.
32. Bordin N, Sillitoe I, Nallapareddy V, Rauer C, Lam SD, Waman VP, Sen N, Heinzinger M, Littmann M, Kim S, et al.: AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. bioRxiv; <https://doi.org/10.1101/2022.06.02.494367>.
33. Krivák R, Hoksza D: **P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure**. *J Cheminf* 2018, **10**:39.
34. Hekkelman ML, de Vries I, Joosten RP, Perrakis A: AlphaFill: enriching the AlphaFold models with ligands and co-factors. bioRxiv; <https://doi.org/10.1101/2021.11.26.470110>.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403–410.
36. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M: **Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions**. *Nucleic Acids Res* 2013, **41**:e121. –e121.
37. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J: **HH-suite3 for fast remote homology detection and deep protein annotation**. *BMC Bioinf* 2019, **20**:473.
38. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices**. *J Mol Biol* 1993, **233**:123–138.
39. Orengo CA, Taylor WR: **[36] SSAP: sequential structure alignment program for protein structure comparison**. In *Methods in enzymology*. Elsevier; 1996:617–635.
40. Zhang Y, Skolnick J, TM-align: **A protein structure alignment algorithm based on the TM-score**. *Nucleic Acids Res* 2005, **33**:2302–2309.
41. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path**. *Protein Eng Des Sel* 1998, **11**:739–747.
42. Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, Bateman A, DePristo MA, Colwell LJ: **Using deep learning to annotate the protein universe**. *Nat Biotechnol* 2022, **40**:932–937.
43. Nallapareddy V, Bordin N, Sillitoe I, Heinzinger M, Littmann M, Waman V, Sen N, Rost B, Orengo C: CATHe: Detection of remote homologues for CATH superfamilies using embeddings from protein language models. bioRxiv; <https://doi.org/10.1101/2022.03.10.483805>.
44. Ilzhoefer D, Heinzinger M, Rost B: SETH predicts nuances of residue disorder from protein embeddings. bioRxiv; <https://doi.org/10.1101/2022.06.23.497276>.
45. Littmann M, Heinzinger M, Dallago C, Weissenow K, Rost B: **Protein embeddings and deep learning predict binding residues for various ligand classes**. *Sci Rep* 2021, **11**, 23916.
46. Endres L, Olenyi T, Erckert K, Weißenow K, Rost B, Littmann M: Refining Embedding-Based Binding Predictions by Leveraging AlphaFold2 Structures. bioRxiv; <https://doi.org/10.1101/2022.08.31.505997>.
47. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Gilchrist CLM, Söding J, Steinegger M: Foldseek: fast and accurate protein structure search. bioRxiv; <https://doi.org/10.1101/2022.02.07.479398>. FoldSeek is a novel tool that performs structural searches with unprecedented speed. Tools such as these will be required to be able to process and interpret the hundreds of millions of predicted protein structures that are now available.
48. Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, Zhu W, Dunham AS, Albanese P, Keller A, et al.: Towards a structurally resolved human protein interaction network. bioRxiv; <https://doi.org/10.1101/2021.11.08.467664>.
49. Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR, et al.: **Computed structures of core eukaryotic protein complexes**. *Science* 2021:374.
50. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Zidek A, Bates R, Blackwell S, Yim J, et al.: Protein complex prediction with AlphaFold-Multimer. bioRxiv; <https://doi.org/10.1101/2021.10.04.463034>. AlphaFold-Multimer is a specialised version of AlphaFold 2.0, which was optimised to predict assemblies instead of single monomeric chains. Prediction tools such as these are on one of the new frontiers of protein structure predictions, i.e. by aiming to accurately model larger molecular assemblies.

51. Lensink MF, Brysbaert G, Mauri T, Nadzirin N, Velankar S, Chaleil RAG, Clarence T, Bates PA, Kong R, Liu B, *et al.*: **Prediction of protein assemblies, the next frontier: the CASP14-CAPRI experiment.** *Proteins* 2021, **89**:1800–1823.
52. Tong AB, Burch JD, McKay D, Bustamante C, Crackower MA, Wu H: **Could AlphaFold revolutionize chemical therapeutics?** *Nat Struct Mol Biol* 2021, **28**:771–772.
53. Thornton JM, Laskowski RA, Borkakoti N: **AlphaFold heralds a data-driven revolution in biology and medicine.** *Nat Med* 2021, **27**:1666–1669.  
\*  
An in-depth overview of the impact the new generation of highly accurate protein structure predictions can have on applied biology and medicine. The most important findings are that the availability of reliable models will accelerate drug discovery and vaccine development.
54. Lindorff-Larsen K, Kragelund BB: **On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins.** *J Mol Biol* 2021, **433**, 167196.
55. Ruff KM, Pappu RV: **AlphaFold and implications for intrinsically disordered proteins.** *J Mol Biol* 2021, **433**, 167208.
56. Biesaga M, Frigolé-Vivas M, Salvatella X: **Intrinsically disordered proteins and biomolecular condensates as drug targets.** *Curr Opin Chem Biol* 2021, **62**:90–100.
57. Moffat L, Greener JG, Jones DT: Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design. bioRxiv; <https://doi.org/10.1101/2021.08.24.457549>.
58. Sen N, Anishchenko I, Bordin N, Sillitoe I, Velankar S, Baker D, Orengo C: **Characterizing and explaining impact of disease-associated mutations in proteins without known structures or structural homologues.** *Briefings Bioinf* 2022, **23**.
59. Bagdonas H, Fogarty CA, Fadda E, Agirre J: **The case for post-predictional modifications in the AlphaFold protein structure database.** *Nat Struct Mol Biol* 2021, **28**:869–870.
60. Buel GR, Walters KJ: **Can AlphaFold2 predict the impact of missense mutations on structure?** *Nat Struct Mol Biol* 2022, **29**:1–2.