comment

AlphaFold heralds a data-driven revolution in biology and medicine

Protein structures predicted using artificial intelligence will aid medical research, but the greatest benefit will come if clinical data can be similarly used to better understand human disease.

Janet M. Thornton, Roman A. Laskowski and Neera Borkakoti

The protein structure prediction problem is the question of how a protein's sequence of amino acids results in its fully folded three-dimensional structure. This has presented a formidable computational challenge for many decades.

At the end of 2020, a significant advance was announced by DeepMind, a London-based artificial intelligence (AI) company now part of Google's parent firm, Alphabet Inc. DeepMind's AlphaFold 2 program had significantly outperformed other methods in the biennial Critical Assessment of protein Structure Prediction (CASP)[1], producing models of a quality approaching that of experimental determination. AlphaFold 2 has since been published[2] and, more recently, the source code and almost 350,000 protein models from various species, including human, have been made public[3]. This trove of protein structures has implications for both experimental and computational structural biology, and beyond[4–7], but here we consider its possible bearing on medicine.

AlphaFold 2 uses data gathered by structural biologists and made publicly available by the worldwide Protein Data Bank (wwPDB)[8]—which currently holds over 180,000 experimentally determined structures. It is commendable that DeepMind has released the code and predictions for everyone to use.

Over 350,000 protein models have been made available on the AlphaFold Protein Structure Database at the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), with tools to view and interrogate the structures[3]. These proteins come from 21 species, including the most common model organisms and some notable pathogens—*Leishmania infantum*, *Mycobacterium tuberculosis*, *Plasmodium falciparum* and *Trypanosoma cruzi*. Before the end of the year, DeepMind expect to release models covering UniRef90, a unique sample of all known protein sequences comprising 130 million proteins.

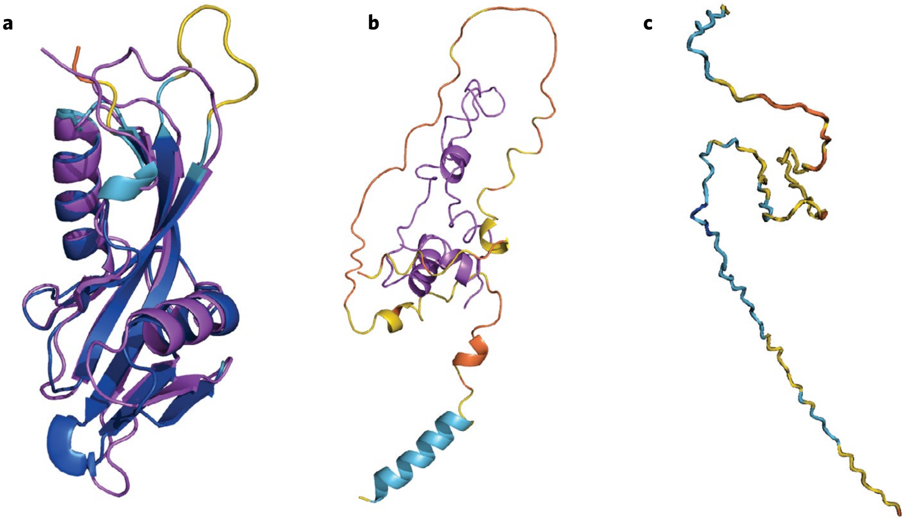

Fig. 1 | The good, the bad and the ugly. a, The good. A superposition of the AlphaFold model of human 14-kDa phosphohistidine phosphatase (UniProt accession Q9NRX4) and the solution NMR structure of the same protein (PDB code 2ai6). The PDB structure is colored purple, while the AlphaFold model is colored according to the pLDDT score: dark blue for the most confidently predicted regions, via light blue and yellow to orange for the regions of very low confidence. The superposition is almost perfect except for the more disordered loop regions. **b**, The bad. Human insulin (UniProt accession P01308) represented by the most complete PDB structure (2kqp) in purple, and the AlphaFold model colored by confidence score (as in **a**). The AlphaFold model bears no resemblance to the PDB structure, possibly because it has missed the disulfide bonds that hold the protein together. **c**, The ugly. The AlphaFold model of human E3 ubiquitin-protein ligase PPP1R11 (UniProt accession O60927), an enzyme classed as EC 2.3.2.27, for which there is no PDB structure, not even of a homolog. One would expect it to be a globular protein, but the AlphaFold model is anything but.

Although protein structures do not of themselves lead to new medicines, they often provide a better understanding of the molecular mechanisms of a protein and in so doing offer insights into how the protein works and how its modulation might lead to a disease or a therapy. Over the past 50 years, protein structures have been an integral part of drug design efforts, with many large pharma companies establishing their own structural biology teams. Structural data have played a critical role both in determining the druggability of a given protein target[9] and then in enabling the design of small-molecule drugs that will bind to it[7].

Variable quality

The AlphaFold AI program rapidly generates models of protein structures from their amino acid sequence more accurately than had previously been achieved. The accuracy of the models is variable (both within and between models) depending on the protein, but, importantly, a confidence measure is provided at each residue position by the predicted local distance difference test (pLDDT) score.

The predictions for single-chain, structured proteins are remarkably good—indeed, comparable in quality to those from experimental structure determination. However, the quality of the predictions depends on the length of the protein and its flexibility.

Not all protein structure predictions are of equal value. Figure 1 highlights three example predictions, showing the good, the bad and the ugly. Figure 2a provides an overview of the coverage (experimental and predicted) and quality of structures for the human proteome. Figure 2b illustrates the distribution of quality scores for the human sequences.

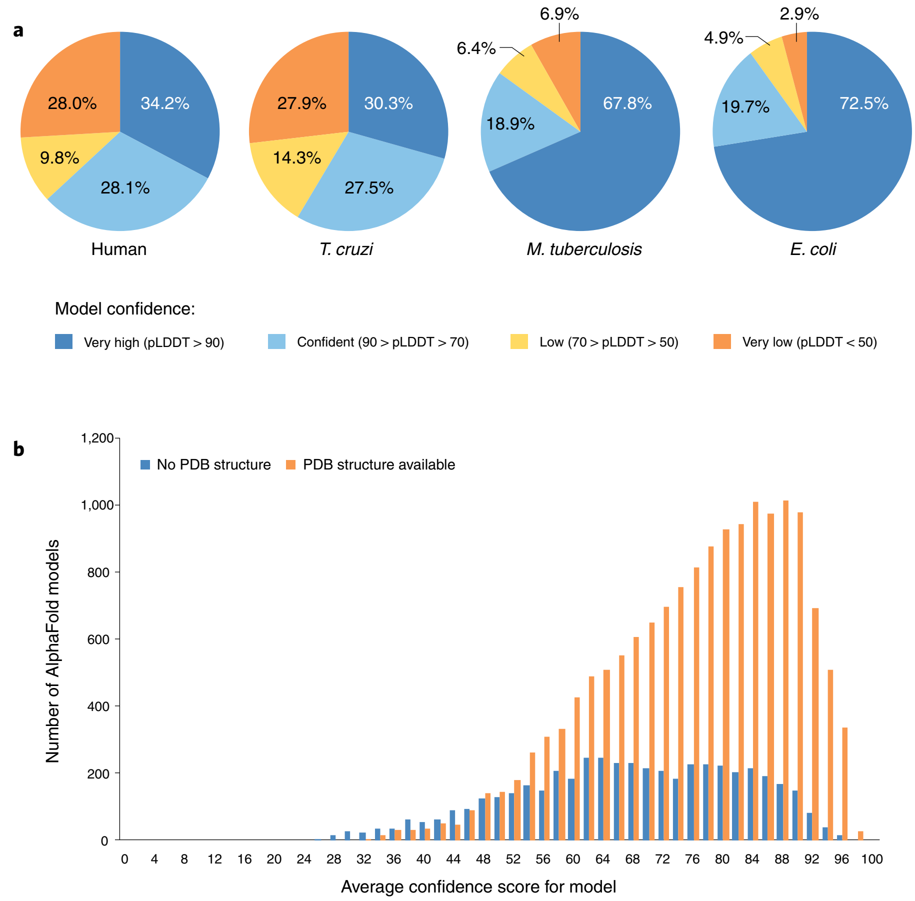

Fig. 2 | Confidence scores for AlphaFold models. a, Distributions of confidence scores for AlphaFold models for four organisms: human, *Trypanosoma cruzi*, *Mycobacterium tuberculosis* and *Escherichia coli*. The scores are classified as very high (dark blue), confident (light blue), low (yellow) and very low (orange). The two bacterial species show over twice as many very highly confident residues as do the other species, possible because they tend to have shorter proteins that can be more confidently predicted. **b**, Distribution of average confidence score per AlphaFold model (obtained by averaging the individual residue confidences over the whole model) for human proteins with no close homolog in the PDB (dark blue) and those in which at least part of the sequence can be homology-modeled from a structure in the PDB (orange). The latter distribution is heavily skewed to higher average confidence scores, suggesting models of higher quality. For long proteins, only the model of the first fragment has been included in the data.

A new structure prediction pipeline

Despite the varying quality of the new structures, SWISS-MODEL[10] has already installed the code from AlphaFold to complement its existing structure prediction pipelines, while other groups have added the models to their databases of protein information, for example UniProt[11] and PDBsum[12]. ColabFold[13] provides tools for modeling multi-chain homo- and hetero-complexes using the AlphaFold and also RoseTTAFold models[14]. Another use of the models is in the interpretation of low-resolution electron microscopy data, especially where the protein shows flexibility between domains.

However, there are major limitations to the relevance of the AlphaFold data to the design of therapeutics. In particular, large multi-domain and flexible proteins still are not modeled very well, and the models lack any ligands (small molecules, DNA, cofactors, metals and other proteins) and therefore do not provide any interaction data, which are especially relevant for elucidating function.

Initially, the AlphaFold models will be used in exactly the same way as experimental structural data (and indeed will be used to help determine low-resolution experimental structures). We see four areas of immediate potential impact for medicine (see Fig. 3).

Therapeutic design

Most small-molecule drugs are designed with the benefit of structural insights[15]. Future design programs (whether for small molecules, biologics, biosimilars or proteolysis targeting chimeras (PROTAC) therapeutics) will use the models from AlphaFold whenever an experimental structure is not available.

For human sequences, the novel coverage is actually rather small (Fig. 2b), especially for those proteins for which drugs have already been developed. It is, of course, invaluable to know the prospective ligand-binding site, preferably with a structure of the complex with a ligand (Fig. 3a). As the predicted models lack all ligands, however, this requires docking approaches, with their varying reliability.

Comparative analyses of the target proteins with AlphaFold models of similar proteins may be used to generate more specific drugs, such as drugs with potentially fewer toxic side effects. In addition, AlphaFold data from different species may be studied to make more informed choices as to the most suitable animal model for testing potential medicines targeted towards humans.

Better drugs and more validated targets are always needed, and although protein structural data may contribute to this, designing small molecules using protein structures at the start of a drug development program is rarely the bottleneck in the time taken to launch a new drug onto the market.

Human pathogenic variants

Structural data help to identify pathogenic variants in humans—that is, those that cause disease[16]. A current challenge is to identify such pathogenic variants (for example, in developmental diseases or cancer progression) among the many variants observed in an individual's genome. Almost 50% of known variants are classified as variants of unknown significance (VUSs) in ClinVar[17], a database of genomic variation and its relationship to human health.

AlphaFold has limited value for modeling the effects of individual mutations, although

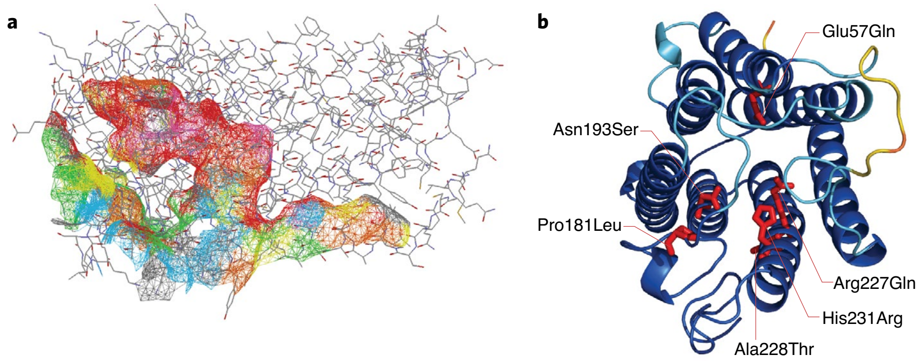

Fig. 3 | Using AlphaFold for drug design and disease-associated variants. a, Application of AlphaFold models to drug design. The protein shown in stick representation is the AlphaFold model of 3-oxo-5-alpha-steroid 4-dehydrogenase 2 (UniProt accession P31213). According to DrugBank, the protein is a target for several drugs, including spironolactone and finasteride. The colored mesh represents the surface of the largest cleft, which forms a deep tunnel. The colors correspond to residue conservation scores (from red for most highly conserved to blue for the least). The large red tunnel suggests a highly conserved binding site that could form a basis for further drug design. The AlphaFold model is virtually identical (root mean squared deviation of 0.4 Å on all C-alpha atoms) to a recent PDB structure of the protein (PDB code 7bw1). **b**, Disease-associated variants. The same protein as in **a**, here shown as a blue cartoon, representing the main chain. The residues shown as red sticks are some of the disease-associated variants (labeled) responsible for pseudovaginal perineoscrotal hypospadias. They mostly line the deep tunnel shown in **a** (here seen end-on) and presumably interfere with the binding of the protein's natural substrate.

reliable models may be used to identify likely binding sites, enzyme active sites, interfaces or structural constraints, and so identify those variants that are more likely to be pathogenic than those that can be benignly replaced by other amino acids (Fig. 3b).

Most functions predicted from sequences or structures rely on close or distant evolutionary relationships. Predicted structures potentially allow one to see further back in evolutionary time, to identify the most distant relatives—from which some functional inference may be drawn.

Drug targets in pathogens

Structural coverage of pathogens in the wwPDB is often much less than for model organisms. With the larger release of data promised for later in 2021, however, predicted structures for many new organisms will be made available.

Protein structures from pathogens such as viruses, bacteria and fungi can be used to assess druggability and possible cross-reactions with human proteins and to aid in the design of medications targeted toward multiple pathogens. Identifying drug targets in infectious agents may provide the most available low-hanging fruit in the short term, and indeed DeepMind is already collaborating with organizations such as Drugs for Neglected Diseases Initiative and other partners.

Enhance vaccine and antibody design

With the COVID-19 pandemic and the development of SARS-CoV-2 vaccines, knowledge of the antigenic spike protein structure has assisted in understanding the surface topology of the virus and its antigenicity.

Amazingly, as of 3 September 2021, there were 1,491 structures of SARS-CoV-2 proteins in the wwPDB[18], contributed by laboratories all around the world. The possibility of predicting viral spike proteins accurately will provide very rapid analysis compared to experimental structure determination for emerging viruses in future pandemics.

A data-driven revolution

The impact of the protein structures from AlphaFold in medicine is potentially substantial. However, AlphaFold is most likely to be just the start of a revolution based on data-driven prediction in biology and medicine. Biological processes at all levels (intracellular, intercellular, organoid and organism) involve interactions between molecules.

Although current AlphaFold predictions are limited to single protein chains and do not provide explicit information about interactions with other molecules, new AI-based tools could predict such interactions across the proteome—delving into different complexes in different cell types, which change with the environment and over time. In the longer term, AI methods will be developed and applied to many aspects of protein structures to improve predictability.

Projects such as the Earth Biogenomes[19] and Darwin Tree of Life[20] that ultimately seek to sequence all living organisms will generate masses of new protein sequence data. AlphaFold2 is the first step to generating the whole structural proteomes for all of these different species. The challenge is then to interpret these genomes in terms of each organism's body shape, development, behavior and natural history, using genotype-to-phenotype studies. Natural products have been the basis for many drugs, so elucidating the genomes of many new species may ultimately lead to novel nature-inspired therapies. No doubt AI methods will be extensively employed in this quest.

From a medical perspective, the opportunities presented by AI are to follow in the footsteps of the DeepMind approach and use clinical data to understand diseases—their diagnosis and prognosis, and determination of what combinations of therapies are best suited for particular patients in a more holistic approach.

Protein Structure Prediction presented the perfect challenge for AI: the data for all known structures were freely available, well curated and organized in the wwPDB. The challenge was very specific, and the success of the outcome measurable and independently assessed in CASP.

The availability of biological research data from institutes such as the US National Center for Biotechnology Information (NCBI) and EMBL-EBI (with the many different types of data and available data resources) has transformed biological research in the last 20 years. The situation for clinical data is entirely different. Like biological data, clinical data are very heterogenous, but they are rarely easily available, often not quantitative, difficult to share across borders and described by limited ontologies and metadata. To add more complexity, such data cannot be made publicly available while maintaining personal confidentiality.

Consequently, to take advantage of the new, powerful AI methods, the imperative with clinical data should be to build the national and international infrastructures necessary to allow clinical data to be collected and shared, collated and standardized.

By analogy with AlphaFold's success in predicting structures, this will accelerate the process of finding therapies that are effective and available to all. In the UK, Health Data Research UK is addressing this challenge by creating Trusted Research

Environments for clinical data, and worldwide, the Global Alliance for Global Health is establishing standards and protocols to enable swifter progress. For this to be successful, multi-disciplinary teams will be needed, involving clinicians, domain experts and machine learning experts, to develop the tools to exploit the data.

It has taken many years to establish the biological databases that are so widely used today—and the challenge for clinical data is even larger. This calls for immediate investment in creating a new health data infrastructure so that patients will be proud to contribute their data to improve human health and the world can face new pandemics with confidence.

Janet M. Thornton ✉,
Roman A. Laskowski and Neera Borkakoti
European Bioinformatics Institute - European Molecular Biology Laboratory EMBL-EBI, South Building, Wellcome Genome Campus, Hinxton, UK.
✉*e-mail: thornton@ebi.ac.uk*

Published online: 12 October 2021
https://doi.org/10.1038/s41591-021-01533-0

References

1. CASP14—14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. Retrieved from https://predictioncenter.org/casp14/ (accessed 27 September 2021).
2. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
3. Tunyasuvunakool, K et al. *Nature* **595**, 590–596 (2021).
4. Jones, D. T. & Thornton, J. M. *Crystallogr. News* **156**, 6–9 (2021).
5. AlQuraishi, M. *Curr. Opin. Chem. Biol.* **65**, 1–8 (2021).
6. Diwan, G. D. et al. *J. Mol. Biol.* 167180 (2021).
7. Workman, P. The Institute of Cancer Research blogs 2021. Retrieved from https://www.icr.ac.uk/blogs/the-drug-discoverer/page-details/reflecting-on-deepmind-s-alphafold-artificial-intelligence-success-what-s-the-real-significance-for-protein-folding-research-and-drug-discovery
8. wwPDB Consortium. *Nucleic Acids Res.* **47**, D520–D528 (2019).
9. Hopkins, A. L. & Groom, C. R. *Nat. Rev. Drug. Discov.* **1**, 727–730 (2002).
10. Waterhouse, A. et al. *Nucleic Acids Res.* **46**, W296–W303 (2018).
11. UniProt Consortium. *Nucleic Acids Res.* **49**, D480–D489 (2021).
12. Laskowski, R. A., Jablonska, J., Pravda, L., Varekova, R. S. & Thornton, J. M. *Protein Sci.* **27**, 129–134 (2018).
13. Mirdita, M., Ovchinnikov, S. & Steinegger, M. Preprint at https://doi.org/10.1101/2021.08.15.456425 (2021).
14. Baek, M. et al. *Science* **373**, 871–876 (2021).
15. Batool, M., Ahmad, B. & Choi, S. *Int. J. Mol. Sci.* **20**, 2783 (2019). (11).
16. Stefl, S. et al. *J. Mol. Biol.* **425**, 3919–3936 (2013).
17. Landrum, M. J. et al. *Nucleic Acids Res.* **48**, D835–D844 (2020).
18. COVID-19 protein structures in the PDB. Retrieved from https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/covid-19.html (accessed 3 September 2021).
19. Lewin, H. A. et al. *Proc. Natl Acad. Sci. USA* **115**, 4325–4333 (2018).
20. Darwin Tree of Life. Retrieved from https://www.darwintreeoflife.org (accessed 27 September 2021).

Author contributions

J.M.T. wrote the first draft of the article, and R.A.L. and N.B. edited and improved it. R.A.L. performed the analyses and created the figures.

Competing interests

J.M.T. sits on the board of Health Data Research UK.

Psychedelic therapy: a roadmap for wider acceptance and utilization

Psychedelics have shown great promise in treating mental-health conditions, but their use is severely limited by legal obstacles, which could be overcome.

Mason Marks and I. Glenn Cohen

The COVID-19 pandemic has exacerbated a national mental-health crisis in the United States. For two decades, drug-overdose deaths have risen exponentially, and suicide rates have steadily increased. These trends reflect deep-seated problems with the healthcare system, including low investment in preventative mental healthcare and a lack of innovation in psychiatry. In search of more effective treatments, clinicians are exploring the therapeutic use of psychedelic compounds, a promising avenue for addressing the mental-health crisis. However, there are social and legal obstacles to making psychedelics a viable treatment option[1].

Schedule I controlled substances

Psychedelics are a class of natural and synthetic compounds that includes psilocybin, MDMA (3,4-methylenedioxymethamphetamine), ibogaine and DMT (dimethyltryptamine). Some psychedelics have been used by Indigenous communities for hundreds or thousands of years. Others were first synthesized in the early 20th century. By the middle of the 20th century, clinicians used psychedelics as adjuncts to psychotherapy, reporting a variety of benefits. However, in the 1970s they were categorized as schedule I controlled substances, which are said to have “no currently accepted medical use and a high potential for abuse”; this blocked mainstream research on these compounds for decades.

In the late 1990s, the US Drug Enforcement Administration (DEA) permitted some researchers to study limited amounts of psychedelics, which allowed research to resume. Clinical trials have now been conducted at leading universities, and a growing body of evidence supports the use of psychedelics, such as psilocybin and MDMA, in the treatment of depression[2], post-traumatic stress disorder[3] and anxiety toward the end of life[4].

The schedule I status of most psychedelics imposes a ceiling on many policy recommendations. The evidence in support of rescheduling is strong, particularly for psilocybin, which is derived from fungi[5]. Unlike other schedule I substances such as heroin, and schedule II compounds, including cocaine and fentanyl, psilocybin exhibits a low risk of toxicity and a very low potential for dependence or addiction[6]. Psilocybin use is not criminalized in several countries, including Portugal and the Netherlands, and a study commissioned by the Dutch Ministry of Health found that over-the-counter sales posed minimal risk to individual people and the public[7].

Acknowledging its therapeutic benefits, the Canadian government made psilocybin available to people with life-threatening illness through compassionate-use regulation. On the basis of clinical-trial data, the US Food and Drug Administration (FDA) designated psilocybin a breakthrough therapy for major depressive disorder and treatment-resistant depression[8].

Rescheduling can occur through several means. The US Congress can amend the Controlled Substances Act, changing the categorization of any controlled substance[9]. Alternatively, the president or the federal