The impact of AlphaFold2 – one year on

David T. Jones[1] and Janet M. Thornton[2]

1., Department of Computer Science and Department of Structural and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, UK
d.t.jones@ucl.ac.uk

2. , European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, CB10 1SD, UK
thornton@ebi.ac.uk

The greatly improved prediction of protein 3-D structure from sequence, achieved by the second version of AlphaFold in 2020, has already started to have a huge impact on biological research; however, in our view a number of challenges remain, and the protein folding problem cannot be considered solved yet. We expect fierce competition to improve the method even further and new applications of machine learning to help illuminate proteomes and their many interactions.

It is now one year since it was announced in the CASP14 meeting, that DeepMind's AlphaFold2 algorithm had done remarkably well[1] in predicting the 3D coordinates of a protein, given its amino acid sequence. Their models proved to be the most accurate of all the submissions for over 90% of the targets, with a fairly large gap to the next best group in most of those cases. AlphaFold2 was the result of using modern machine learning approaches and 170,000 protein structures in the public Protein Databank (PDB) to train the model. Furthermore, they published their method[2], made it freely available and released the predicted models for over 350,000 proteins for 21 model organism proteomes. These are available in the AlphaFold Protein Structure Database hosted at EMBL-EBI, which provides tools to view and interrogate the structures[3]. Furthermore, they intend to release models for UniRef90[4] – i.e. over 130 million protein models by the end of 2021 – ultimately providing more than 700 times as many models as the experimental structures that are currently available in PDB.

This effort has generated an enormous amount of interest in the life sciences community and beyond – not least because this was a problem of many years' standing and it potentially has many applications in proteomics in all domains of life, such as the design of new molecules – either as therapeutics or the design of proteins with new functions. Here we discuss what has happened since the publication of their papers and the release of the models; assess the strengths and weaknesses of the models; identify what has been achieved and what has not been achieved and look ahead to how this might evolve in the future.

The AlphaFold2 Method

The complete description of the AlphaFold2 system takes up 62 pages of supplementary material, and as further details on how AlphaFold2 works can be found elsewhere in this issue, we are not going to detail the inner workings of it here. The basics, however, are that

the system was constructed as a linked set of transformer neural networks, using the concept of attention[5]. These are very different from the standard feedforward neural networks that have been used in bioinformatics for many years, in that they work on strings of input tokens. These tokens are simply vectors which can represent anything from different English words to different amino acids (as they do in AlphaFold2). What makes transformers so powerful is that the attention mechanism considers the relationships not just between tokens found close together in the input sequence, but between any pair of tokens. This gives them an essentially global view of how every token relates to every other token. For AlphaFold2, this gives it a complete view of the interplay between amino acid substitutions modulated by the underlying tertiary (and quaternary) structure.

The overall system architecture of AlphaFold2 has two main processing "tracks", with the inputs to one track representing the rows and columns of a multiple sequence alignment (MSA), and those of the other track essentially representing the interatomic distances between each amino acid in the model. The MSA path allows the network to keep track of amino acid conservation and covariation features, whilst the distance matrix provides the 3-D spatial information for every pair of amino acids. Information is exchanged between these two tracks, which means that the MSA can be reinterpreted as the distance information is improved. Similarly, the distance information can be improved as the MSA is reinterpreted. At the end, information from the two tracks is fed into the so-called structure module, which attempts to construct a 3-D model of the protein i.e., it directly outputs 3-D coordinates for the amino acid residues without needing an external modelling program (called end-to-end prediction). The job of the structure module is not just to produce a single set of coordinates, but also to make iterative improvements to the initial 3-D model, again using an attention-based mechanism, though using a special geometric representation that is invariant to rotations of the structure. Other groups had speculated (wrongly as it turns out) that DeepMind had made use of some new developments in geometric machine learning, called SE(3) equivariant attention[6], which could ensure that the results produced are not affected by rigid body rotations of the protein chain. It turned out, however, that this rotational invariance was achieved using a much older idea from structural bioinformatics where local coordinate frames are defined around each amino acid, based on standard covalent geometry[7].

In some respects, seeing the final complete description of the method was a tiny bit disappointing, after the huge anticipation that built up following the CASP14 meeting. Not because the method, when laid bare, wasn't sophisticated or well thought-out, but simply because there appeared to be no radical new biological insights that were essential to the method's success. In many respects, AlphaFold2 is 'just' a very well-engineered system that takes many of the recently explored ideas in the field, such as methods to interpret amino acid covariation, and splices them together seamlessly using attention processing.

Open questions remain as to how AlphaFold2 has managed to reach such a high accuracy. Some explanation can be had from the ablation study results presented in the final paper. Ablation in machine learning relates to removing critical parts of a machine learning model to see how important they are. In one experiment, the detailed MSA track was effectively removed from the model, leaving just the pairwise distance representation track to produce the predictions. This radical surgery only reduced the performance of AlphaFold2 by a few

percentage points. That's really surprising, because it seemed to most people that the interplay between the two tracks must be the key driver of AlphaFold2's success, and yet, when almost half of the neural network model is deleted, the performance (e.g. on the CASP14 targets) was seen to drop only very slightly. There are two ways we can interpret this. One possibility is that AlphaFold2 is currently over-engineered. In this scenario it should be possible for researchers to simplify the model, and perhaps further improve the accuracy by removing the less important aspects and replacing them with new better ideas. The other, perhaps less optimistic possibility, is that AlphaFold2 is just a collection of many individually small ideas, each of which is adding a tiny percentage to the final performance of the model. An analogy to this might be the design of a Formula 1 racing car, where although none of the little aerodynamic tricks and engine gimmicks contribute much individually, when you remove any of them, the final performance drops just enough to lose the race. Further improvement in this case would be a matter of just piling on even more such tricks.

The quality of AlphaFold2 models can be very variable

AlphaFold2 models have variable quality, often with major differences in reliability across different parts of the chain. This means that the models must be used with great care, with a full understanding of their strengths and weaknesses. Critically, AlphaFold2 provides two measures of confidence, which provide good information about the local reliability (predicted local-distance difference test, pLDDT[8-9]) and the reliability of pairwise interactions between different residues in the chain. It is essential to take both of these measures into account when trying to use the models.

By comparison with proteins for which experimental structures are already available in the PDB, it is clear that most of the AlphaFold2 models generated are of very high quality, with good side chain placement and very low RMSDs. In contrast, for proteins without clear homologues, the pLDDT scores are usually lower and show a very broad range of predicted reliability. (See Fig. 1).

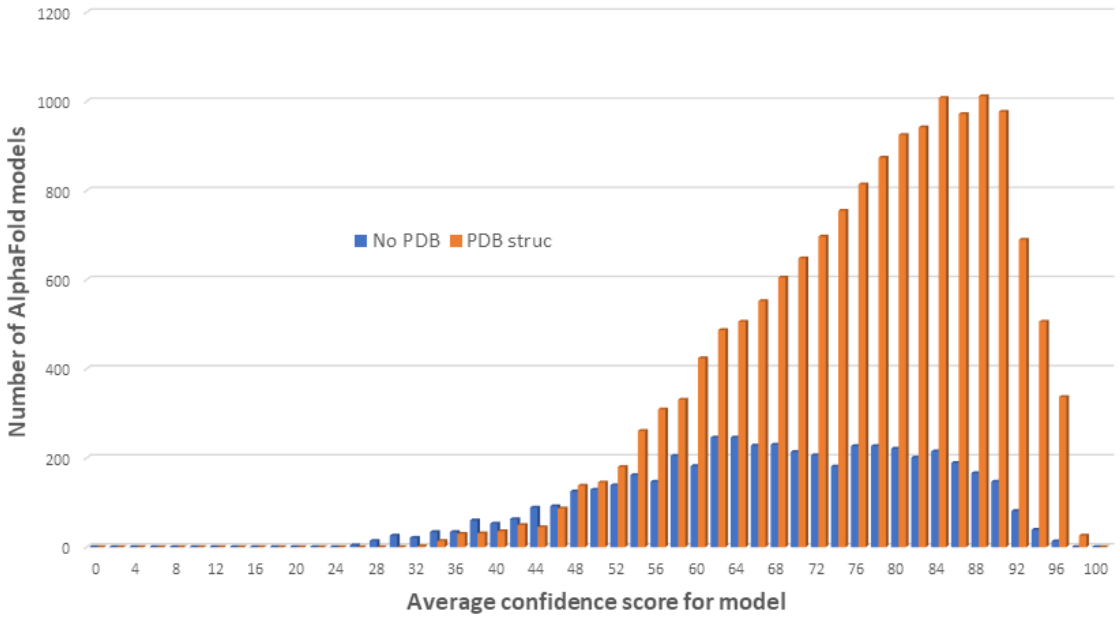

Fig. 1 Distribution of average confidence score per AlphaFold2 model (obtained by averaging the individual residue confidences over the whole model) for human proteins with no close homologue in the PDB (dark blue), and proteins where at least part of the sequence can be homology modelled from a structure 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. (See https://www.nature.com/articles/s41591-021-01533-0)

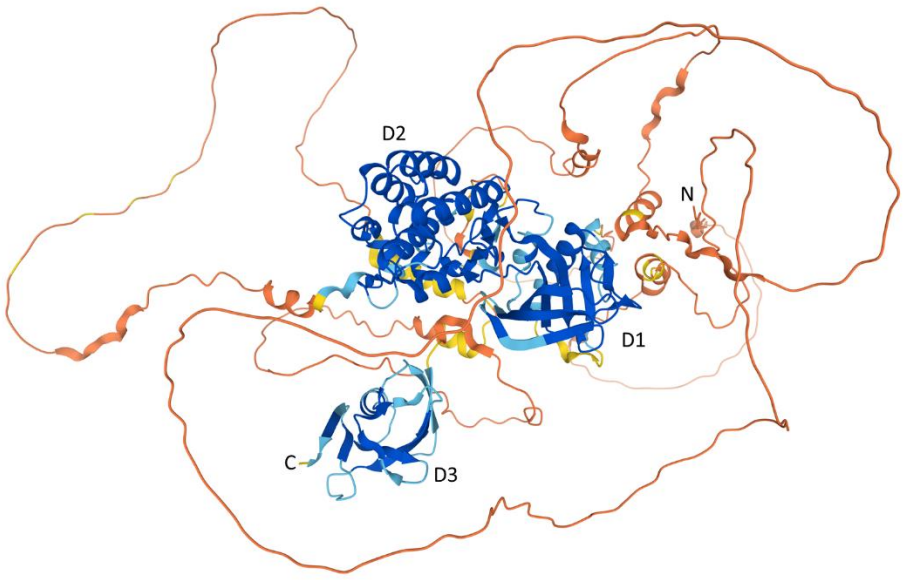

Fig. 2 Entry Q99558 (MAP3K14) from the EBI AlphaFold Database. The model is coloured by confidence score, blue being most confident and orange being least, and three compact structured domains can be seen (D1-D3). The N and C termini are also labelled.

Fig.2 shows a model for a typical large human protein (MAP3K14), taken from the EBI AlphaFold Database. The first two domains (D1 & D2) have already been experimentally determined (e.g. PDB code 6z1t), and so are just being recapitulated in this model, but another compact domain (D3), without any obvious homologues in PDB, is visible at the C-terminus, near the bottom of the figure. This single *de novo* modelled domain is not packed against the other domains, and it is unclear why AlphaFold2 is only sometimes able to pack domains together. One possibility is that it relies on some aspect of similarity to known structures, not necessarily simple homology, but perhaps common motifs or other conserved structural features and without these constraints, the domains are placed seemingly arbitrarily.

The parts of the model where the predictions have very low reliability, are modelled as long loops, which project from the structured core. These regions do not appear to 'obey' the stereochemical rules for polypeptides and should be regarded as arbitrary linkers. For over half of the structure of MAP3K14, AlphaFold2 produces such random coil-like structure at low confidence, interspersed with a few elements of secondary structure at slightly higher confidence. There are minimal interactions between these secondary structures and the well-modelled domains, but some tantalising loose interactions might be in line with the native structure, though it's impossible to say more without further experimental investigation.

Another way to characterise the models is to look at their stereochemistry – in particular the ϕ,ψ distribution, or Ramachandran plot, which depicts the conformation of the main chain of the protein. This distribution was shown to strongly reflect the resolution of an experimental structure[11] and is often used as part of the wider validation of an experimentally determined structure provided in the PDB. These plots can be used to explore the variation of reliability for individual structures, and also reflect how those parts of the chain for which adequate information is not available, have been constructed. The distribution of ϕ,ψ values for the AlphaFold2 human models is shown in Fig. 3.

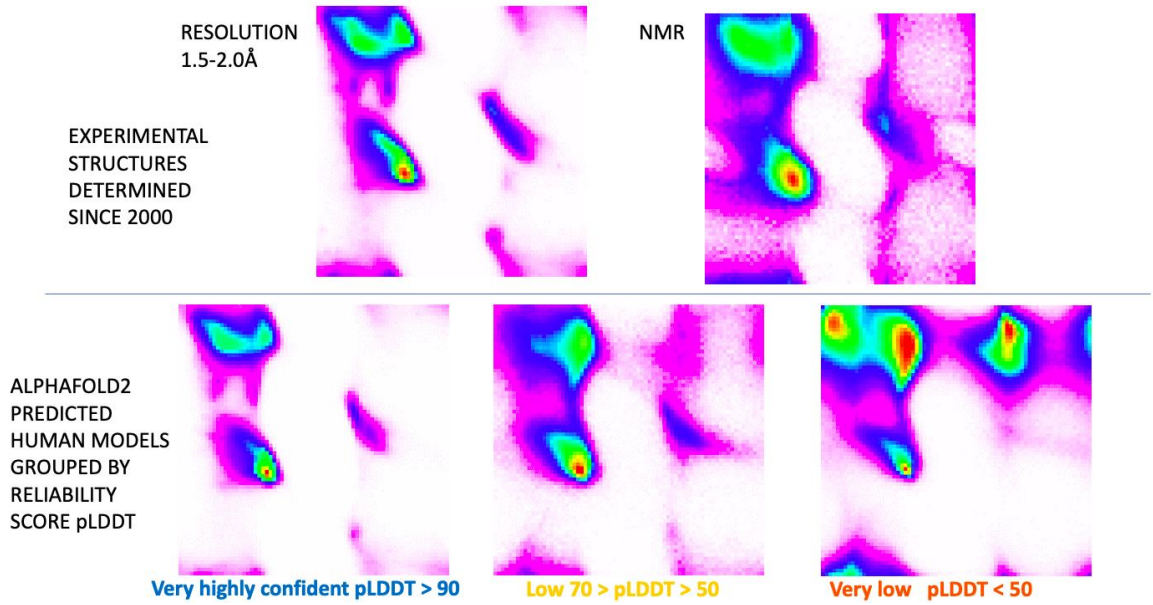

Fig 3. Ramachandran plots of the ϕ,ψ main chain torsion angles for experimentally determined protein structures (a and b) and AlphaFold2 models (c to e). a. Structures solved since 2000 by X-ray crystallography at resolutions of 1.5-2.0Å. b. NMR structures solved since 2000. c-e. Residues in AlphaFold2 models of human proteins with (c) very high pLDDT scores (>90), (d) low scores (70 > pLDDT > 50), and (e) very low scores (pLDDT < 50). For very low pLDDT scores the distribution is radically different from expected. This has implications for running programs based on coordinates of models with large unstructured and poorly resolved loops.

These plots show that below a pLDDT score of 70, the ϕ,ψ distributions of AlphaFold2 models differ from those observed for highly resolved experimental structures, and for very low pLDDT scores, the distribution is radically different from what we expect to see. The low resolution 'loop' regions are clearly not at all physically realistic and will cause errors when using the complete coordinate model for some calculations (eg accessibility, pockets, electrostatic potential etc). An obvious explanation for these spaghetti-like regions in larger AlphaFold2 models is that these regions are disordered[13]. That probably does account for some of it, but it is unlikely to be an explanation for all of it. Many disordered regions become ordered upon binding, and this is an important aspect of their function. Quite possibly, many of these regions may depend on obligate interactions from other chains to create stability, and in this aspect the exclusion of multimeric information from the database is a critical limitation.

Despite this, without very thorough testing, we cannot dismiss the simplest of all explanations for these regions, and that is simply that AlphaFold2 is unable to find the right information in its internal knowledge-base to model these regions. As a rule, when AI systems are given inputs that are far outside the distribution of their training data, they tend to behave in unpredictable ways. One common failure modality is that they simply output either the most common or "average" values, and this is a quite likely explanation for how these low confidence regions are being produced - AlphaFold2 may simply be

outputting structures with close-to-average main chain torsion angles, with maybe some small adjustment for the amino acid type. For sure, in our own tests on AlphaFold2, when we present it with just a long sequence of alanine-residues, we see a single long alpha helix predicted with high confidence, as we might expect. However, presented with a similar sequence of isoleucine residues, which would not be a common feature in the program's training set, it produces the same high confidence alpha helix prediction, which is probably not what we might expect for this amino acid.

Outstanding Challenges Remain

Although the AlphaFold2 software predicts the coordinates of a typical folded protein, there are several related challenges which it is not designed to address. Firstly, AlphaFold2 models do not include any ligands (neither small molecules nor other macromolecules, such as proteins or nucleic acids). Interestingly, AlphaFold2 does sometimes reproduce the holo form of a protein without the ligand being present, which again suggests that it is carrying out some kind of comparative modelling process, albeit a very arcane one. Presumably, in these cases, the majority of structures AlphaFold2 is sampling from have the ligands present, and so the model includes the correct binding site as a useful artefact so to speak. Secondly the method does not aim to elucidate the folding pathway, nor the dynamics of the structure[18]. It may be possible to couple AlphaFold2 with dynamics simulations in the future, however. Lastly, and perhaps the more immediate problem is that AlphaFold2 models cannot be explained or externally validated. From our human perspective, it's essentially 'alien' technology that is currently beyond our understanding, so "asking" it why it predicted something in a particular conformation is clearly not feasible. Other than experimental validation in a lab, there currently seems little prospect of designing computer programs which can independently validate AlphaFold2 models. Its own internal confidence estimates are likely to remain the best indicators of whether a model is right or wrong. This is going to increase as a problem when AlphaFold2 gets serious competition, which it inevitably will. In that case, we will need to be able to compare confidence estimates between different models, and that will be a major challenge. One possible way forward in that respect will be to very carefully curate standardized test sets of protein structures, against which different prediction programs can be calibrated.

New work Inspired by AlphaFold

The AlphaFold2 success and media hype has caused a huge flurry of excitement within the structural biology field and already MANY papers inspired by or using the AlphaFold2 models have been published (mainly as preprints). In fact, the excitement has been such that even preprints were deemed not speedy enough, and quite a lot of useful information ended up being exchanged via social media. One interesting development from this has been the very rapid deployment of AlphaFold2 on Google's Python notebook service called the Google Colaboratory or Colab for short[19].

The flood of AlphaFold2 associated papers, so far, are generally in the area of benchmarking i.e. attempting to validate it using various benchmark sets. Our earlier comments on the importance of proper cross-validation need to be re-emphasised here, and in some respects a lot of this early work does look somewhat rushed. In general, these papers assess

prediction accuracy for particular types of protein domains (eg transmembrane regions) or specific families of proteins; some consider specifically the regions which are not predicted with any accuracy – attempting to use this as a marker of potential disorder[20]. Attempts to use AlphaFold2 to predict the impact of variants on function have not produced convincing results[21-22], and probably more work will need to be done for it to have a bearing on this important area. One other interesting application of AlphaFold2 has been inverse protein folding, or protein design[23-24].

Overall, perhaps the biggest effort to evaluate AlphaFold2 has come from trying to "hack" the software to predict protein-protein interactions. The basic idea is simply to tack the two sequences together and pretend that they are two domains. Surprisingly, this has even had reasonable success at predicting obligate homomeric interactions accurately[20,25-26]. Surprising, because as far as anyone knows, AlphaFold2 was only trained on single protein chains, and so in theory should not have any knowledge of protein-protein interfaces. It's still puzzling how the original AlphaFold2 model, trained only on single chains, can predict multimeric structures at all, but our best guess here, is that there probably is much more in common between domain-domain packing interactions and multimer packing interactions than perhaps we previously thought. This unexpected application has been so popular that DeepMind have already brought out a version of AlphaFold2 that was trained to predict multimeric structures[27].

What are the models most useful for?

Perhaps the single most useful role for the AlphaFold2 models is to seed and solve the determination of experimental structures – especially for large complexes or even tomograms for whole cells. Already several (unpublished) structures have been determined using the models, whilst others (which had already been solved but not published) have proven the accuracy of the predictions. There have been several methodological papers on using the models for molecular replacement[29] . Looking forward with the advent of high resolution cryo EM, the possibility to solve large complexes using these models is very attractive. Indeed, some have already been published[30]

The challenge of solving complexes of proteins with small molecule ligands is greater, since there are a huge number of metabolites (especially in plants) and accurate interaction data are only available for a relatively small number of molecules. The field of docking remains challenging for this reason and it will be interesting to see if these new AI methods can improve docking and energy calculations.

Perhaps the last big question is whether these methods will improve our ability to design proteins with new functions. Great strides have been made in the last decade in designing very stable protein folds with a given structure[31] and even integrating functions. But this remains challenging. The hope is that improved structure prediction will in turn allow us to focus more on protein function and eventually to tame it to the benefit of humankind.

Acknowledgements

We acknowledge help from Roman Laskowski, who generated Figure 3. JMT acknowledges funding from EMBL.

Conflict of Interest
DTJ & JMT have no competing interests.

References

1. Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., & Lupas, A. N. (2021). High-accuracy protein structure prediction in CASP14. Proteins: Structure, Function, and Bioinformatics.
2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A and others. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–589.
3. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Zidek A, Bridgland A, Cowie A, Meyer C, Laydon A and others. Highly accurate protein structure prediction for the human proteome. Nature 2021;595:590-596.
4. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, *23*(10), 1282-1288.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).
6. Fuchs, F. B., Worrall, D. E., Fischer, V., & Welling, M. (2020). Se (3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503*.
7. Taylor, W. R., & Orengo, C. A. (1989). Protein structure alignment. *Journal of molecular biology*, *208*(1), 1-22.
8. Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, *29*(21), 2722-2728.
9. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706-710.
10. Morris, A. L., MacArthur, M. W., Hutchinson, E. G., & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, *12*(4), 345-364.
11. Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography*, *26*(2), 283-291.
12. Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., ... & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, *13*(7), e1005659.

13. Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., ... & Babu, M. M. (2014). Classification of intrinsically disordered regions and proteins. *Chemical reviews*, *114*(13), 6589-6631.
14. Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2021). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 1-16.
15. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093-1109.
16. Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, *247*(4), 536-540.
17. Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., ... & Grishin, N. V. (2014). ECOD: an evolutionary classification of protein domains. *PLoS computational biology*, *10*(12), e1003926.
18. Rubiera, C. O., Deane, C., & Nissley, D. A. (2021). Current protein structure predictors do not produce meaningful folding pathways. *bioRxiv*.
19. https://colab.research.google.com/
20. Akdel, M., Pires, D. E., Pardo, E. P., Jänes, J., Zalevsky, A. O., Mészáros, B., ... & Beltrao, P. (2021). A structural biology community assessment of AlphaFold 2 applications. *bioRxiv*.
21. Diwan GD, Carlos Gonzalez-Sanchez J, Apic G, Russell RB. Next generation protein structure predictions and genetic variant interpretation. J Mol Biol 2021:167180.
22. Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., ... & Ivankov, D. N. (2021). Using AlphaFold to predict the impact of single mutations on protein stability and function. *bioRxiv*.
23. Moffat, L., Greener, J. G., & Jones, D. T. (2021). Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design. *bioRxiv*.
24. Jendrusch, M., Korbel, J. O., & Sadiq, S. K. (2021). AlphaDesign: A de novo protein design framework based on AlphaFold. *bioRxiv*.
25. Bryant, P., Pozzati, G., & Elofsson, A. (2021). Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments. *bioRxiv*.
26. Yin R, Feng BY, Varshney A, Pierce BG. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. bioRxiv; 2021. DOI: 10.1101/2021.10.23.465575.
27. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A. W., Green, T., ... & Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*.
28. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD and others. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373(6557):871-876.
29. McCoy, Airlie J., Massimo D. Sammito, and Randy J. Read. "Possible Implications of AlphaFold2 for Crystallographic Phasing by Molecular Replacement." *bioRxiv* (2021).
30. Beckham, Katherine SH, et al. "Structure of the mycobacterial ESX-5 type VII secretion system pore complex." *Science advances* 7.26 (2021): eabg9923.
31. Leman, Julia Koehler, et al. "Macromolecular modeling and design in Rosetta: recent methods and frameworks." Nature methods 17.7 (2020): 665-680.

Box 1 - Benchmarking AlphaFold2 requires Care and Attention (no pun intended)
Since the AlphaFold2 software was made available, over 30 different benchmarking studies have been made available as preprints through BioRxiv. We can't comment individually on all these studies, but considering them as a whole, a few common concerns have become apparent. As seen in Fig. 1, AlphaFold2's confidence scores are clearly correlated with whether the target structure has homologues in PDB or not. Again, this arises from either the explicit use of templates or from the occurrence of similar structures in AlphaFold2's training set. Removing the effects of directly using templates is easy to control for - that option simply needs to be turned off. Accounting for the effects of bias from the original training data is a wholly different thing. It's already clear that AlphaFold2 is able to build homology models better than any previous homology modelling approach. This means that designing an experiment to properly evaluate its ability to predict new structures is not at all easy. It should be obvious that simply checking the degree of sequence identity (say less than 30%) between the test samples and the training set (effectively all of PDB released prior to May 2018) is not going to be sufficient to avoid bias[14]. Nevertheless, this is the most common approach people seem to have used to produce test sets for AlphaFold2, rendering those studies highly unreliable. Better options exist, such as fold classification databases like CATH[15], SCOP[16] and ECOD[17]. No benchmarking of systems like AlphaFold2 should ever be done without reference to at least one of these resources, which provide evidence for very distantly related homology between protein domains.

Box 2 - The Task of Sisyphus
The EBI AlphaFold Database is a very nice resource to have. Hundreds of thousands of models generated by the current state-of-the-art method in protein modelling, and a nice viewer to look at the models quickly. Also, the models can be downloaded by anyone and therefore used in further studies. However, some issues are likely to emerge in the future.

Firstly, AlphaFold2 is not a single well-defined protocol. Comparing the results from the default AlphaFold2 on CASP14 targets with the results that DeepMind produced themselves, shows quite substantial differences on the harder targets. The AlphaFold Database likely contains a substantial number of currently bad models which could be improved with some changes to the inputs. These changes might arise from new structures being deposited in PDB (providing new templates) or more certainly, additional protein sequences becoming available from metagenomics studies.

Secondly, without any doubt, AlphaFold2 will end up spawning many implementations and even serious competitor methods from other labs. Before long there may be a whole family of methods, some of which may be better (on average) than the current version of AlphaFold, but there then being no obvious way to pick just one "best" model from the alternatives available.

These two concerns will create a huge dilemma. Should the AlphaFold Database be updated? How can it not be? But if so, how often? If the database does grow to 130 million entries, is it feasible (or even desirable) to regenerate that amount of data? A back-of-envelope calculation (assuming roughly 5 minutes of computation per model) gives an estimated total amount of computation of well over 1000 (GPU or TPU) years. Despite carbon offsetting, this is still a lot of energy to use. And unlike the job of painting the Forth

Bridge, which now only needs to be done every 25 years, updating the AlphaFold Database will need to be done much more often. And none of this accounts for the fact that in 5 years' time, UniProt is likely to have twice as many proteins in it, that also will need to be modelled. Of course, computer hardware should get faster in that time, but even if it is just about feasible to keep rerunning AlphaFold, can we afford to ignore the emergence of perhaps better methods that get developed by other labs? Should those models also be stored in the database, and if so, how often should they be updated? Answers to some of these questions will be needed really very soon. We need to know if updating the AlphaFold Database will be merely a labour of Hercules or whether it will end up being a task fit for Sisyphus.

Box 3 - Open Source but not Open Science

One of the most surprising things that came from the AlphaFold2 paper was that it was accompanied by fairly complete source code, released under a standard open source license. DeepMind were rightly applauded for doing this, but there was a fly in the ointment. Although the source code was complete enough to allow users to run the model, either on cloud systems or on local computer systems, which was a definite improvement over the first version of AlphaFold, the all-important neural network parameters were not released under the same license terms. Instead, these parameters were released under a non-commercial license. Of course, as a company, it is entirely up to DeepMind whether or not they release their source code, and under what terms. However, publication of computational work in many of the top journals is increasingly predicated on full open release of the software used to generate the results in the published article. This is particularly important with something as complex as AlphaFold2, where having the source code to look at has been essential to understanding the method fully. Indeed, some minor discrepancies have already been reported between what's written in the paper and what's actually in the source code. In some cases, actually running the code was essential to check exactly how a particular subroutine worked, thus, the ability for all researchers to be able to test computer software is an absolutely vital aspect of the proper academic publishing of computational science research. As things stand, AlphaFold2 cannot be used by commercial researchers, and worse still, the AlphaFold2 models stored at the EBI, although available to download without any restrictions, cannot be validated by commercial users either.

Now, it's somewhat debatable whether the weights of a neural network can be protected as intellectual property, and it's also worth noting that DeepMind are not alone in using this possible loophole to prevent commercial usage of ostensibly open-source machine learning software. The RosettaFold software[28] from David Baker's laboratory has similar restrictions on using its neural network weights. As we say, Open Source certainly, but not truly Open Science.