CASP14: what Google DeepMind's AlphaFold 2 really achieved, and what it means for protein folding, biology and bioinformatics

Disclaimer: this post is an opinion piece based on the experience and opinions derived from attending the CASP14 conference as a doctoral student researching protein modelling. When provided, quotes have been extracted from my notes of the event, and while I hope to have captured them as accurately as possible, I cannot guarantee that they are a word-by-word facsimile of what the individuals said. Neither the Oxford Protein Informatics Group nor I accept any responsibility for the content of this post.

You might have heard it from the scientific or regular press, perhaps even from DeepMind's own blog. Google 's AlphaFold 2 indisputably won the 14th Critical Assessment of Structural Prediction competition, a biannual blind test where computational biologists try to predict the structure of several proteins whose structure has been determined experimentally — yet not publicly released. Their results are so incredibly accurate that many have hailed this code as the solution to the long-standing protein structure prediction problem.

Protein structure is at the core of biochemistry, and has profound implications for medicine and technology. Establishing the structure of a protein is a bottleneck in structure-based drug discovery, and accurate structure prediction is expected to improve the productivity of pharmaceutical research pipelines (although it is only one factor, and we will need to get other things right before truly revolutionary changes happen — check Derek Lowe's posts here and here). Structural information of proteins is also essential in biology, where it helps to elucidate function — many key papers in biochemistry derive insight from experimental advances in structure determination.

Given the importance of the problem, and the wide network of resources that have slowly been advancing for decades, I think nobody was expecting that a solution would be presented too quickly. I myself decided to focus my PhD research in the field of structure prediction, thinking like many others that several years of work across many research lines would be necessary before we could achieve something close to a solution. I may now need to change topics.

How much of the press release is true, what has actually happened, and how significant is it? There has been

endless discussion about this topic in multiple forums. Frankly, I haven't been able to think about anything else for the past 72 hours. In an attempt to clear my own thoughts, I have decided to write this blog post detailing everything I have learned since my scientific world was turned upside down around 3 pm GMT on Monday. I hope this is useful for my fellow protein bioinformaticians who could not attend CASP14, but also to anyone who wants to hear a little bit more about this topic.

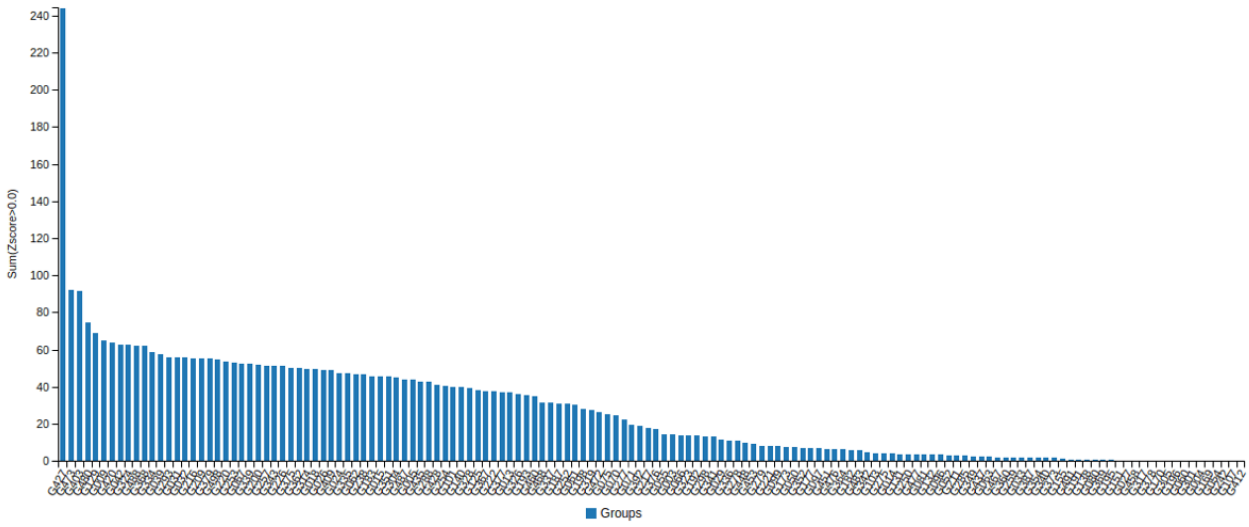
Please bear in mind that my report of the CASP14 assessment and conference will necessarily be interspersed with conjecture. The details of how AlphaFold 2 works are still unknown, and we may not have full access to them until their paper is peer-reviewed (which may take more than a year, based on their CASP13 paper). The magnitude of the breakthrough is undeniable — but we need more details to gauge its potential impact.

This is going to be a long post. Don't say I didn't warn you.

How good is AlphaFold 2, exactly?

Astoundingly so.

Let me tell you the story as it happened last Monday. A handful of hours before the start of the CASP14 meeting — around noon GMT — the organisers released the results of the assignment. Almost immediately, comments started to circulate around Twitter. This is the image everyone was sharing:

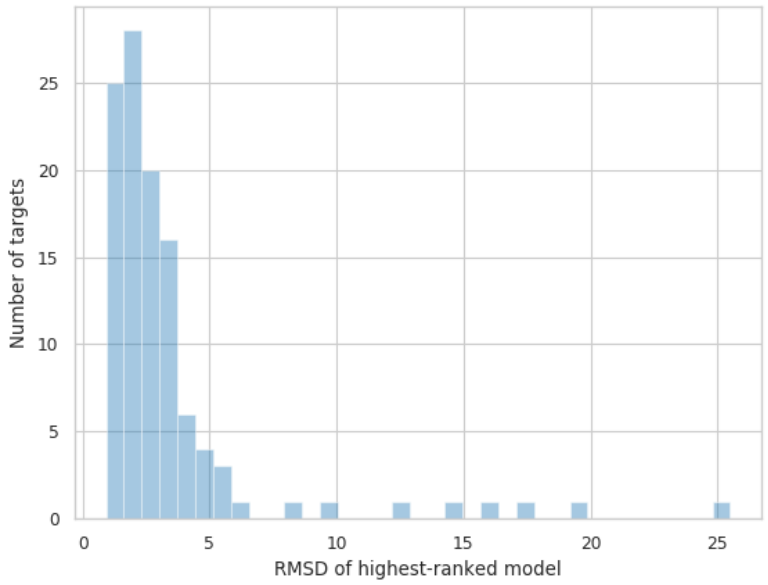


Ranking of participants in CASP14, as per the sum of the Z-scores of their predictions (provided that these are greater than zero). One group, 427, named AlphaFold 2, shows an incredible improvement with respect to the second best group, 473 (BAKER). This figure was obtained from the official CASP14 webpage on Tuesday 1st December, 2020.

This bar plot describes the sum of Z-scores representing the predictions from the different groups. Remember that the Z-score is just the difference of a sample's value with respect to the population mean, divided by the

standard deviation; a high value represents a large deviation from the mean, and is commonly used as a outlier detection procedure. In other words, the groups that are markedly better than the average will have larger Z-scores. In this graph we see that one group performs a lot better than the rest: Group 427, whose average Z-score was around 2.5 when considering all targets, and rose to 3.8 in the hardest ones. If this was an intelligence test, AlphaFold 2 would score an intelligence quotient (IQ) above 160.

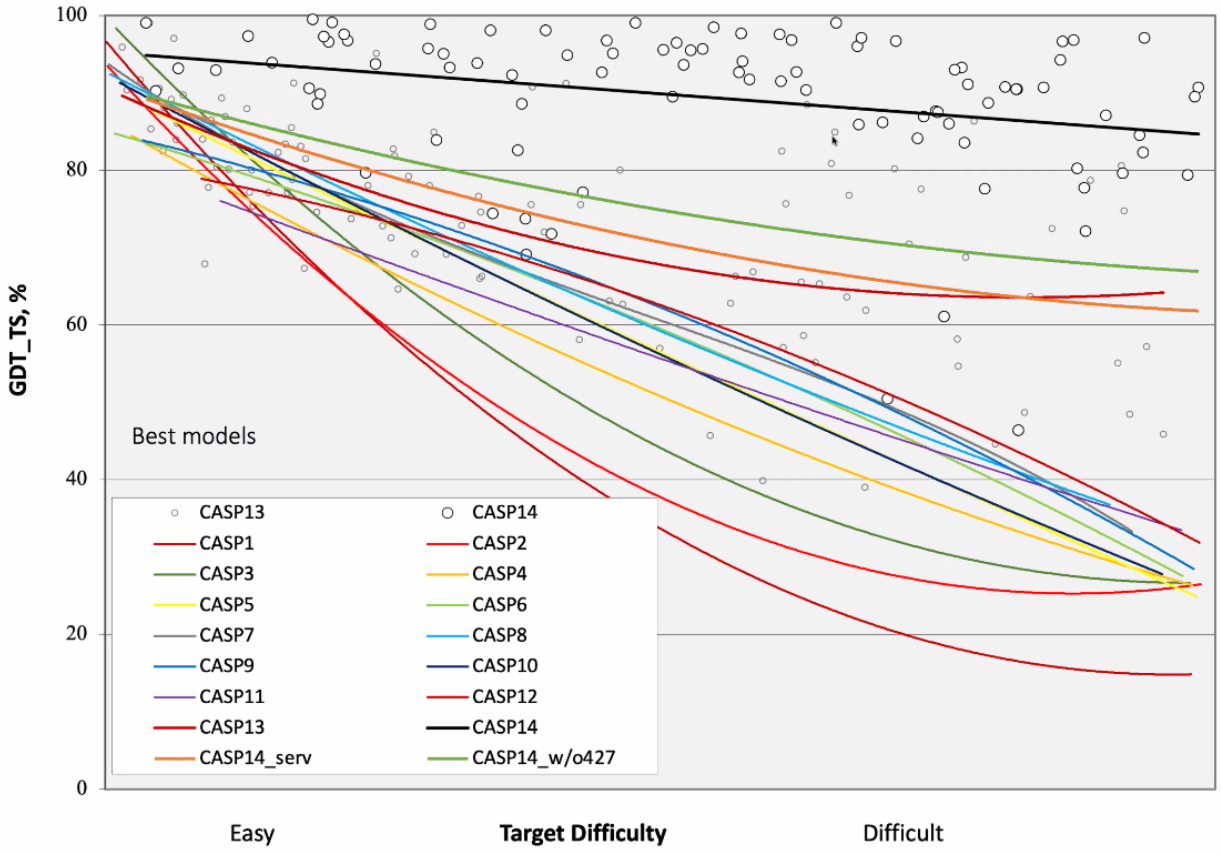
If the relative comparison is astounding, the actual performance is just as impressive. I am going to consider a typical metric in structural biology, the root-mean-square deviation (RMSD) of atomic positions. If you are not much of a protein folder, these numbers may not say much to you. Don't worry — in next section I will show a few graphical examples. Just keep in mind that (1) a lower RMSD represents a better predicted structure, and that (2) most experimental structures have a resolution around 2.5 Å (**updated 8th Dec:** although, as many have pointed out in Twitter, this is an apples to oranges comparison). Taking this into consideration, about a third (36%) of Group 427's submitted targets were predicted with a root-mean-square deviation (RMSD) under 2 Å, and 86% were under 5 Å, with a total mean of 3.8 Å.



Distribution of RMSDs for the highest-ranked models submitted by AlphaFold 2.

Data obtained from the CASP14 website on Tuesday 1st December, 2020.

We were still digesting this information when the conference started and, oh boy, did we suffer the first half an hour. Claims that this year's competition was "a little unusual" were followed by suggestions that one particular group had produced impressive results. Finally, John Moult, who has chaired every CASP since 1994, masterfully delivered a nail-biting exposition of the history of the competition, slowly feeding us information until he finally showed the graph we were all expecting. Here it is:



Combined results of all the CASP competitions. The dark orange line (CASP14_serv) corresponds to the predictions made by fully automated servers, the olive green line (CASP14_w/o427) includes all predictions assisted by humans except for the highest performing group; and the black line (CASP14) represents the predictions by the best performing team: Group 427, or AlphaFold 2. This plot uses the GDT_TS score, where 100 represents perfect results and 0 is a meaningless prediction.

As a rule of thumb, a GDT_TS around 60% represents a "correct fold", meaning that we have an idea of how the protein folds globally; and over 80% we start seeing side chains that closely resemble the model. As you can see, AlphaFold 2 achieves this objective for all but a small percentage of the tasks.

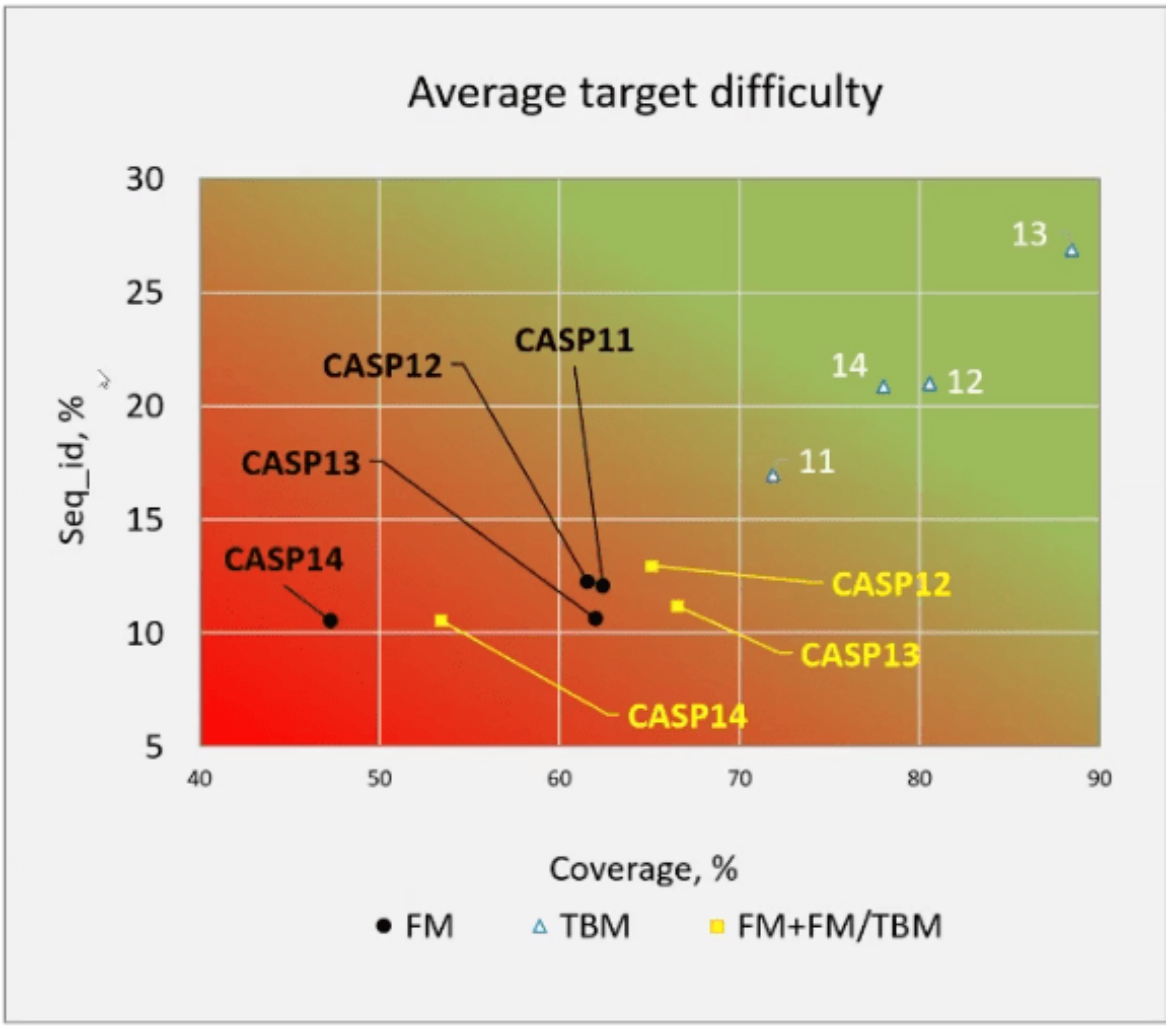
And then, after three decades of competitions, the assessors declared that AlphaFold 2 had succeeded in solving a challenge open for 50 years: to develop a method that can accurately, generally and competitively predict a protein structure from its sequence (or, well, a multiple sequence alignment, as we will see later). There are caveats and edge cases, as in any application — but the magnitude of the breakthrough, as well as its potential impact, are undeniable.

The story does not end there. The models produced by AlphaFold 2 were so good that in some cases defied the results of the experiment. I will provide two brief examples, based on examples mentioned in the conference. The first example comes from the group of Osnat Herzberg, who were studying a phage tail protein. After appreciating the excellent agreement of DeepMind's model with their structure, they noticed that they had

a different assignment for a cis-proline. Upon reviewing the analysis, they realised they had made a mistake in the interpretation and corrected it.

The second comes from the group of Henning Tidow, who was studying an integral membrane protein, the reductase FoxB (apparently related to iron uptake in Gram-negative bacteria). Prof. Tidow's group worked on this model for about two years, trying different methodologies to solve the crystal structure, including experimental phasing methods. When they were given the models from DeepMind's prediction, they managed to solve the problem by molecular replacement in a matter of hours.

There is one last point to clear. Some people have wondered if Google's incredible success is not perhaps related to an easier set of target proteins this year. This is, *per se*, a difficult claim to sustain (after all, wouldn't other groups with much more experience have also benefited from this?), but to disprove this point the assessors have concluded that the targets for CASP14 were the most difficult to date, based on the similarity of existing protein structures to the targets:



Comparison of the targets of the last four CASPs in terms of the coverage and sequence identity of the available templates. On both counts, CASP14 includes the most difficult free-modelling (FM) targets yet provided.

TBM stands for Template Based Models.

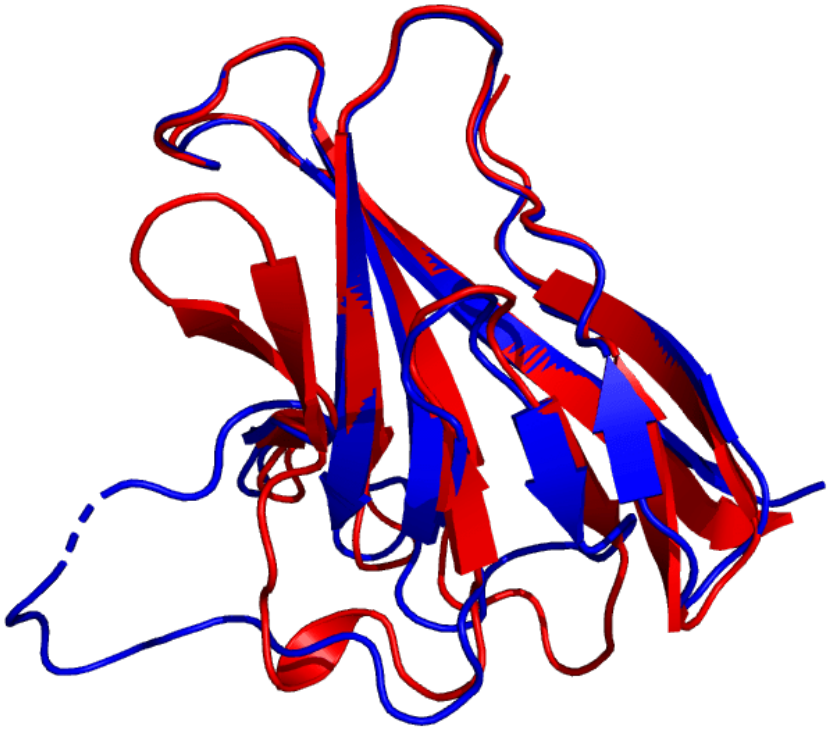
There are still many interesting points of discussion. Many will argue that the set of targets studied at CASP14 is not representative of *all* interesting structural prediction problems — and they will be right. And, yes, certainly there are some problems where AlphaFold 2 hasn't performed that well. I will give you my thoughts on some of the caveats later, when we near the end of this blog post. But, for now, let's be clear on something: AlphaFold 2 is a tool that can solve the protein structure prediction *for a very significant number of targets*.

How does this compare to other methods?

I might have convinced you that AlphaFold 2 is a massive breakthrough. Now it is time that we move from the solid terrain of what the assessors found out, down a gradient of increasing conjecture that explores how and what other groups did, followed by AlphaFold 2's methodology, and then to forecasts of what this might mean for biology, and particularly for bioinformatics.

I am going to have a closer, albeit brief look at two of the targets in the competition, comparing AlphaFold 2 with two of the best ranked groups: David Baker's and Yang Zhang's. They have both (1) consistently performed really well in past CASP competitions, and (2) given fantastic talks this Tuesday, so I have a decent idea of what is happening under the hood.

The first target I am going to look at the ORF8 protein, a viral protein involved with the interaction between SARS-CoV-2 and the immune response (PDB: 7JTL, preprint available on bioRxiv). In CASP14, it was labelled as T1064. Let's have a glance at how the structure predicted by AlphaFold (red) compares to the crystal structure (blue).



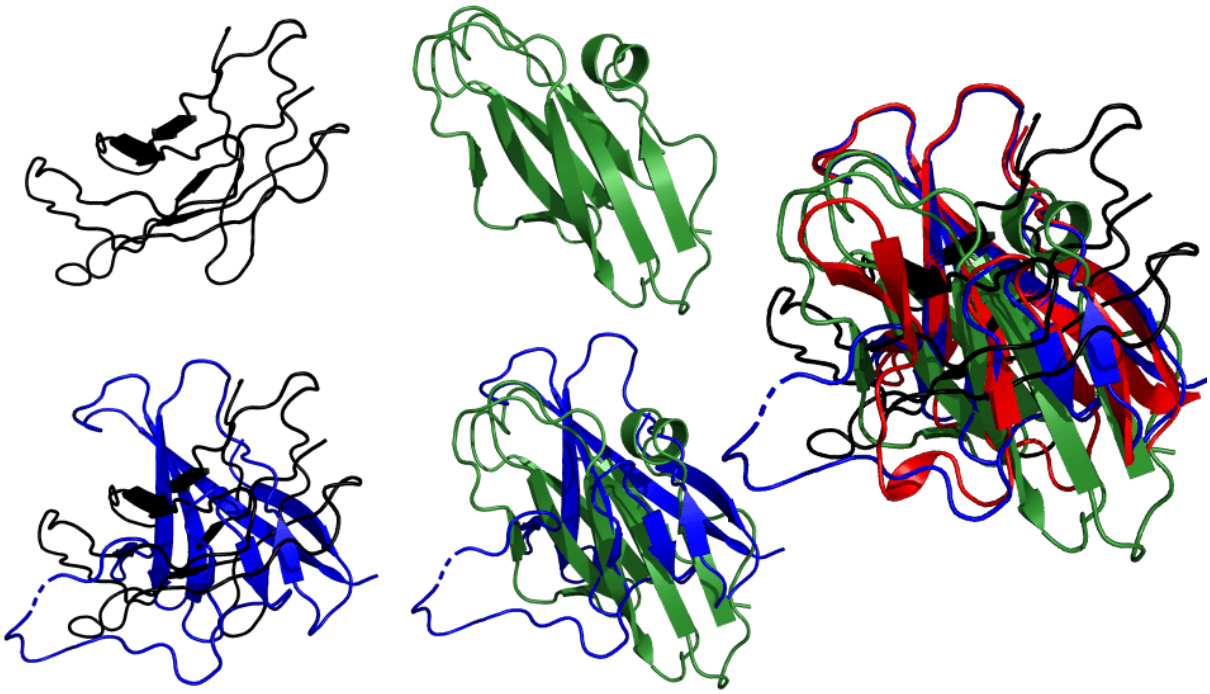
Top group 427 model for the T1064 target (red), superimposed onto the 7JTL_A structure (blue). DeepMind's structure was obtained from the CASP14 webpage on Tuesday 1st December, 2020.

The prediction of the core of the protein is in excellent agreement with the experiment, closely reproducing the structure of the antiparallel β-sheets, and more impressively, the loops that connect them. Remember that loop regions are characterised by a lack of secondary structure, meaning that there is not a scaffold of hydrogen bonds that maintains the structure together, as in α-helices and β-sheets. For this reason, loops are generally considered hard to predict, and compared to usual methods, AlphaFold 2's performance is quite impressive.

Mo's postNotice, however, that there is a large loop region, on the bottom left corner of the image, that is very significantly different from the crustal structure. Besides the overall shape of the loop, the hydrogen bonding pattern is clearly wrong, with PyMOL recognizing a significant portion as a β-sheet. Although this 30-residue long loop region is in the wrong position, the model does bear some resemblance to the loop, and its performance is still better than most common methods. More importantly, since loop regions are commonly very flexible, the failure of the program may just be pointing out that this region is mobile (as happens sometimes in distance prediction models). Furthermore, even with a double-digit percentage of the structure gone wrong, the overall RMSD is barely above 1 Å.

How did the other groups do? Both the Baker and Zhang groups used a similar pipeline, which incorporates

many of the ideas of the CASP13 AlphaFold: build a multiple sequence alignment, potentially incorporating metagenomics sequences; predict a potential using deep learning and find a minimum using their lab-branded method (ROSETTA for Baker group, I-TASSER for Zhang group), and apply some refinement, potentially also using deep learning. I will not discuss the details in depth — look out for their papers in the special issue for CASP14 — instead let's see how they performed (Baker group in green and Zhang group in black):



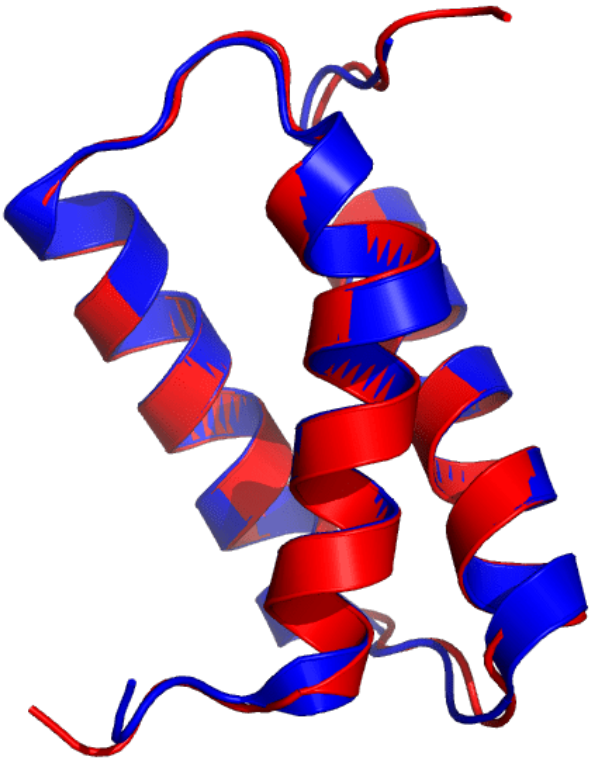
Top: highest-ranked models for the target T1064 submitted by the Zhang (black) and Baker (green) human groups. Bottom: models aligned with the crystal structure. Right: all three models (Zhang, Baker and AlphaFold 2) aligned with the crystal structure. The submissions were obtained from the CASP14 webpage on Tuesday 1st December, 2020.

We can see clear differences between the models and the crystal structure. Both models get the topology of the core wrong: Baker's group shows more β-sheets than the crystal structure, and their topology is wrong, combining parallel and antiparallel sheets; Zhang's group barely captures the structure of the core. In both cases, the loops connecting the β-sheets are all around the place, and the large 30-residue loop region that AlphaFold 2 didn't model correctly is modelled even worse by these two submissions.

Don't get me wrong — this is a difficult target, Baker's and Zhang's work has been excellent, and their predictions would be state-of-the-art in any other CASP. The second best model for this target, by Xianming Pan at Tsinghua, is only slightly better. But something is clear: when we compare against the best performing groups in the protein structure prediction community, AlphaFold 2's accuracy is simply on a whole different level.

While certainly not the worst of their models, the ORF8 protein was highlighted during DeepMind's talk as one of the targets where "they didn't do that well". Well, let's look at some of the models where they actually did re-

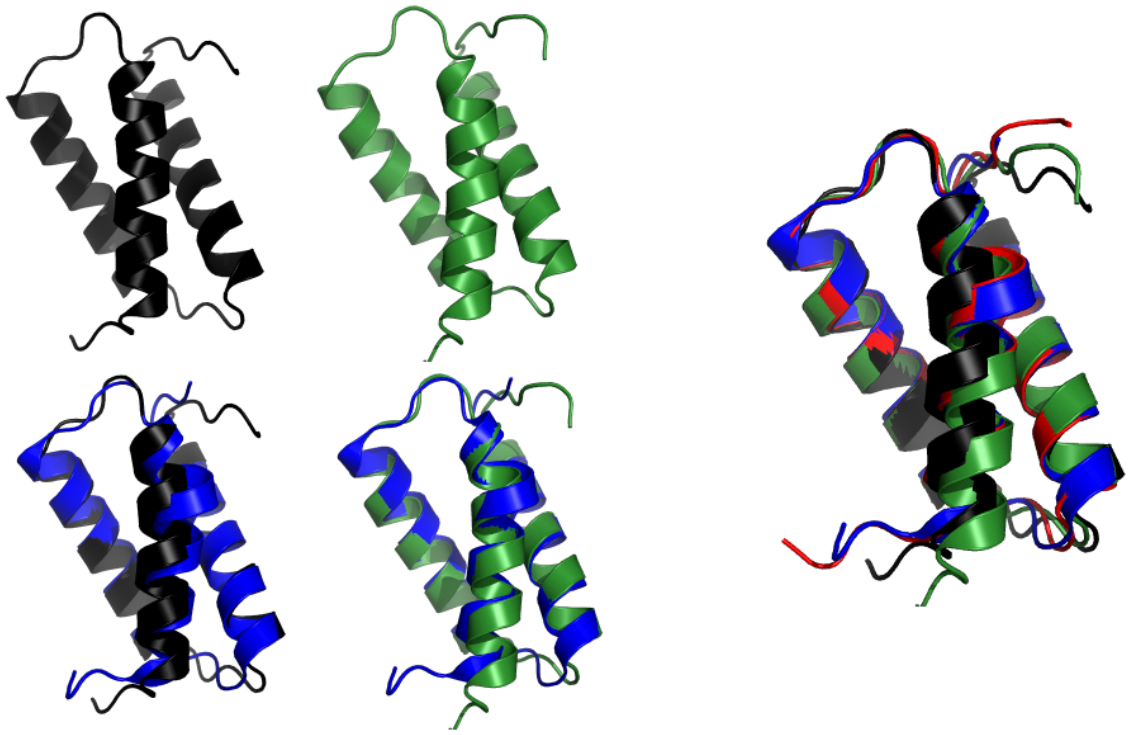
ally well. We are now going to look at target T1046s1 (PDB: 6x6o, chain A).



Top Group 427 model for the T1046s1 target (red), superimposed onto the 6X6O_A structure (blue). DeepMind's structure was obtained from the CASP14 webpage on Tuesday 1st December, 2020.

Here, the AlphaFold model is virtually indistinguisable from the crystal structure, with a total RMSD of 0.48 Å. The α-helices are represented with fantastic accuracy, and in particular the kink of the first α-helix (the one closest to the viewer on the 3D image) is reproduced with outstanding accuracy. As in the previous target, the loops connecting major secondary structure portions are barely distinguisable from the crystal structure. The only areas that show any appreciable differences are the N- and C-termini, and these are very small indeed.

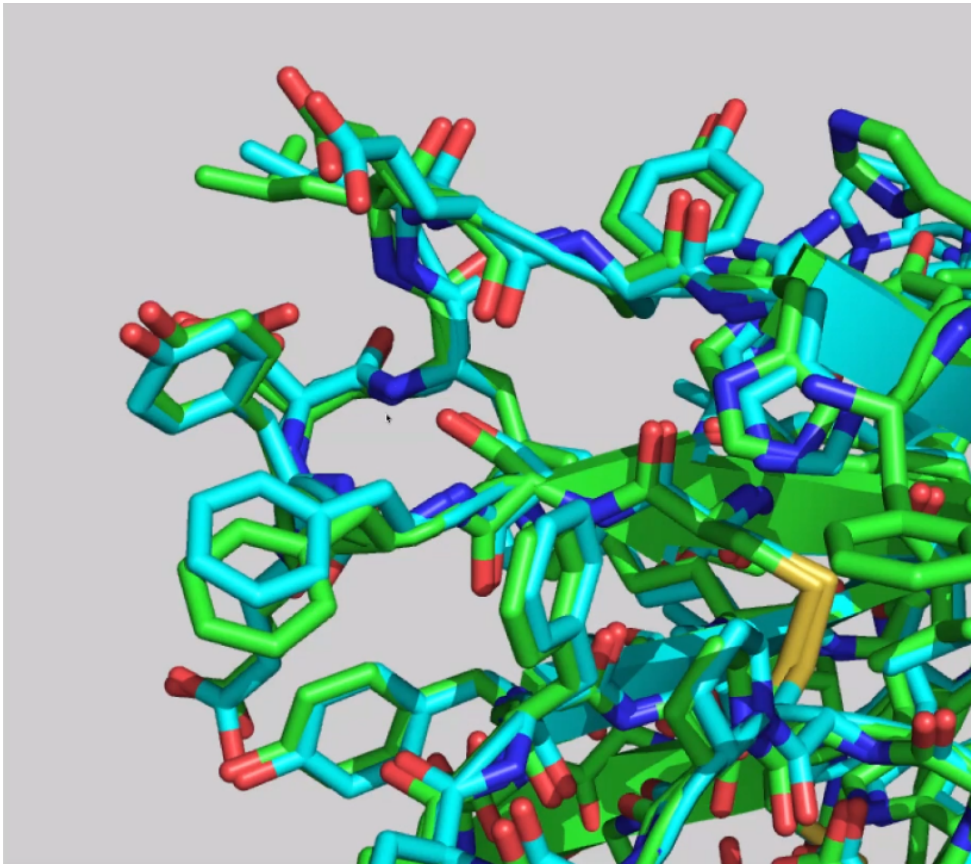
Since this is a relatively simple protein (a small, all-α protein), it should not come as a surprise that both the Baker and the Zhang groups built models that accurately reproduce the fold:



Top: highest-ranked models for the target T1046s1 submitted by the Zhang (black) and Baker (green) human groups. Bottom: models aligned with the crystal structure (blue). Right: all three models (Zhang, Baker and AlphaFold 2, red) aligned with the crystal structure (blue). The submissions were obtained from the CASP14 webpage on Tuesday 1st December, 2020.

These are really good models, and in particular the performance on the loops connecting the α-helices is outstanding. However, a close examination reveals some discrepancies. The kink in the first α-helix is not reproduced accurately: Zhang's group models it as an essentially straight helix, while Baker groups shows a smaller kink; in comparison, AlphaFold 2 modelled the kink to perfect accuracy. Moreover, the magnitude of the deviations is a lot larger than in the AlphaFold 2 model.

One of the assessors, Nick Grishin, summarized this uncanny performance in a quote that goes more or less like this: *"What did AlphaFold 2 get right, that other models did not? The details"*. In fact, the agreement is so good that it extends to side chains:

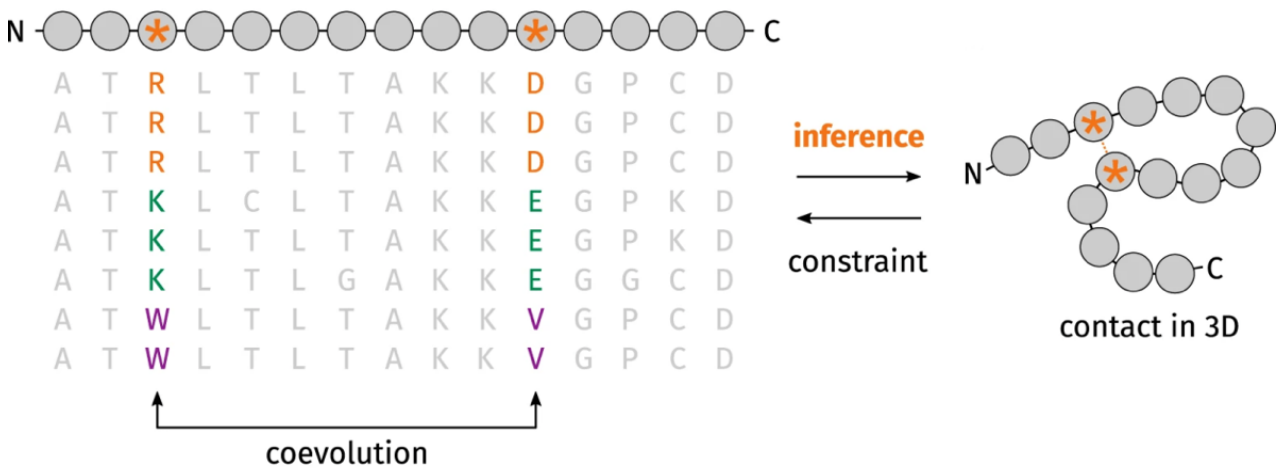


AlphaFold 2 not only predicts the global structure of the protein with high accuracy — it also produces incredibly accurate forecasts of the side chain structure. Image taken from the CASP14 slides.

How did they do it? Part 1: technical details

This is going to be a difficult one. DeepMind's description of their protocol in the CASP14 book of abstracts is sparing in details, and while their talk did unveil some interesting information, much is still unknown. We won't know exactly what they did until they release the corresponding paper, which will take months if not more than a year. However, I can tell you what they have said so far and we can try to speculate what is going on under the hood.

AlphaFold 2 relies, like most modern prediction algorithms, on a multiple sequence alignment (MSA). The sequence of the protein whose structure we intend to predict is compared across a large database (normally something like UniRef, although in later years it has been common to enrich these alignments with sequences derived from metagenomics). The underlying idea is that, if two amino acids are in close contact, mutations in one of them will be closely followed by mutations of the other, in order to preserve the structure.



Schematic of how co-evolution methods extract information about protein structure from a multiple sequence alignment (MSA). Image modified from doi: 10.5281/zenodo.1405369.

Consider the following example. Suppose we have a protein where an amino acid with negative charge (say, glutamate) is near to an amino acid with positive charge (say, lysine), although they are both far away in the amino acid sequence. This Coulombic interaction stabilises the structure of the protein. Imagine now that the first amino acid mutates into a positively charged amino acid — in order to preserve this contact, the second amino acid will be under evolutionary pressure to mutate into a negatively charged amino acid, otherwise the resulting protein may not be able to fold. Of course, real situations are rarely as clear-cut as this example, but you get the idea.

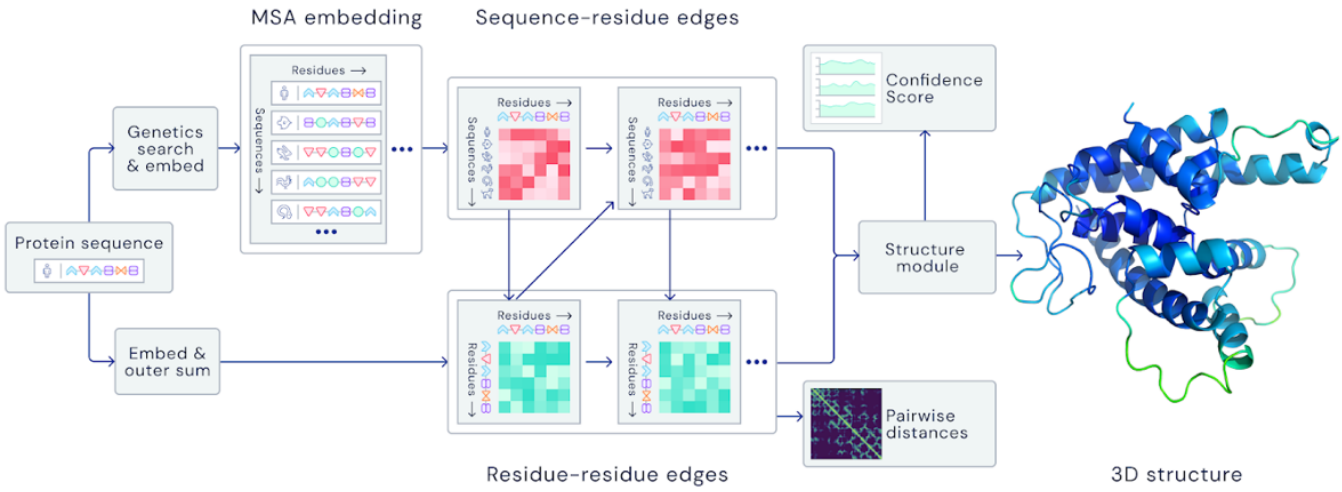
This principle has inspired very many algorithms to predict structural properties of proteins, from contacts to secondary structure. AlphaFold's own success in CASP13 did in fact use deep learning to predict interresidue distances from a MSA (and many other features, that included the output of some co-evolution software). These predictions would then be transformed into a potential, that would be minimised (with a simple gradient descent algorithm like L-BFGS) to find a good structure. This idea has been adopted by many research groups in CASP14, including the very best ones.

This time, however, DeepMind decided to develop an end-to-end model. Instead of using the MSA to predict constraints, they produced a deep learning architecture that takes the MSA as input (plus some template information, but that is another story) and outputs a full structure at the end. Their motivation is simple. Given that the ~170,000 structures available in the PDB constitute a small dataset, they want to make the most of inductive bias — introducing constraints into the model architecture that ensure that the information is assimilated rapidly and efficiently.

To understand what DeepMind's team was after, let us consider for a moment the case of convolutional neural networks, a deep learning architecture behind many successes in computer vision. Many believe that the success of CNNs is due to the way they restrict the flow of information: owing to their design, the information corresponding to a pixel is mixed with its neighbours, and this locality flows throughout the layers, extracting in-

formation from different regions in a hierarchical manner . The network does not need to use a large amount of data or training time to learn that local information is important — instead, this is learned naturally due to the constraints imposed by the architecture.

How they use inductive bias is less clear. We know that the input information is embedded into an embedding space, for which we do not have much information. John Jumper, who represented DeepMind at the conference, explained it "learns sequence-residue edges and residue-residue edges", and mentioned that the model "employs an attention-based network, that identifies which edges are important" (in comparison, the original AlphaFold weighed all distances equally). While we have little information about the actual architecture, we know that an important piece is a 3D equivariant transformer — a novel piece of deep learning architecture widely known for its role in famous models like GPT-3 and BERT — which are in charge of updating the protein backbone and building the side chains.



DeepMind's diagram (taken from their blog) provides an overview of the architecture of AlphaFold 2, but lacks the details that would be required to reproduce it.

The predictive process proceeds in an iterative fashion, "passing information back and forth between the MSA and the internal representation of the protein". I guess this means that the information obtained from a forward pass through the network is somehow fed back to the input features, and then rerun until convergence — but that is, of course, a conjecture. From the graphs shown at the conference, the first predictions are often very good (around 70-80 GDT_TS) and after a few iterations converge to the impressive 90+ GDT_TS predictions that we have seen in CASP14. The final result is not guaranteed to obey all stereochemical constraints, so the final structure is relaxed by coordinate-restrained gradient descent using the Amber ff99SB force field and OpenMM.

There is certainly not enough information to attempt to create a similar model. I suspect the rest of the protein informatics community is experiencing a scientific cliff-hanger, awaiting DeepMind's paper with more enthusiasm than *Cyberpunk 2077*. In the meantime, it is unclear in which direction we are going to work.

How did they do it? Part 2: not-so-technical details

Of course, the success of the DeepMind team is not only related to deep learning. There is more, a lot more.

Many of the factors are reminiscent of Mohammed AlQuraishi's celebrated piece after the last CASP — that DeepMind organised a nimble, well-funded group that could try many ideas quickly and exchange information at a much faster rate than academic groups, that only communicate every two years. I do not wish to discuss this, since I am expecting AlQuraishi to write a similar piece after this CASP (**updated 8th Dec**: here is Mo's post). Instead, I would like to discuss two questions which I think are important not only to understand their success, but also to consider how this success will impact academic computational research: (1) the influence of DeepMind's virtually limitless compute power, and (2) the large amount of data, in terms of structure and methodologies, that has been produced and published by academic research groups.

Let us first talk about computational resources. While John Moult was introducing the impressive performance of AlphaFold 2, and the first press releases were starting to come out, one topic seemed to dominate the CASP14 Discord channel: how many resources went into training this model. DeepMind's blog post states about their model that:

> *It uses approximately 128 TPUv3 cores (roughly equivalent to ~100-200 GPUs) run over a few weeks, which is a relatively modest amount of compute in the context of most large state-of-the-art models used in machine learning today*
>
> *AlphaFold: a solution to a 50-year-old grand challenge in biology, in DeepMind's blog*

Tensor Processing Units (TPUs) are a proprietary Application-Specific Integrated Circuit (ASIC) developed by Google to accelerate the training of neural networks. Unlike GPUs that were originally conceived to process graphics and then repurposed, TPUs have been designed from the ground up for deep learning, and they have been featured in most of DeepMind's recent successes.

There is not a clear equivalence between TPUs and GPUs (as there isn't one between GPUs and CPUs), since performance depends on the problem, but in the right hands they can deliver quite the speedup. Perhaps more importantly, an 8-core TPU v3 chip has 128 GB of vRAM, which is necessary for some architectures that have high memory cost — like attention models. Just for reference, the GPU with the largest RAM I am aware of is the NVIDIA A100, with 40 GB (although an 80 GB version of this GPU has recently been announced). This is quite a difference.

If you think that GPUs are expensive, consider that renting 128 TPUv2 cores has an annual cost of half a million dollars as per Google Cloud's pricing page. Reproducing a replica of DeepMind's experiment using cloud ser-

vices would take anywhere between $25,000 and $200,000, depending on the conditions — and this is, of course, not accounting for the computational effort in exploring the architecture, debugging, optimising the hyperparameters or running several replicas. The total computational cost is likely to be in the region of several million dollars.

This is all very nice, but, how does it compare with the rest of participants? During one of the Q&As, the Baker and Zhang groups said they used roughly 4 GPUs to train their models for a couple of weeks. This means the DeepMind team had roughly two orders of magnitude more computational resources. And, certainly, figures like the one we estimated in the previous paragraph are beyond the capabilities of even the best funded computational research groups.

Is this massive computational power the only factor behind DeepMind's success? I don't think so. The work of this talented team displays novel ideas and creative problem-solving, and the differences cannot be attributed merely to processor muscle. At the same time, it cannot be ignored. The massive compute power means not only that they can work with larger models — they can also achieve a much higher throughput than any academic group would. What the Baker group needed a month to test in their 4 Titan GPUs might only take a few hours for DeepMind, allowing for rapid prototyping and testing of ideas. And, of course, ideas like the architecture that finally resulted in AlphaFold 2 would not even be considered in absence of appropriate hardware.

Looking forward, one can only wonder how this imbalance in resources will impact academic computational research. There is a clear trend of models becoming larger and more complex, and this is happening a lot faster than the decrease in price of the hardware. Unless we come up with a way to ameliorate the need for rapidly increasing computational resources at an affordable price, we might end up in the senseless situation where academic research can not pursue the bold, blue-sky ideas they are supposed to entertain — merely because they are limited to run highly simplified models.

We might of course learn strategies to reduce the impact of limited resources. Tricks like gradient checkpointing, for example, might help reducing the memory footprint. Alternatively, the existence of limitations may well drive our creativity designing other models that achieve similar or better performance at a reduced cost — as the Baker group did with trRosetta, that out-performs CASP13's AlphaFold with a smaller architecture. However, it is clear that those with more computational muscle will always have the upper hand.

This might lead to a future where computational research groups require significant investments in infrastructure to be viable — much like our colleagues in experimental biosciences, albeit with a much faster rate of equipment obsolescence. The success of AlphaFold might convince funding agencies that computational research can do great things with enough resources, and make this possible. Alternatively, we may all have to pool up our resources in a massive international consortium that purchases hardware at scale — in the same way that high-energy physicists had to team up to build massive projects like CERN.

This is starting to get a bit gloomy, so I am going to use the excuse that we are now talking about research funding to stop inadvertently complaining and discuss a different topic. That is, the role played by the vast

amount of data and information about protein structure that has been gathered, mostly by academic research groups, for several decades.

A major piece of DeepMind's success is the availability of techniques, and especially data that has been painstakingly collected by structural biology groups for decades. The Protein Data Bank, which they used for training, accounts for ~170,000 structures, most of them produced by academic groups, as happens with the UniRef protein database, or the BFD and MGnify clusters of metagenomics sequences. The software tools employed, like HHblits, JackHMMER and OpenMM were also developed by government-funded academic initiatives. Also important — most of these initiatives were funded with public money. Big as DeepMind's war chest might be, the taxpayers' investment that has made their achievement possible is several orders of magnitude larger.

This is no less true for the wide body of research about protein structure prediction that is available in peer-reviewed articles, which has been conducted, written and reviewed by academics. This includes many of the ideas that AlphaFold incorporates, from exploiting a multiple sequence alignment to predict protein structure, to incorporating templates into modelling. This is intending in no way to diminish DeepMind's work. They have developed a novel way of tackling protein structure prediction which combines many creative ideas with superb engineering. But, if they have managed to see further, it is because they stood on the shoulders of giants.

This raises many interesting questions about the ethics of research, and of artificial intelligence. Consider, for example, the possibility that Alphabet decides to commercially exploit AlphaFold, for example — is it reasonable that they make profit off such a large body of research paid almost exclusively by the taxpayers? To what extent is the information created by publicly available research — made public, mind you, to stimulate further public research — belong to the public, and under what conditions could it be used in for-profit initiatives? There are many questions to ask if we want to keep science being the open, collaborative machine that it ought to be.

What will this mean for biology?

There are two important questions that are circulating amongst most protein bioinformaticians right now. The first question is: will they [DeepMind] make their code available, and if so, how? And the second, only slightly less important, is: what will it take to run it?

The first question is of utmost importance. When asked about code availability (just over a third of the questions in the virtual [CASP14] Q&A chat-box), John Jumper claimed they were having "internal discussions" at DeepMind, about "making their work available to the community" and that they were hoping to make an announcement in early January.

There are multiple ways that this could happen. Alphabet is ultimately a private, for-profit company, so they may decide to exploit AlphaFold 2 commercially — very much like OpenAI decided to do with GPT-3, the cele-

brated language model unveiled earlier this year. This would also likely mean that the code would remain private, which would quite frankly stall progress in protein informatics, at least for some time. Then there is the possibility that they decide to open source the code — probably with some sort of license for commercial users — which is what everyone is hoping they do after their paper is peer-reviewed, so that the community can build on this incredible success.

Making their code available does not mean that anyone could run it. When their Nature paper was published last January, there was a crucial piece lacking: the code to build the input features to the neural network. While they did provide a description of these features, some OPIGlets and I have been unable to produce a meaningful result despite significant efforts in this direction — and from the discussion in CASP14's Discord channel, it seems many other scientists made similar attempts with equally disappointing results. However, with the architecture it should be possible to retrain the model completely, perhaps by pooling resources from several sources, and deploy a system that while slightly worse than AlphaFold 2, would still be useful for practical applications.

This leads to a related question. We know that DeepMind employed a massive amount of compute power to produce AlphaFold 2, but how long does it actually take to run? When asked how long did it take to train and run their model, John Jumper repeated DeepMind's blog post information about the resources used to produce the final model — but avoided saying how long it took to run the code, and under what conditions. "Several days", as mentioned in the press release, in 128 TPU-v3 cores might very well mean months of computation with the average computational group's resources.

Updated 4th Dec: *Demis Hassabis (CEO of DeepMind) confirmed that the model requires "between a few hours to a few days" on 5-40 GPUs, depending on the protein. This is not very informative, since DeepMind has looked at quite a wide range of lengths and sequence alignment depths, and since we do not know which type of GPUs they used — if they are 40 Titan GPUs, for example, would be a £100,000+ investment in infrastructure. However, even this price is at least an order of magnitude lower than experimental methods, and a lot faster. Things look promising.*

If the code is fast to run, then it could be loaded into an API and be used, much like GPT-3, by anyone with an internet connection. If it requires special hardware, it might be restricted to computational groups that have the acumen to maintain a high-performance computing cluster. My personal intuition is that the latter is more likely, given that the original AlphaFold took several days to run on a general-purpose GPU… and AlphaFold 2 is seemingly a lot bigger than its predecessor. Whatever the cost, it is very likely that will be much faster and cheaper than experimental determination of protein structure, which often takes of the order of years and millions of dollars.

A classical roadblock for drug discovery — assuming that we know a reliable target for the disease, which is an entirely different story — is the lack of reliable crystal structures. If the structure of a target is known, it is possible to design a compound that optimally binds to the active site — a process known as *structure-based drug design* — as well as engineer said molecule to have properties of solubility and low toxicity that makes it a useful drug. Unfortunately, there are entire families of targets — think of G protein-coupled receptors (GPCRs),

whose members are targets of a third of FDA-approved drugs — whose structures are not known accurately. The promise of rapid and accurate protein structure predictions could massively improve the productivity of drug discovery.

Another interesting application of accurate structure predictions will be accurate protein design and engineering. If we can predict the structure of a protein in a short timeframe — days, rather than months or years — we could conceive of an inverse design process by which we define a scaffold of interest (for example, an active center for an enzyme), and try to fine-tune a protein sequence to adopt this structure. Several groups, particularly David Baker's, have been working with similar ideas for some time. This, however, will be contingent on some technical details, for example the ability of AlphaFold to extract information from shallow multiple sequence alignments, as we will briefly discuss in the next section.

Overall, cheap and accurate structural prediction will be a win for biology. Protein function is dependent on structure — the ability to generate structures on-demand promises to further our understanding of life.

What will this mean for structural bioinformatics?

It means we can focus on other problems beyond structural prediction.

Because that is what AlphaFold 2 has solved and not, as many press releases have claimed, the *protein folding problem*. DeepMind's code will provide no information of how a polypeptide, or an ensemble of chains, assembles within seconds into the intricate structure it requires to function. It can just provide an accurate estimation of the crystal structure, which is just a snapshot of the conformation of the protein. But proteins are much more — and the crystal structure doesn't necessarily tell us the whole story (see, *e.g.,* this paper).

More importantly, while AlphaFold 2 provides a *general* solution for protein structure prediction, this does not mean that it is *universal*. Several of the CASP14 targets were not predicted successfully, suggesting that there are some protein families that require further work; and of course, these targets are not fully representative of a proteome. The model was trained on the Protein Data Bank, which has a well-known bias towards proteins that are easy to crystallize. Furthermore, since AlphaFold takes a multiple sequence alignment as input, it remains to be seen if it can tackle problems where these are shallow or not very informative, as happens for example in the very important problem of protein design, in mutated sequences or sometimes in antibody sequences.

Folding is itself a fascinating question, of interest for basic biology but also for biomedicine, where it may better our understanding of the many diseases where misfolding is a cause or a terrible consequence. The success of AlphaFold 2 might provide us with some insight, if we can analyse how the neural network infers the folded structure — but it might also provide very little knowledge, because of difficulty of interpretation or simply because the network's inference is just not very representative of the dynamic folding process.

Protein movement, including flexibility and allostery, is another obvious next step in protein informatics. These mechanisms are fundamental to the way in which proteins function and transmit signals, but the available computational techniques to model these phenomena are still very limited. Dominik Schwarz, one of our OPIGlets, has recently shown that distance predictions by deep learning encode some information about the flexibility of protein regions. It may be that AlphaFold 2 can provide similar insights.

Another area, that I am very interested about, is the study of protein-protein interactions. Think of antibodies, for example: there is a set of interactions between the paratope (antibody portion of the interface) and the epitope (antigen portion) that are fundamental to maintain bindings. Protein-protein docking remains widely unsuccessful despite significant work, and the existence of a CASP-like regular assessment: CAPRI. Lessons learned from AlphaFold 2 may stimulate this area, although we know from CASP14 that it often fails to predict lattice contacts.

Finally, the increased availability of protein structures will only heighten the interest in protein-ligand docking, the prediction of how a ligand will interact with a protein — and how strongly. There has been notable progress in this area, especially with the relative success of novel deep learning methods, although there is still a pervasive lack of well-annotated data that enables major progress in this field.

Conclusion

What Google just achieved might very well be among the most important scientific achievements this century, in terms of impact if not epistemologically. The long sought-after ability to predict the structure of a protein from its sequence (and, as of yet, availability of similar mutated sequences) will unlock applications spanning the entirety of the life and medical sciences, from basic biology to pharmaceutical applications. The prospects are truly astounding.

That said, this statement has to be taken carefully. While we have a *general* solution to the protein structure prediction problem, we do not yet have an *universal* one. Some of the structures in CASP were predicted with low accuracy even by AlphaFold 2, suggesting that further work might be required in particular target families. The Protein Data Bank, which was used for training, displays a well-known bias towards easy to crystallize proteins, and it is unclear how this will affect its usefulness for the dark proteome. Furthermore, since prediction relies on a multiple sequence alignment, it remains to be seen whether this method works when there are few or no sequences in the alignment, as might happen with designed proteins, or when it is not very informative, as in antibodies.

The success of DeepMind also raises a number of points that we, the scientific community, need to consider quite seriously. While nimbler and better funded than most individual research groups, this achievement elicits deep-rooted questions about the way we conduct and communicate research, and whether our community, which collectively has more resources and accumulated knowledge, has really been using their potential efficiently. We also need to reflect on our responsibility as scientists to ensure that science remains open, and that

the research pursued with the support of the public remains useful for the public.

These concerns aside, the solution of the structure prediction problem will finally stimulate novel pathways of research. For too long we have focused on reproducing the static picture of protein structure that we capture through X-ray crystallography. We can now dedicate more efforts to other equally interesting questions: how do proteins fold into these fantastically complicated conformations? How do they move, and how is that movement regulated? How do they interact with other proteins, and with ligands? This is only the start of a very exciting time for protein informatics.

I would like to thank Mark Chonofsky, Fergus Imrie, Constantin Schneider, Javier Pardo Díaz, Matthew Raybould, and Garrett M. Morris, who painstakingly reviewed, identified typos of, and provided invaluable feedback for the first and second drafts of this piece.