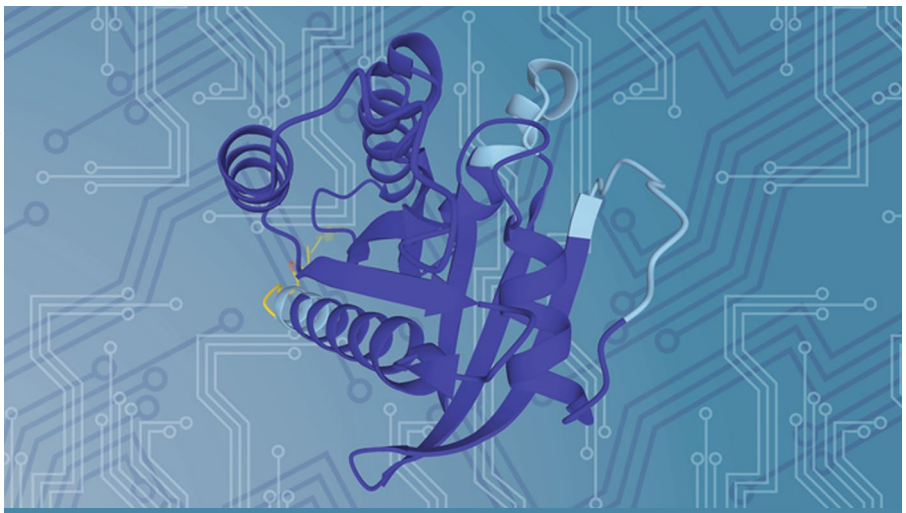

Source: *Nature* **596**, 583–589 (2021)/Springer Nature Limited

What does AlphaFold mean for drug discovery?

AlphaFold and RoseTTAFold have delivered a revolutionary advance for protein structure predictions, but the implications for drug discovery are more incremental. For now.

Asher Mullard

The email came out of the blue, titled only "Zoom invite". But DeepMind's John Jumper breathed a sigh of relief on opening it. After months of stressing about CASP14 — a biennial competition to predict the structure of proteins on the basis of only their amino acid sequences — the results were clear. His group had "performed amazingly well … both relative to other groups and in absolute model accuracy," the conference organizers had written.

CASP, the 'Critical Assessment of Structure Prediction' competition, launched in 1994 as a means of benchmarking the computational prediction of protein structure. Teams have long struggled to achieve decent results. DeepMind, a Google subsidiary, joined the fray in 2018 and bested its peers. In 2020, its 'AlphaFold2' aced the competition.

Due to COVID restrictions, Jumper shared the good news with his team on a video conference call. It was a better one than most. "You could see people's eyes light up," he recalls.

Now, DeepMind has published details of its approach in two papers in *Nature* and released its source code for others to explore. Teaming up with the EMBL-EBI, it has released predicted structures for more than 350,000 proteins for all to access. Inspired by AlphaFold's predictive power, the University of Washington's David Baker and colleagues have developed an alternative called RoseTTAFold that also accurately predicts protein structures.

Drug developers and others are eagerly taking these for a spin.

"There's no question that these approaches have taken a giant leap forward in generating models of protein structures. They enabled everybody to become a structural biologist, which is great fun," says Fiona Marshall, Head of Discovery, Preclinical and Translational Medicine at Merck & Co. and a structure-based drug design pioneer.

These programmes streamline some aspects of the drug discovery workflow, she and others are finding. For instance, they can make it easier to solve structures experimentally — at the front-end by facilitating the design of stable protein constructs that form crystals, and at the back end by helping to make sense of X-ray data. "There's a really nice synergy between the virtual world and the experimental world here," says Marshall.

But uncertainty about the accuracy of the predictions in active sites remains a key limitation, as does the inability to define which conformational state of a protein the programme will predict.

"Where possible, we will still try and get co-crystal structures of ligands bound to proteins to do structure-based drug design," says Marshall. Whereas this was once a 3–4 year undertaking, cryo-EM has helped make this possible within a matter of months for some types of unsolved proteins.

There is plenty of promise for the longer-term future, adds Mark Murcko, board member and strategic advisor to Dewpoint Therapeutics and co-founder of Relay Therapeutics. "AlphaFold has opened a toolbox, and made apparent to the whole world what might be possible," he says. "Now we'll have dozens of labs, each thinking about slightly different problem sets." These include the prediction of protein–ligand structures, the druggability of allosteric pockets, protein–protein interactions and RNA targets, and the design of vaccine immunogens and de novo therapeutic proteins.

"Any success in science immediately makes you think okay, this is great. What do we do next? How do we build on it?," says Murcko.

Pocket predictions

DeepMind's advances were made possible by the Protein Data Bank (PDB), an open access repository of experimentally

> There's no question that these approaches have taken a giant leap forward in generating models of protein structures. They enabled everybody to become a structural biologist, which is great fun

solved structures. Established in 1971, the PDB now includes over 180,000 entries, for around 100,000 unique proteins. DeepMind, like other teams in the CASP competition, used these solved structures as a training set for their deep-learning system. By leveraging both computational expertise and massive computing power, DeepMind developed the pattern-matching neural network system that has now pushed protein structure prediction into the mainstream.

On an amino acid residue level, this means AlphaFold has 'high confidence' in the predicted placement of 36% of the residues in the human proteome, and 'confidence' in the placement of another 22% of the residues. On a per-protein level, it is 'confident' in the prediction of at least 75% of a protein's sequence for 44% of the human proteome.

This is a massive improvement over the status quo. Before AlphaFold, experimental and other modelling approaches provided useful structural insights into 47% of the human proteome, found one analysis. AlphaFold brings this up to 75%. Whereas there used to be 4,832 human proteins for which there was no structural information, AlphaFold lowers this to 29–1,336 perplexing proteins, depending on where the usefulness threshold is set.

"In general, we're pretty excited that we're going to have access to a much larger number of protein structures," says Karen Akinsanya, Chief Biomedical Scientist at Schrödinger.

But drug hunters are particularly concerned with specific parts of proteins — the active and allosteric pockets where small molecules can bind. The question for them, consequently, is how reliable the predictions are for those regions?

"At one level I'm really impressed," says Brian Shoichet, a chemist and virtual drug screener at the University of California, San Francisco. His preliminary comparisons of predicted versus solved structures show that both AlphaFold and RoseTTAFold perform "remarkably well" on the overall folds. "But when it comes to the binding sites, it's more of a mixed bag," he says.

> AlphaFold has opened a toolbox, and made apparent to the whole world what might be possible

Critically, active sites of proteins tend to break the 'protein-folding rules', because they have to be flexible enough to bind one or more ligands that aren't always there. For instance, they are often not as well packed as other parts of a protein. "The reason why these rules can be broken is that the rest of the protein is well folded. And that makes active sites hard for these predictive methods," says Shoichet.

"The community is still at a point where it needs to play with this and see how useful it is," he says. It could be more helpful for some protein families than for others, he adds.

Bryan Roth, a pharmacologist and GPCR expert at University of North Carolina, is even more circumspect. His lab overlayed AlphaFold and RoseTTAFold predictions onto the experimentally solved structures of 20 GPCRs that were not yet in the PDB. "In about 50% of the cases, it was pretty good. And in about 50% of the cases it was not, as far as I can determine, particularly useful," says Roth. "The problem, of course, is that you don't know which 50% your structure is in," he says.

In this analysis, AlphaFold's built-in confidence metric did not predict success.

"For my projects, it's not useful. It's not going to change how we do things," he adds. "Just get a structure. That's what we're doing."

For Roth and Shoichet, prospective validation experiments are needed to gain real insight into the utility of these programmes. To this end, the collaborators plan on running ultra-large computational screens of hundreds of millions of virtual ligands against a set of novel active sites that have been both experimentally solved and computationally predicted. They will then buy the top few hundred hits, test these for activity in the lab, and compare the success rates.

"With retrospective studies, you can convince yourself of anything. You need prospective testing," says Shoichet.

Drug hunters eyeing novel targets that have not been solved structurally will likely take a more project-based approach. AlphaFold predictions offer "a perfectly reasonable starting point," says Murcko, so long as teams can feed in other structural insights and make the most of physics-based molecular dynamic models to refine the details of an active site before progressing.

"Personally I think that it's a great way to start. What we'll have to learn, just from practicing, is where the structures are the most and the least accurate," he adds.

> The community is still at a point where it needs to play with this and see how useful it is

Moving parts

In part, the utility of the predictions will depend on the conformational state of the structures they generate. After all, proteins are in motion in the cell. Both individual X-ray structures and predictions, however, capture just one snapshot of the possibilities.

At companies such as Relay, researchers are laser-focused on understanding how protein motion can create drug discovery opportunities. Predictive algorithms that can reliably facilitate this work would speed things up.

"Imagine having the ability to not just say 'here's how this protein folds', but also 'pay very close attention to this amino acid, which is near the catalytic site and can exist in several different conformations," says Murcko. "That would be the next level of development for these algorithms."

AlphaFold and RoseTTAFold are not there. Rather, they can't yet even discriminate between 'active' and 'inactive' conformations of a protein. Instead, "the current version of AlphaFold gives you the structure that it believes is the most likely one to appear in PDB," explains Jumper.

This can be a problem when it comes to virtual screening efforts, he adds. Whereas researchers want to run docking experiments on the active, ligand-bound conformation of a protein, the current version of AlphaFold often returns an inactive state with an empty pocket and misaligned side chains.

"Our feeling is [molecular docking] might be worth a try if you have a very high-confidence prediction, but in general we expect there will need to be more tool development before this can be a reliable procedure," said DeepMind's Kathryn Tunyasuvunakool in a recent EMBL-EBI webinar on how to interpret AlphaFold structures.

Proteins at the extreme end of the conformational scale — those with floppy intrinsically disordered regions that can adopt multitudes of shapes — are even more troublesome. These regions can have important activity, including in the formation and dissolution of transient membraneless organelles called biomolecular condensates.

> For my projects, it's not useful. It's not going to change how we do things. Just get a structure. That's what we're doing

But AlphaFold scores particularly poorly on predicting the structures of these regions, with good reason given how hard they are to characterize with traditional crystallographic methods.

At Dewpoint, a biotech company at the forefront of biomolecular condensate biology, researchers are exploring whether they can turn this weakness into a strength. 'Low-confidence' predictions might be useful for pinpointing the location of intrinsically disordered regions, a key step to studying their biology. "It's like an independent check step on what regions of a protein are actually disordered," says Murcko. "Dewpoint is especially interested in understanding what this methodology teaches us about disordered regions of proteins."

Others are too. Sometimes AlphaFold can identify functional sites in intrinsically disordered regions and predict their complex structures, noted the EMBL's Bálint Mészáros on the EMBL-EBI webinar.

The longer wish list

Better active-site accuracy and conformational control would certainly make these predictive programmes more useful for drug hunters. But deep learning enthusiasts hope that these systems could one day be even more transformative. For instance, perhaps they can be used to one day reliably predict the structures of protein–ligand interactions.

"If we can crack the protein–ligand problem, that changes the world of drug discovery, instantly," says Murcko. "It would be time well spent."

The hurdles here are massive.

For one, AlphaFold was trained on over 170,000 structures in the Protein Data Bank. But many of these are not bound to physiologically relevant ligands. A reduction in the size of the training set will reduce the accuracy of the resulting predictions.

More problematically, proteins are built with the standard 20 amino acids over and over again — with defined atomistic interactions that can take place between the different building blocks. Small-molecule space, by contrast, is an enormous 10^{60} molecules, filled with galaxies of uncharted atomistic possibilities.

"You have to train a machine learning model to predict the atomistic interactions between all of those ligands — or a sizeable portion of those — and a protein at its many binding sites. I think it's an enormous problem," says Akinsanya.

Further complicating matters, proteins can bind multiple ligands. So how much of their binding potential has to be captured in a training set? Is one protein–ligand structure sufficient, or do individual pockets need to be solved bound to tens, hundreds or even thousands of different ligands? Similarly, is that depth of data needed across hundreds or thousands of proteins?

"We're just underpowered," says Shoichet. "It's really difficult for me to imagine getting to a point where we have enough observations."

Ever pragmatic, Murcko is focused on the next steps. "The question is, what can be done to accelerate the deposition of additional data?," he asks.

If industry groups could pool their structural data together, that might facilitate progress. Those at the forefront of structure-based drug design are amassing troves of X-ray data that are not entered into the PDB. Boehringer Ingelheim, in its hunt for KRAS inhibitors, has now solved 580 structures of KRAS bound to different ligands, for example. "We are getting co-crystals solved within an hour or two," says Darryl McConnell, Research Site Head of Boehringer Ingelheim in Austria, where he is developing an 'X-ray first' approach to medicinal chemistry.

This dataset is a competitive advantage for BI in the KRAS field, but large collections of solved ligand-bound structures from legacy programmes are locked away in internal databases throughout industry.

"This is an opportunity," says McConnell. "Maybe there's a need for a PDB plus."

Baker hopes such an effort can be pulled off. "If these datasets were made available, then there very well could be enough data to solve the protein–ligand problem," he says.

This dataset might also help future algorithms to better predict the ligand-bound conformations of proteins of interest.

> Maybe there's a need for a PDB plus

Other features on industry's wish list include the capacity to predict the structures of protein–protein, protein–DNA and protein–RNA complexes, as well as the ability to forecast the effects of point mutations.

Some of these may come sooner than others. AlphaFold was not explicitly trained to predict the structures of protein–protein interactions, but Baker's work shows that this is possible. As a case study for RoseTTAFold, his team predicted a structure of the IL-12 cytokine bound to the IL-12 receptor, a key step towards identifying ways to intervene in the interaction.

Small-molecule drug hunters and de novo therapeutic protein designers may able to make use of these insights. "We're certainly doing a lot of that now," says Baker, a leading de novo protein designer.

A few years ago Baker's lab designed a de novo mimetic of IL-2/IL-15, which licensee Neoleukin Therapeutics has since advanced into the clinic. But the identification of de novo candidates that fold up as expected remains a rate-limiting research step. His team is assessing whether the new fold-predicting algorithms can lower this experimental overhead. "Ask me in a few months. But we are expecting a significant step up in success rates," he says.

Deep questions

DeepMind, having set these possibilities in motion, has yet to disclose the research or business plans for its deep-learning system. But it is considering its options. "A lot of our time has really been focused on getting this out. But we're taking stock now, and trying to get really situated on where we're going and future directions," says Jumper. "We're not packing our toys up and going home."

Pushmeet Kohli, head of research at DeepMind, adds that there might already even be enough data to take on some of the more ambitious applications. "What amount of data is needed is sort of a tricky question, because it depends on your machine learning model."

If future models can be taught physics and chemistry, they may yet offer even bigger structural biology benefits.

This opportunity, combined with the pace of progress, is fueling optimism even from the experimentalists. "It's exciting to see how fast the field is moving forward," says Marshall. "Given the trajectory from the first AlphaFold to AlphaFold2, I expect to see a lot of rapid development over the next 2–5 years."