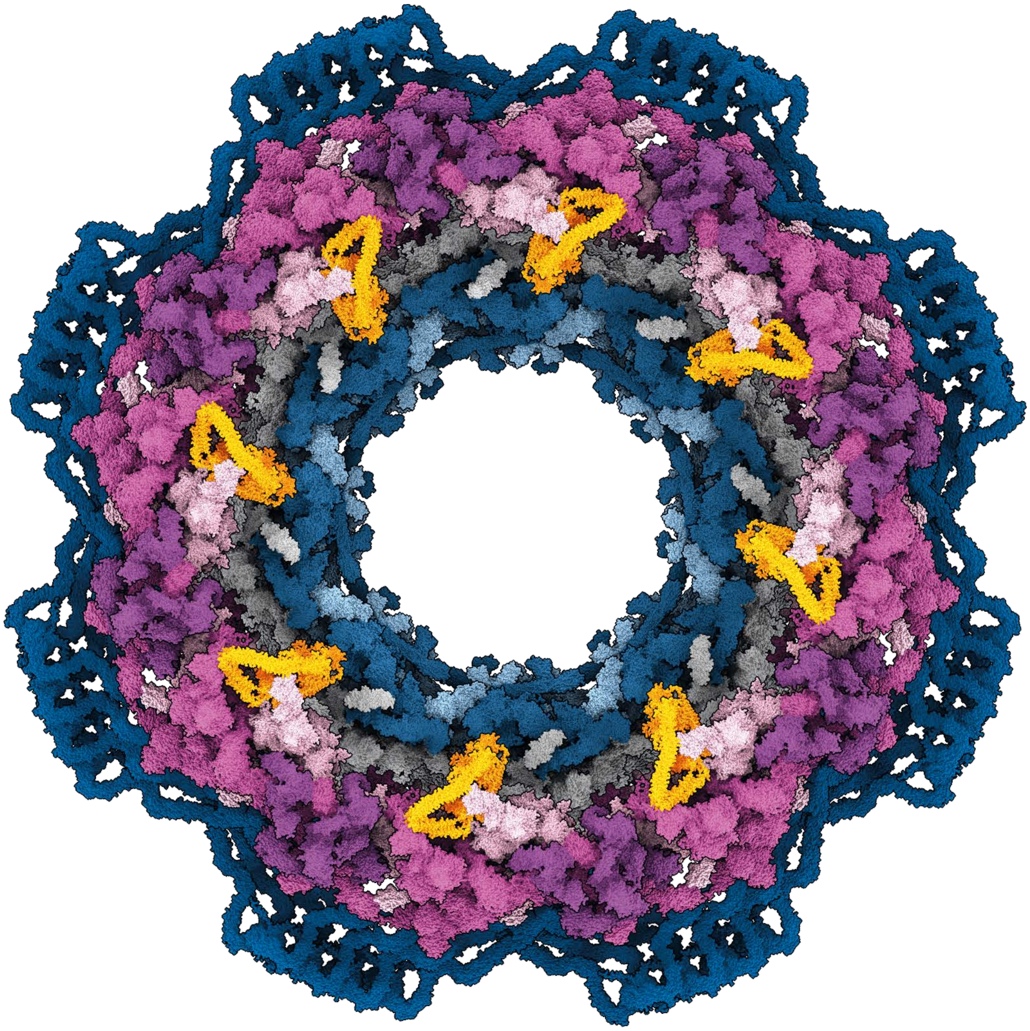


ADAPTED FROM REF. 3/ AGNIESZKA OBARSKA-KOSINSKA

A top-down view of the human nuclear pore complex, the largest molecular machine in human cells.

WHAT'S NEXT FOR THE AI PROTEIN-FOLDING REVOLUTION

AlphaFold, software that can predict the 3D shape of proteins, is already changing biology.

By Ewen Callaway

For more than a decade, molecular biologist Martin Beck and his colleagues have been trying to piece together one of the world's hardest jigsaw puzzles: a detailed model of the largest molecular machine in human cells.

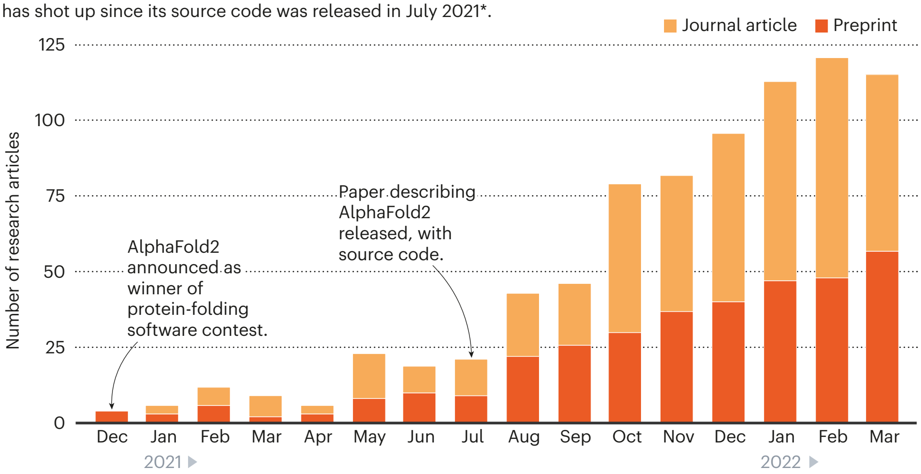
This behemoth, called the nuclear pore complex, controls the flow of molecules in and out of the nucleus of the cell, where the genome sits. Hundreds of these complexes exist in every cell. Each is made up of more than 1,000 proteins that together form rings around a hole through the nuclear membrane.

These 1,000 puzzle pieces are drawn from more than 30 protein building blocks that interlace in myriad ways. Making the puzzle even harder, the experimentally determined 3D shapes of these building blocks are a potpourri of structures gathered from many species, so don't always mesh together well. And the picture on the puzzle's box – a low-resolution 3D view of the nuclear pore complex – lacks sufficient detail to know how many of the pieces precisely fit together.

In 2016, a team led by Beck, who is based at the Max Planck Institute of Biophysics (MPIB)

ALPHAFOLD MANIA

The number of research papers and preprints citing the AlphaFold2 AI software has shot up since its source code was released in July 2021*.



**Nature* analysis using Dimensions database; removing duplicate preprints and papers/R. Van Noorden, E. Callaway.

in Frankfurt, Germany, reported a model[1] that covered about 30% of the nuclear pore complex and around half of the 30 building blocks, called Nup proteins.

Then, last July, London-based firm DeepMind, part of Alphabet – Google's parent company – made public an artificial intelligence (AI) tool called AlphaFold[2]. The software could predict the 3D shape of proteins from their genetic sequence with, for the most part, pinpoint accuracy. This transformed Beck's task, and the studies of thousands of other biologists (see 'AlphaFold mania').

"AlphaFold changes the game," says Beck. "This is like an earthquake. You can see it everywhere," says Ora Schueler-Furman, a computational structural biologist at the Hebrew University of Jerusalem in Israel, who is using AlphaFold to model protein interactions. "There is before July and after."

Using AlphaFold, Beck and others at the MPIB – molecular biologist Agnieszka Obarska-Kosinska and a group led by biochemist Gerhard Hummer – as well as a team led by structural modeller Jan Kosinski, at the European Molecular Biology Laboratory (EMBL) in Hamburg in Germany, could predict shapes for human versions of the Nup proteins more accurately. And by taking advantage of a tweak that helped AlphaFold to model how proteins interact, they managed to publish a model last October that covered 60% of the complex[3]. It reveals how the complex stabilizes holes in the nucleus, as well as hinting at how the complex controls what gets in and out.

In the past half-year, AlphaFold mania has gripped the life sciences. "Every meeting I'm in, people are saying 'why not use AlphaFold?'," says Christine Orengo, a computational biologist at University College London.

In some cases, the AI has saved scientists time; in others it has made possible research that was previously inconceivable or wildly impractical. It has limitations, and some scientists are finding its predictions to be too unreliable for their work. But the pace of experimentation is frenetic.

Even those who developed the software are struggling to keep up with its use in areas ranging from drug discovery and protein design to the origins of complex life. "I wake up and type AlphaFold into Twitter," says John Jumper, who leads the AlphaFold team at DeepMind. "It's quite the experience to see everything."

A startling success

AlphaFold caused a sensation in December 2020, when it dominated a contest called the Critical Assessment of Protein Structure Prediction, or CASP. The competition, held every two years, measures progress in one of biology's grandest challenges: determining the 3D shapes of proteins from their amino-acid sequence alone. Computer-software entries are judged against structures of the same proteins determined using experimental methods such as X-ray crystallography or cryo-electron microscopy (cryo-EM), which fire X-rays or electron beams at proteins to build up a picture of their shape.

The 2020 version of AlphaFold was the software's second edition. It had also won the 2018 CASP, but its earlier efforts mostly weren't good enough to stand in for experimentally determined structures, says Jumper. However, AlphaFold2's predictions were, on average, on par with the empirical structures.

It wasn't clear when DeepMind would make the software or its predictions widely available, so researchers used information from a public talk by Jumper, and their own insights, to develop their own AI tool, called RoseTTAFold.

Then on 15 July 2021, papers describing RoseTTAFold and AlphaFold2 appeared[2,4], along with freely available, open-source code and other information needed for specialists to run their own versions of the tools. A week later, DeepMind announced that it had used AlphaFold to predict the structure of nearly every protein made by humans, as well as the entire 'proteomes' of 20 other widely studied organisms, such as mice and the bacterium *Escherichia coli* – more than 365,000 structures in total (see 'What's known about proteomes'). DeepMind also publicly released these to a database maintained by the EMBL's European Bioinformatics Institute (EMBL–EBI), in Hinxton, UK. That database has since swelled to almost one million structures.

"People are solving structures that, for years, had not been solved."

This year, DeepMind plans to release a total of more than 100 million structure predictions. That is nearly half of all known proteins – and hundreds of times more than the number of experimentally determined proteins in the Protein Data Bank (PDB) structure repository.

AlphaFold deploys deep-learning neural networks: computational architectures inspired by the brain's neural wiring to discern patterns in data. It has been trained on hundreds of thousands of experimentally determined protein structures and sequences in the PDB and other databases. Faced with a new sequence, it first looks for related sequences in databases, which can identify amino acids that have tended to evolve together, suggesting they're close in 3D space. The structure of existing related proteins provides another way to estimate distances between amino-acid pairs in the new sequence.

AlphaFold iterates clues from these parallel tracks back and forth as it tries to model the 3D positions of amino acids, continually updating its estimate. Specialists say the software's application of new ideas in machine learning research seems to be what makes AlphaFold so good – in particular, its use of an AI mechanism termed 'attention' to determine which amino-acid connections are most salient for its task at any moment.

The network's reliance on information about related protein sequences means that AlphaFold has some limitations. It is not designed to predict the effect of mutations, such as those that cause disease, on a protein's shape. Nor was it trained to determine how proteins change shape in the presence of other interacting proteins, or molecules such as drugs. But its models come with scores that gauge the network's confidence in its prediction for each amino-acid unit of a protein – and researchers are tweaking AlphaFold's code to expand its capabilities.

By now, more than 400,000 people have used the EMBL-EBI's AlphaFold database, according to DeepMind. There are also AlphaFold 'power users': researchers who've set up the software on their own servers or turned to cloud-based versions of AlphaFold to predict structures not in the EMBL-EBI database, or to dream up new uses for the tool.

Solving structures

Biologists are already impressed with AlphaFold's ability to solve structures. "Based on what I've seen so far, I trust AlphaFold quite a lot," says Thomas Boesen, a structural biologist at Aarhus University in Denmark. The software has successfully predicted the shapes of proteins that Boesen's centre has determined but not yet published. "That's a big validation from my side," he says. He and Aarhus microbial ecologist Tina Šantl-Temkiv are using AlphaFold to model the structure of bacterial proteins that promote the formation of ice – and which could contribute to the cooling effects of ice in clouds – because biologists haven't been able to fully determine the structures experimentally[5].

As long as a protein curls up into a single well-defined 3D shape – and not all do – AlphaFold's prediction can be hard to beat, says Arne Elofsson, a protein bioinformatician at Stockholm University. "It's a one-click solution to get probably the best model you're going to get."

Where AlphaFold is less confident, "it's very good at telling you when it doesn't work", Elofsson says. In such cases, predicted structures can resemble floating spaghetti strands (see 'The good, the bad and the ugly'). This often corresponds to regions of proteins that lack a defined shape, at least in isolation. Such intrinsically disordered regions – which make up around one-third of the human proteome – might become well defined only when another molecule, such as a signalling partner, is present.

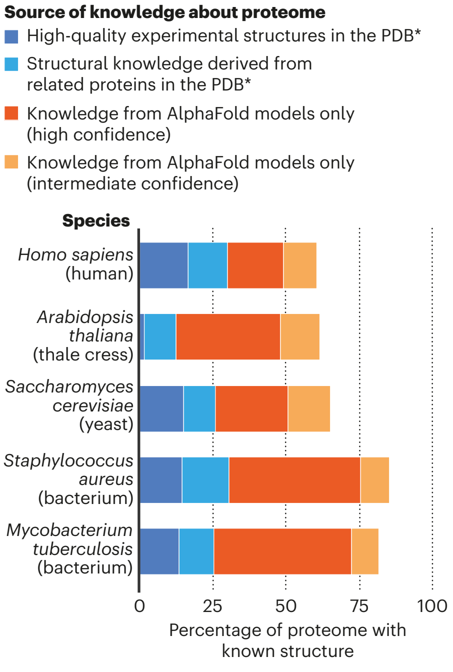
Norman Davey, a computational biologist at the Institute of Cancer Research in London, says AlphaFold's ability to identify disorder has been a game-changer for his work studying the properties of these regions. "Instantly there was a huge increase in the quality of the predictions we had, without any effort on our part," he says.

AlphaFold's dump of protein structures into the EMBL-EBI database is also immediately being put to use. Orengo's team is searching it to identify fresh kinds of proteins (without experimentally verifying them) and has turned up hundreds, perhaps thousands, of potentially new protein families, expanding scientists' knowledge of what proteins look like and can do. In another effort, the team is scouring databases of DNA sequences harvested from the ocean and waste water, to try to identify new plastic-eating enzymes. Using AlphaFold to quickly approximate the structures of thousands of proteins, the researchers hope to better understand how enzymes evolved to break down plastic, and potentially to improve them.

The ability to transform any protein-coding gene sequence into a reliable structure should be especially powerful for evolution studies, says Sergey Ovchinnikov, an evolutionary biologist at Harvard University in Cambridge, Massachusetts. Researchers compare genetic sequences to determine how organisms and their genes are related across species. For distantly related genes, comparisons might fail to turn up evolutionary relatives because the sequences have changed so much. But by comparing protein structures – which tend to change less rapidly than genetic sequences – researchers might be able to uncover overlooked ancient relationships. "This opens up an amazing opportunity to study the evolution of proteins and the origins of life," says Pedro Beltrao, a computational biologist at the Swiss Federal Institute of Technology in Zurich.

WHAT'S KNOWN ABOUT PROTEOMES

AlphaFold's predictions have greatly increased the proportion of confidently known structures in the human proteome — the collection of all human proteins. The software is even more useful for other species.



*PDB: Protein Data Bank. AlphaFold can also be used to calculate these structures — but doesn't add significantly to what's already known.

E. PORTA-PARDO *ET AL. PLOS COMPUT. BIOL.* **18**, E1009818 (2022).

> "Because it looks nice doesn't mean it is correct. You need some experimental data that show you're right."

To test this idea, a team led by Martin Steinegger, a computational biologist at Seoul National University, and his colleagues used a tool they developed, called Foldseek, to look for relatives of the RNA-copying enzyme of SARS-CoV-2 – the virus that causes COVID-19 – in the EMBL-EBI's AlphaFold database[6]. This search turned up previously unidentified possible ancient relatives: proteins across eukaryotes – including slime moulds – that resemble, in their 3D structure, enzymes called reverse transcriptases that viruses such as HIV use to copy RNA into DNA, despite very little similarity at the genetic-sequence level.

Experimental assistant

For scientists who want to determine the detailed structure of a specific protein, an AlphaFold prediction isn't necessarily an immediate solution. Rather, it provides an initial approximation that can be validated or refined by experiment – and which itself helps to make sense of experimental data. Raw data from X-ray crystallography, for instance, appear as patterns of diffracted X-rays. Typically, scientists need a starting guess at a protein's structure to interpret these patterns. Previously, they'd often cobble together information from related proteins in the PDB or use experimental approaches, says Randy Read, a structural biologist at the University of Cambridge, UK, whose lab specialized in some of these methods. Now, AlphaFold's predictions have rendered such approaches unnecessary for most X-ray patterns, Read says, and his lab is working to make better use of AlphaFold in experimental models. "We've totally refocused our research."

He and other researchers have used AlphaFold to determine crystal structures from X-ray data that were uninterpretable without an adequate starting model. "People are solving structures that, for years, had not been solved," says Claudia Millán Nebot, a former postdoc in Read's lab who now works at the analytics firm SciBite in Cambridge. She expects to see a glut of new protein structures submitted to the PDB, in large part as a result of AlphaFold.

The same is true for labs specializing in cryo-EM, which captures pictures of flash-frozen proteins. In some instances, AlphaFold's models have accurately predicted unique features of proteins called G-protein-coupled receptors (GPCRs) – which are important drug targets – that other computational tools got wrong, says Bryan Roth, a structural biologist and pharmacologist at the University of North Carolina at Chapel Hill. "It seems to be really good for generating first models, which we then refine with some experimental data," he says. "That saves us some time."

But Roth adds that AlphaFold isn't always that accurate. Of the several dozen GPCR structures his lab has solved, but not yet published, he says, "about half the time, the AlphaFold structures are fairly good, and half the time they're more or less useless for our purposes". In some instances, he says, AlphaFold labels predictions with high confidence, but experimental structures show that it is wrong. Even when the software gets it right, it cannot model how a protein would look when bound to a drug or other small molecule (ligand), which can substantially alter the structure. Such caveats make Roth wonder how useful

AlphaFold will be for drug discovery.

It's increasingly common in drug-discovery efforts to use computational-docking software that screens billions of small molecules to find some that might bind to proteins – one indication that they could make useful drugs. Roth is now working with Brian Shoichet, a medicinal chemist at the University of California, San Francisco, to see how AlphaFold's predictions compare with experimentally determined structures in this exercise.

Shoichet says they are limiting their work to proteins for which AlphaFold's prediction chimes with experimental structures. But even in these instances, the docking software is turning up different drug hits for the experimental structure and AlphaFold's take, suggesting that small discrepancies could matter. "That doesn't mean we won't find new ligands, we'll just find different ones," says Shoichet. His team is now synthesizing potential drugs identified using AlphaFold structures, and testing their activity in the lab.

Critical optimism

Researchers at pharmaceutical companies and biotechnology firms are excited about AlphaFold's potential to help with drug discovery, says Shoichet. "Critical optimism is how I'd describe it." In November 2021, DeepMind launched its own spin-off, IsoMorphic Labs, which aims to apply AlphaFold and other AI tools to drug discovery. But the company has said little else about its plans.

Karen Akinsanya, who leads therapeutics development at Schrödinger, a drug-discovery firm headquartered in New York City that also publishes chemical-simulations software, says she and her colleagues are already having some success using AlphaFold structures, including for GPCRs, in virtual screens and compound design for drug candidates. She finds that, just as with experimental structures, extra software is needed to get at the fine details of amino-acid side chains or locations where individual hydrogen atoms might sit. Once this is done, AlphaFold structures have proved good enough to guide drug discovery – in some cases.

"It's hard to say 'this is a panacea'; that because you can do it very well for one structure – surprisingly and excitingly well – that it is eminently applicable to all structures. It clearly isn't," Akinsanya says. And she and her colleagues have found that AlphaFold's accuracy predictions don't show whether a structure will be useful for later drug screening. AlphaFold structures will never fully replace experimental ones in drug discovery, she says. But they might speed up the process by complementing experimental methods.

Drug developers curious about AlphaFold received good news in January, when DeepMind lifted a key restriction on its use for commercial applications. When the company released AlphaFold's code in July 2021, it had stipulated that the parameters, or weights, needed to run the AlphaFold neural network – the end result of training the network on hundreds of thousands of protein structures and sequences – were for non-commercial use only. Akinsanya says this was a bottleneck for some in industry, and there was a "wave of excitement" when DeepMind changed tack. (RoseTTAFold came with similar restrictions, says Ovchinnikov, one of its developers. But the next version will be fully open-source.)

AI tools are not just changing how scientists determine what proteins look like. Some researchers are using them to make entirely new proteins. "Deep learning is completely transforming the way that protein design is being done in my group," says David Baker, a biochemist at the University of Washington in Seattle and a leader in the field of designing proteins, as well as predicting their structures. His team, with computational chemist Minkyung Baek, led the work to develop RoseTTAFold.

Baker's team gets AlphaFold and RoseTTAFold to "hallucinate" new proteins. The researchers have altered the AI code so that, given random sequences of amino acids, the software will optimize them until they resemble something that the neural networks recognize as a protein (see 'Dreaming up proteins').

In December 2021, Baker and his colleagues reported expressing 129 of these hallucinated proteins in bacteria, and found that about one-fifth of them folded into something resembling their predicted shape[7]. "That's really the first demonstration that you can design proteins using these networks," Baker says. His team is now using this approach to design proteins that do useful things, such as catalyse a particular chemical reaction, by specifying the amino acids responsible for the desired function and letting the AI dream up the rest.

Hacking AlphaFold

When DeepMind released its AlphaFold code, Ovchinnikov wanted to better understand how the tool worked. Within days, he and computational-biology colleagues, including Steinegger, set up a website called ColabFold that allowed anyone to submit a protein sequence to AlphaFold or RoseTTAFold and get a structure prediction. Ovchinnikov imagined that he and other scientists would use ColabFold to try and 'break' AlphaFold, for instance, by supplying false information about a target protein sequence's evolutionary relatives. By doing this, Ovchinnikov hoped he could determine how the network had learnt to predict structures so well.

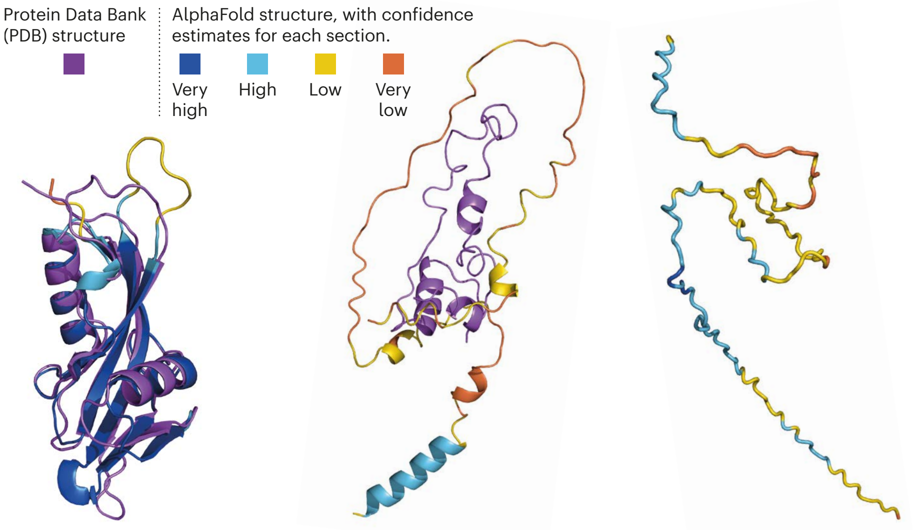
As it turned out, most researchers who used ColabFold just wanted to get a protein structure. But others used it as a platform to modify the inputs to AlphaFold to tackle new applications. "I didn't expect the number of hacks of various types," says Jumper.

By far the most popular hack has been to wield the tool on protein complexes comprised of multiple, interacting – and often intertwined – chains of peptides. Just as with the nuclear pore complex, many proteins in cells gain their function when they form complexes with multiple protein subunits.

AlphaFold was designed to predict the shape of single peptide chains, and its training consisted entirely of such proteins. But the

THE GOOD, THE BAD AND THE UGLY

AlphaFold's predictions of a folded protein's structure come with confidence estimates. Superimposing each model on the experimentally determined structure (if available) shows the accuracy of the prediction.


Protein Data Bank (PDB) structure
AlphaFold structure, with confidence estimates for each section.
Very high
High
Low
Very low

Good
AlphaFold model of phosphohistidine phosphatase overlaps closely with PDB structure.

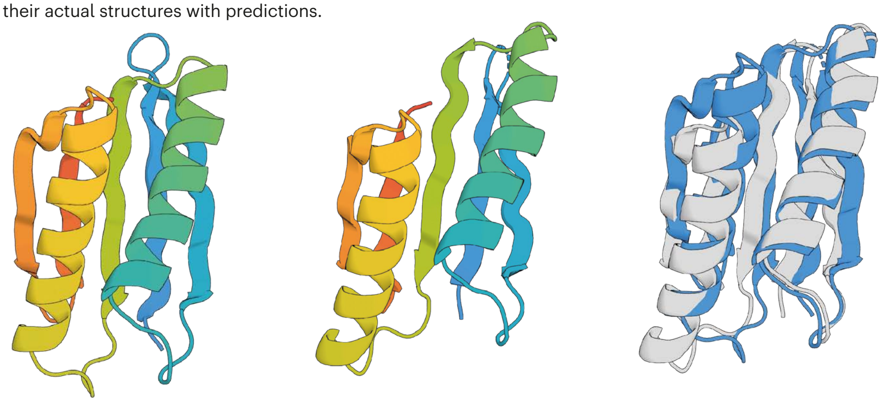
Bad
AlphaFold model of human insulin bears no relation to the PDB structure.

Ugly
AlphaFold has little confidence across much of its prediction for this human ubiquitin-protein ligase. There is no PDB structure to compare it with.

J. M. THORNTON *ET AL. NATURE MED.* **27**, 1666–1669 (2021).

DREAMING UP PROTEINS

Researchers used deep neural networks to invent, or 'hallucinate', sequences of amino acids that could fold into proteins; in some cases they have synthesized these proteins to compare their actual structures with predictions.



'Hallucinated' protein (software prediction)

Actual structure (experimentally determined)

Superposition of hallucinated (blue) and actual (grey) structures

REF. 7

network seems to have learnt something about how complexes fold together. Several days after AlphaFold's code was released, Yoshitaka Moriwaki, a protein bioinformatician at the University of Tokyo, tweeted that it could accurately predict interactions between two protein sequences if they were stitched together with a long linker sequence. Baek soon shared another hack to predict complexes, gleaned from developing RoseTTAFold.

ColabFold later incorporated the ability to predict complexes. And in October 2021, DeepMind released an update called AlphaFold-Multimer[8] that was specifically trained on protein complexes, unlike its predecessor. Jumper's team applied it to thousands of complexes in the PDB, and found that it predicted around 70% of the known protein–protein interactions.

These tools are already helping researchers to spot potential new protein partners. Elofsson's team used AlphaFold to predict the structures of 65,000 human protein pairs that were suspected to interact on the basis of experimental data[9]. And a team led by Baker used AlphaFold and RoseTTAFold to model interactions between nearly every pair of proteins encoded by yeast, identifying more than 100 previously unknown complexes[10]. Such screens are just starting points, says Elofsson. They do a good job of predicting some protein pairings, particularly those that are stable, but struggle to identify more transient interactions. "Because it looks nice doesn't mean it is correct," says Elofsson. "You need some experimental data that show you're right."

The nuclear pore complex work is a good example of how predictions and experimental data can work together, says Kosinski. "It's not like we take all the 30 proteins, throw them into AlphaFold and get the structure out." To put the predicted protein structures together, the team used 3D images of the nuclear pore complex, captured using a form of cryo-EM called cryo-electron tomography. In one instance, experiments that can determine the proximity of proteins turned up a surprising interaction between two components of the complex, which AlphaFold's models then confirmed.

Kosinski sees the team's current map of the nuclear pore complex as a starting point for experiments and simulations that examine how the pore complex functions – and how it malfunctions in disease.

AlphaFold's limits

For all the progress made with AlphaFold, scientists say that it is important to be clear about its limitations – particularly because researchers who don't specialize in predicting protein structures use it.

Attempts to apply AlphaFold to various mutations that disrupt a protein's natural structure, including one linked to early breast cancer, have confirmed that the software is not equipped to predict the consequences of new mutations in proteins, since there are no evolutionarily-related sequences to examine[11].

> **"I think we're going to find new applications of structure that we haven't conceived of yet."**

The AlphaFold team is now thinking about how a neural network could be designed to deal with new mutations. Jumper expects this would require the network to better predict how a protein goes from its unfolded to its folded state. That would probably need software that relies only on what it has learnt about protein physics to predict structures, says Mohammed AlQuraishi, a computational biologist at Columbia University in New York City. "One thing we are interested in is making predictions from single sequences without using evolutionary information," he says. "That's a key problem that does remain open."

AlphaFold is also designed to predict a single structure, although it has been hacked to spit out more than one. But many proteins take on multiple conformations, which can be important to their function. "AlphaFold can't really deal with proteins that can adopt different structures in different conformations," says Schueler-Furman. And the predictions are for structures in isolation, whereas many proteins function alongside ligands such as DNA and RNA, fat molecules and minerals such as iron. "We are still missing ligands, we are missing everything else about proteins," says Elofsson.

Developing these next-generation neural networks will be a huge challenge, says AlQuraishi. AlphaFold relied on decades of research which generated experimental structures of proteins that the network could learn from. That volume of data is currently not available to capture protein dynamics, or the shapes of the trillions of smaller molecules that proteins could interact with. The PDB includes structures of proteins as they interact with other molecules, but this captures just a sliver of chemical diversity, Jumper adds.

Researchers think that it will take time for them to determine how best to wield AlphaFold and related AI tools. AlQuraishi sees parallels with the early days of television, when some programmes consisted of radio broadcasters simply reading the news. "I think we're going to find new applications of structure that we haven't conceived of yet."

Where the AlphaFold revolution is ends up is anybody's guess. "Things are just changing so fast," says Baker. "Even in the next year, we're going to see really major breakthroughs made using these tools." Janet Thornton, a computational biologist at the EMBL-EBI, thinks one of AlphaFold's biggest impacts might be simply to convince biologists to be more open to insights from computational and theoretical approaches. "To me, the revolution is the mindset change," she says.

The AlphaFold revolution has inspired Kosinski to dream big. He imagines that AlphaFold-inspired tools could be used to model not just individual proteins and complexes, but entire organelles or even cells down to the level of individual protein molecules. "This is the dream we will follow for the next decades."

Ewen Callaway writes for *Nature* from London.

1. Kosinski, J. *et al. Science* **5**, 363–365 (2016).
2. Jumper, J. *et al. Nature* **596**, 583–589 (2021).
3. Mosalaganti, S. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2021.10.26.465776 (2021).
4. Baek, M. *et al. Science* **373**, 871–876 (2021).
5. Hartmann, S. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.01.21.477219 (2022).
6. van Kempen, M. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.02.07.479398 (2022).
7. Anishchenko, I. *et al. Nature* **600**, 547–552 (2021).
8. Evans, R. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2021.10.04.463034 (2021).
9. Bryant, P., Pozzati, G. & Elofsson, A. *Nature Commun.* **13**, 1265 (2022).
10. Humphreys, I. R. *et al. Science* **374**, eabm4805 (2021).
11. Buel, G. R. & Walters, K. J. *Nature Struct. Mol. Biol.* **29**, 1–2 (2022).