Method of the Year: protein structure prediction

Nature Methods has named protein structure prediction the Method of the Year 2021.

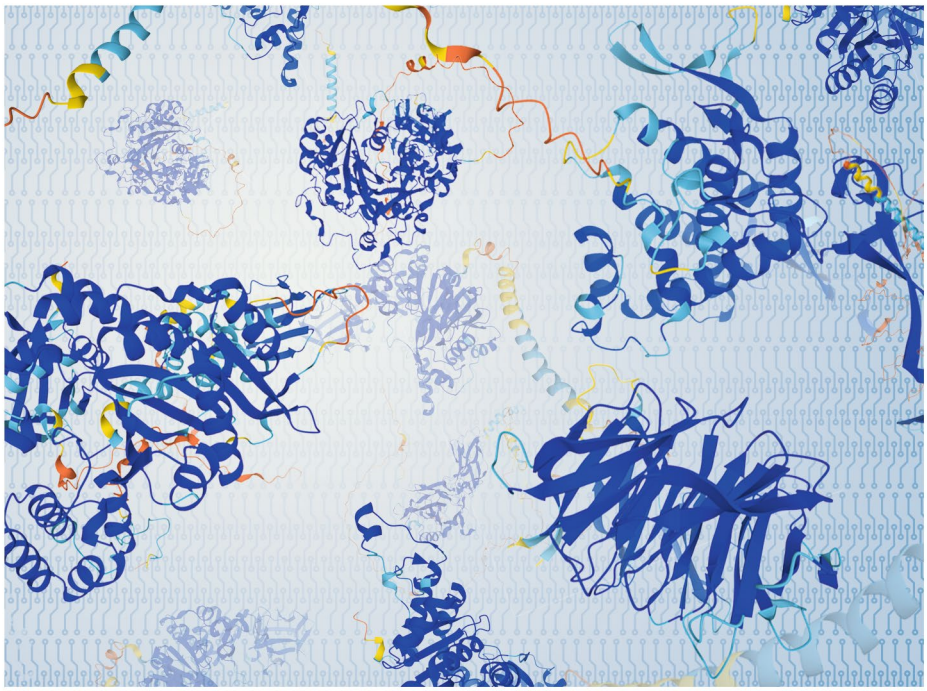
Vivien Marx

If the Earth moves for you, among other reasons, the causes can be geologic or romantic. In science, in the context of predicting protein structure, you might have felt the ground tremble in late 2020 as you perused the results of the 14th Critical Assessment of Protein Structure Prediction (CASP). In this competition, scientists regularly test the prowess of their methods that computationally predict the intricate twirly-curly three-dimensional (3D) structure of a protein from a sequence of amino acids.

A pleasant frisson may have set in more recently as you browsed the new and rapidly growing AlphaFold Protein Structure Database or perused papers[1–3] about a method called AlphaFold and its application to the entire human proteome, or when you dug into the code that drives this inference engine, with its neural network architecture that yields the 3D structure of proteins from a given amino acid sequence. The team behind AlphaFold is DeepMind Technologies, launched as an AI startup in 2010 by Demis Hassabis, Shane Legg and Mustafa Suleyman and now part of Alphabet after being acquired by Google in 2014. DeepMind has presented AlphaFold1[4] and AlphaFold2 and, more recently, AlphaFold-Multimer[5] for predicting the structures of known protein complexes.

AlphaFold has received much attention, but there are many other recent tools from academic labs, such as RoseTTAFold[6], a method with a 'three-track' network architecture developed in the lab of David Baker and colleagues at the University of Washington along with academic teams around the world. It can be used to, for example, predict protein structures and generate models of protein-protein complexes, too. In their paper, the authors note that they had been "intrigued" by the DeepMind results and sought to increase the accuracy of protein structure prediction as they worked on their architecture.

At CASP14 in 2020, AlphaFold2 blew away its competitors. The difference between the DeepMind team results and those of the group in second place "was a bit of a shock," says University College London researcher David Jones. "I'm still processing that a bit, really." Only some months later, when DeepMind gave a glimpse of its method and shared the code, were scientists able to begin looking under the hood. No new information was used to transition AlphaFold1 to AlphaFold2; there was no "clever trick," says Jones. The team used what academics had been doing for years but applied it in a more principled way, he says.

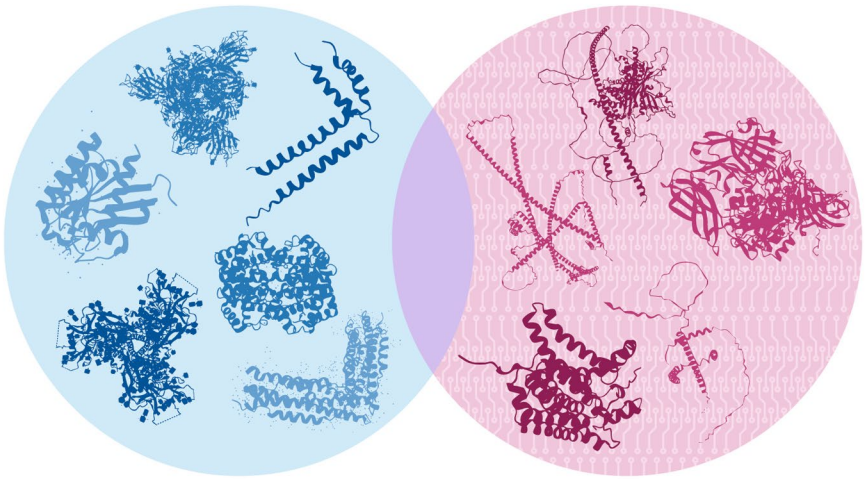


AlphaFold2, developed by DeepMind Technologies, is predicting protein structures on a massive scale. Google acquired the company in 2014. The structures generated by AlphaFold2 are being shared in the AlphaFold Protein Structure Database developed by DeepMind and EMBL-EBI. Credit: T. Phillips, Springer Nature

In the lead-up to the 2018 CASP13 competition, which the DeepMind team won quite handily with AlphaFold1, Jones had consulted for DeepMind. Especially after machine-learning-based methods were introduced in 2016, CASP results had been steadily improving, says Dame Janet Thornton, DBE, from the European Bioinformatics Institute (EBI). Thornton is the former EBI director and has long worked on the challenges of protein structure determination. She was interviewed jointly with Jones. At CASP13, she had been delighted to see progress taking place with protein structure prediction methods. Now, as Thornton considers the possibilities AlphaFold2 opens up, by having solved a big methods puzzle in science, "it gives me a spring in my step." She says she hadn't thought "we'd get quite this far in my lifetime."

Historical build

The way AlphaFold2 can predict a protein structure is the culmination of a scientific journey that began with the work[7,8] of Max Perutz and John Kendrew of the Cavendish Laboratory at the University of Cambridge, says Aled Edwards of the University of Toronto and the Structural Genomics Consortium, a public–private venture. Perutz and Kendrew received the Nobel Prize in 1962 for the way they used X-rays passing through crystallized protein and onto film to painstakingly decipher the structures of proteins such as hemoglobin and myoglobin.



The Protein Data Bank is reserved for structures resolved experimentally. Discussions are underway, but are being kept under wraps for now, on how existing data in the PDB and data generated by AlphaFold2, RoseTTAFold and other computational approaches should be stored and served to the research community. Credit: T. Phillips, Springer Nature

Structural biologists have since followed in their footsteps to experimentally determine structures of many proteins. The research community has deposited structures and accompanying data in the Protein Data Bank (PDB)[9], an open resource founded in 1971 that holds 185,541 structures as *Nature Methods* goes to press.

The PDB's holdings stem from labs around the world that toiled with X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR) or electron microscopy to determine the complex structure of a 3D protein. AlphaFold2's machine-learning algorithm was trained on the PDB's data to assess the patterns with which amino acids become the many combinations of helices, sheets and folds that enable a protein to do its specific tasks in a cell.

Converting experimental signals into structures has been the realm of physicists and mathematicians who devoted time, perseverance and sweat to determine protein structures, says Edwards. In the early days, this work involved assessing measurements on photographic film. The fact that they, and those who followed, have been so committed to data quality enabled the continued work in protein structure determination. Speaking more generally, he says, experimentally solving protein structures is "a pain in the (expletive)." It's why he applauded the foresight of University of Maryland researcher John Moult, who launched CASP in 1994 to highlight and advance community activity related to methods for computationally predicting protein structure. Edwards and many others were part of the NIH-funded Protein Structure Initiative that ran from 2000 to 2015. The project set out to systematically add to PDB's experimentally determined structures and has certainly contributed to AlphaFold's success, says Edwards. When the project's funding ceased, many labs were dismayed. The PSI had been sampling the still-unexplored "structure space," he says. After the PSI ended, the PDB kept growing as labs continued to add their structures.

The PDB's main database has been reserved for structures resolved experimentally and by single methods such as X-ray crystallography, NMR or cryo-electron microscopy (cryo-EM), says Helen Berman, who co-founded the Protein Data Bank. Over time, computational models emerged that used multiple sequence alignments and, later, also machine learning to predict structures. PDB-Dev was set up as a digital home for structures determined with "integrative methods," which means it's for structures generated using experimental methods combined with computational ones. The strictly in silico structures are held in the ModelArchive.

"AlphaFold is a triumph," says Berman. But it "would never ever have succeeded, ever," she says, if models had been improperly mixed with experimentally determined structures. The training set for AlphaFold's neural network has been PDB's well-curated experimental data. DeepMind, in collaboration with EBI, is now filling the AlphaFold Protein Structure Database with hundreds of thousands of computationally generated human protein structures and those from many other organisms, including the 'classic' research organisms maize, yeast, rat, mouse, fruit fly and zebrafish.

Every day, the PDB sees around 2.5 million downloads of protein coordinates, says Berman. Biotech and pharma companies regularly download the database for research performed behind their firewalls. Around the time of CASP13 in 2018, Berman noticed massive downloads that seemed unlike the typical downloads from the structural biology community. Usage is not monitored in detail, and all of it, be it from academia or companies, has made her happy about the resource. "If you don't have people use it, then why have it?" she says. As a child of the 1960s, her personal commitment has been to the "public good" that the resource provides. Over time, the PDB team has navigated expanding its global reach and managing structural data generated by emerging methods. "Now

Confidence measures

Each AlphaFold2 structure is accompanied by a "confidence score," which, as Janet Thornton says, will help and guide users, be they structural biologists or scientists working in other areas. The per-residue confidence score (pLDDT) is between 0 and 100.

Indeed, says Aled Edwards, confidence scores are important pieces of information, but they likely matter more to structural biologists than to other biologists. A diabetes researcher with a hypothesis about a protein who has downloaded a structure with a confidence score of 82% will not be deterred from an experiment he or she is planning, he says. The confidence score could be a point that a reviewer might critically note: the paper authors had chosen to use a "maybe structure," one with a lower confidence score.

Janos Hajdu sees value in confidence scores. Just as one dresses differently for a weather forecast of a 5% or a 95% probability of rain, confidence scores are important and need to be sufficiently well developed. After all, different parts of a predicted structure can have different quality and reliability. The reliability of interpretation also has a human factor to contend with, says Hajdu: even though a lottery win and a lightning strike of a person walking in a storm have similar probabilities, people generally feel less fearful about lightning strikes and more hopeful about their chances of hitting the jackpot.

we have to make a new kind of decision," says Berman, whose workload belies the fact that she recently retired from PDB and her position on the Rutgers University faculty. She has daily calls about how the existing data—the data in the PDB, data generated by AlphaFold2, data generated by RoseTTAFold and other platforms, and other computationally generated data—should be stored and, separately, how they should be served to the community.

Rather than make a centralized behemoth of a database, says Janos Hajdu, who splits his time between the European Extreme Light Infrastructure at the Academy of Sciences of the Czech Republic and the Laboratory of Molecular Biophysics at Uppsala University and who is not involved in these discussions, he would like to see "independent databases that talk to each other."

The actual database plan is still emerging, and details are under wraps until it's worked out, says Berman. It will take six months to a year to hammer out these details and find a solution that works for all.

Into the machine

Compared to other software developed in the academic community, says Thornton, AlphaFold's advances include more accurate placement of side chains in the protein models and an improved approach to integrating machine learning with homology modeling, which looks at protein structure in the context of evolutionarily related proteins. The software uses homology modeling at an "ultrafine" level, says Jones. "It's taking little pieces of everything it needs from the whole of PDB." Instead of taking an homologous 3D structure, building a model from that and then including the side chains and loops, the system finds all the right pieces in high-dimensional space. In a way, he says, it's solving "the worst jigsaw puzzle in history made up of tiny little pieces."

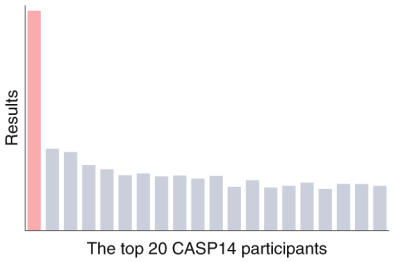
In CASP13, DeepMind entered its AlphaFold1, and then in CASP14 the team entered AlphaFold2. A big difference between CASP13 and CASP14, says Jones, was the way the DeepMind team applied language modeling, specifically the self-attention model, to reduce the need for computing steps run sequentially. The leap made the academic community look like "we'd all been spending 30 years staring at the wall doing nothing," says Jones, which, of course, is not the case. DeepMind's computational approach is based on one that Google Brain scientists presented at the 2017 Conference on Neural Information Processing Systems called 'Attention is all you need'[10]. It's had great impact on AlphaFold, bioinformatics and the computer science community, he says.

Applying this approach in AlphaFold pares back the recurrent layers that the encoder–decoder architectures in machine learning apply and replaces them with "multi-headed self-attention," which interconnects many operations at the same time. These attention models "can just mix data across all the data you feed in," says Jones. Such data-mixing on a scale larger than previously accomplished, lifted constraints that academic groups had faced. Removing computational constraints gives AlphaFold its power to juggle data. "They can mix it up in any way necessary to solve the problem," he says. At the time of CASP14, bioinformaticians were not yet applying this technology, but since then, says Jones, in machine-learning circles he encounters many scientists who work on variations of attention models.

'Attention' is indeed part of a big change in this field, says Burkhard Rost, a computational biologist at the Technical University of Munich who was previously at Columbia University. Both AlphaFold1 and AlphaFold2 rely on multiple sequence alignment and on machine learning. When combining these techniques, academics have used standard feedforward networks with a network of processing units, or nodes, that are arranged in layers with outputs from one layer leading to the next. Training weights the nodes. By including natural-language processing techniques in AlphaFold and in academic labs such as his and others, researchers have enabled machines to 'learn' the grammar of a given protein sequence, says Rost, and the grammar gives context. Based on sentences from Wikipedia, a neural network can extract grammar rules for general language. In the same way, the network can extract the grammar of "a protein language," he says, one that is learned from input amino acid sequence and the corresponding 3D output.

> CASP14 felt like being "hit by a truck or a freight train," says Burkhard Rost. "I'm utterly impressed by what they did."

A platform can learn, for example, that the amino acid alanine might be both at position 42 and 81 in a protein. But it's the 3D environment around these amino acids that affects the protein in different ways. Even though this computational approach does not teach 3D structure or evolutionary constraints, systems can learn rules such as physical constraints that shape protein structure. Rost says that



In the 14th Critical Assessment of Protein Structure Prediction (CASP14), the performance of AlphaFold2 (first column) was far better than that any of the other participants.

never before has there been a CASP winner from outside the field of protein structure prediction. CASP14 felt like being "hit by a truck or a freight train," he says. He found AlphaFold1's predictions to be "amazingly accurate." AlphaFold2 is "a completely different product" in which he sees "so much novelty" he says. "I'm utterly impressed by what they did."

To train the system, the DeepMind approach used tensor processing units (TPUs), which are Google's proprietary processors. They are not for sale; academics can only access them through the Google Cloud. Indeed, DeepMind has "great hardware," says Juan Restrepo-López, a physicist who has turned to biology as a PhD student in the lab of Jürgen Cox at the Max Planck Institute of Biochemistry. AlphaFold2 is likely inconceivable without that hardware, says Restrepo-López. AlphaFold1, with its convolutional neural networks (CNNs), is "for sure much easier to understand due to its simpler architecture." Both AlphaFold1 and AlphaFold2 were trained on TPUs. AlphaFold1 could be run on graphics processing units (GPUs), and this has also eventually become true for AlphaFold2, he says. In AlphaFold2, DeepMind no longer used CNNs but rather transformers, says Restrepo-López. The main advantage for AlphaFold2 came from Google's huge computing clusters, which made it possible to run many types of models. "You can go crazy and run 200,000 experiments because you have unlimited resources," he says. To generate structures, DeepMind first uses multiple sequence analysis, which originated in academia. The core of the algorithm uses transformers, developed at Google. Transformers originated in the field of natural-language processing and are now being applied in many areas. "They are particularly interesting because they can detect long correlations," he says.

This AlphaFold2 architecture with transformers makes it possible, as previously mentioned, to process many aspects of the sequence in parallel and figure out long-term dependencies very well,says Restrepo-López. For example, residues far apart in a sequence can be very close in a folded protein, and this concept has to be introduced into a model.

Scooped

For decades, academic groups around the world have been predicting structures using the millions of amino acid sequences in databases and integrating evolutionary information as part of homology modeling. But DeepMind has used many more sequences plus a different way of scaling computation, says Rost.

When he saw CASP13 results, Konstantin Weissenow, now a PhD student in the Rost lab, was a master's degree student working on a protein structure prediction method. It seemed to him that DeepMind was taking a "traditional" deep learning approach not unlike his. At the time, DeepMind was not sharing the code, but Weissenow felt he could reverse engineer the method and "this is essentially what I tried to do," he says. He incorporated what he gleaned into his method. But CASP14 and AlphaFold2 "was a different story." A few months later, Deep Mind made the AlphaFold2 code public. Michael Heinzinger, another graduate student in the Rost lab, was wrapping up protein language modeling as he watched the livestream of CASP14, which the Rost lab was competing in with Weissenow's tool. When experimentalists began saying that this computational system was reaching close to the quality of experimentally generated results and structures, Heinzinger felt like it was a moment that "people might actually then read in the history books years or decades after this point," he says. "This was just mind blowing."

"The big impact came with CASP14," says Weissenow. By then he had started his PhD work in the Rost lab. He and others had entered their software tool, called EMBedding-based inter-residue distance predictor (EMBER), for CASP14. It's geared toward predicting protein structures for which there are few evolutionary relatives, and computationally it uses a many-layered convolutional network similar to that of AlphaFold1. EMBER allows the team to predict structures on a large scale, and it can predict the human proteome on a typical computer. It was not going to be as good as AlphaFold2, says Rost, but it has a lower carbon footprint. After CASP14, says Weissenow, some participants got together to consider reverse engineering AlphaFold2, but they soon realized that was not going to work. Then, DeepMind published predictions for 98.5% of the human proteome[2]. This was a few weeks before Weissenow had planned to present his tool at a conference and show how it could generate structures of the human proteome. "Scooped again," says Rost, who was interviewed jointly with Heinzinger, Weissenow and postdoctoral fellow Maria Littmann, who works on ways to predict, from amino acid sequence, which residues bind DNA, metal or small molecules.

One issue Littmann faced around 2018 and 2019, says Rost, was the lack of experimental data. It will now be interesting, says Litmann, to see how she and others can integrate the availability of these models into their work and extend it. When predicting residues only from sequence but without a structure, "you don't know what the actual binding site looks like," she says. In the folded structure, residues may be close together or far apart, and it's impossible to know, for example, if two residues are part of the same or a different DNA-binding site. "For that she needs a model," says Rost. Now, given AlphaFold2, Littmann feels she can move beyond the task of predicting which residues bind to being able to predict binding sites.

> "This is a game-changer for several applications we are pursuing in the lab," says Jürgen Cox.

AlphaFold has immense value for work in his lab, says the MPI's Cox. He finds AlphaFold2 is enabling for proteomics more generally. His team integrates structure information into the lab's computational-mass-spectrometry-based proteomics workflows, and Restrepo-López is integrating AlphaFold2 predictions into the Cox lab's MaxQuant algorithms. AlphaFold has trumped a number of existing tools in the protein prediction space, but many of them had been close to retirement age, says Cox. The best way to predict structural information along the protein sequence such as secondary structure or solvent accessibility "is to just do the 3D structure prediction and project these properties from the structure onto the sequence." With the advent of AlphaFold2, says Cox, it's become possible to assume that a structure—either a computationally generated or an experimentally deciphered one—is at researchers' fingertips for nearly every protein and organism and that a computationally generated structure is similar in quality to one determined by X-ray crystallography. "This is a game-changer for several applications we are pursuing in the lab," he says.

Science or engineering?

To some, AlphaFold's achievement is more an engineering feat than a scientific one. AlphaFold2's utility is indisputable, says Jürgen Cox. Every achievement in the development of algorithms and computational tools runs into the issue of being perceived as 'just' engineering as opposed to 'real' science," he says. But it's not justified in this case or in other aspects of computational biology. "Think of the BLAST algorithm. Is it science or engineering?" he asks. Bioinformatics supports life science research and, in so doing, enables findings not achievable through other means. Advances in machine-learning methods are science unto themselves, he says. Differentiating between science and engineering doesn't matter, says Hajdu, given that tool-making is an integral part of science. "A drill, an XFEL, various algorithms, mathematical breakthroughs" can all turn into tools in some fields, he says, referring to X-ray free electron lasers. "Someone's science today is someone's tool tomorrow."

"You can't do the science without the engineering," says Jones. He is essentially 'split' across engineering and science in that he holds a double appointment at University College London: in computer science and in structural and molecular biology. "If the science is wrong, it doesn't matter how good your engineering is," he says. And if there is bad engineering, no correct answers are to be had. "Engineering makes things a reality," he says. "And the science builds the foundations on which that happens."

When a company does it

Some researchers have been irked that a commercial venture achieved this goal of large-scale protein prediction, as opposed to an academic lab or consortium. "I was just pleased overall," says Thornton, who feels the achievement will benefit the entire field. "In a way it was quite disappointing," she says, but it's a company with access to "a lot of compute" and one positioned at the forefront of machine learning.

To Hajdu, it makes no difference that a company and not an academic group reached this goal, he says. Going forward, scientists now have access to many more protein structures, most of them

computationally generated. The situation is comparable to one with the sequencing of the human genome, which both a company and an academic consortium worked on. "The important thing is that it is done," he says. And it matters that the results and tools are or will be available to all. That, he hopes, is an aspect the research community will be able to shape.

When Littmann first saw the CASP14 results, she assumed that because a company had developed the method, it would be kept "behind closed doors," which would prevent the academic community from ever figuring out how the team had achieved what they had. She also assumed one would have to pay to obtain structures, meaning that the academic community would still have needed other methods to predict structure. Her eye-opener moment was when the DeepMind team announced that they are publishing for the research community's benefit the structures from UniProt, which is the database of protein sequences, for the entire human proteome. "That's something that I never expected," she says. Gone was the situation of a lack of high-resolution structures for most proteins. Now, she says, researchers can revisit projects done with sequences and see if they can improve them by adding a structure to their analysis.

Isabell Bludau, a postdoctoral fellow and computational biologist in the lab of Matthias Mann at the Max Planck Institute of Biochemistry, picked up on the excitement in the research community about AlphaFold, but its "real impact" on her and for her work, she says, also occurred when DeepMind published structures for the entire human proteome, dramatically expanding the structures available. "This information can now be easily integrated into any systems biology analysis that I do," says Bludau. As she explores patterns in proteomic data, she can now complement information about the presence and quantity of proteins with structural information, meaning that her analysis can provide a more complete picture. "This is, for me, probably the most exciting part of it," she says.

A landscape of change

AlphaFold is poised to change the structural biology community in a number of ways. The AlphaFold–EBI database gives scientists around the world a "global picture" of the data, says Jones, and this might change the discipline of biology itself.

Early in Jones's career, when he interacted with biologists, he heard them say that protein structure mattered little to their work. Proteins are, he says as he recalls their words, "just blobs that do things and they stick to other blobs." As a PhD student in Thornton's lab, he felt differently about protein structures and began working on computational tools for predicting and analyzing them. Labs these days that use cryo-electron tomography (cryo-ET) and cryo-EM are revealing ever more about the structure of 'blobs', says Thornton. Resolution with cryo-ET is improving and can reach 1.2 Å, she says, although it's still generally "relatively low." For some biological questions, "a blob is enough," she says. But she and Jones both believe the computationally generated models can help many labs to assess proteins, for instance by fitting the computational structure onto the 'blob' they captured with cryo-ET or cryo-EM experiments. What will change overall because of the wealth of computationally generated structures that are becoming available, says Jones, is that the field of structural biology will need to spend less time on technology and thus have more time for assessing why solving structures matters. It will be possible, he says, to appreciate the power of models and the predicted protein structure coordinates for exploring deeper questions.

Max Perutz (left), shown with his wife, and John Kendrew of the University of Cambridge received the Nobel Prize in 1962 for resolving the structure of proteins such as hemoglobin and myoglobin. The way AlphaFold2 can predict a protein structure is the culmination of a scientific journey that began with this work, says Aled Edwards. Credit: Left, Keystone Press/Alamy Stock Photo; right, reprinted from ref. [7] with permission from Perutz, M. F. et al. *Nature* 185, 416–422 (1960), Springer Nature.

> As Janet Thornton considers the possibilities AlphaFold2 opens up by having solved a big methods puzzle in science, "it gives me a spring in my step."

Jones and Thornton have many entries on their to-do list of things they wish to understand: the protein folding pathway, protein–protein and protein–DNA interactions, intrinsically disordered proteins, the interactions of proteins with small molecules, questions of drug design, protein complexes, molecular machines and the overarching question of what proteins do. Having a complete proteome of structures opens entirely new avenues for research questions involving the complexity of protein function. When trying to, for instance, explore and understand protein–protein interactions, it's "quite difficult if you don't have protein structures," says Thornton. "It's not easy when you have them," says Jones, and, says Thornton, "it's impossible when you don't have them." They both laughed as they said this.

Among the problems Cox and his team want to tackle is predicting the effect of post-translational modifications on the structure of proteins and complexes. Speaking more generally, Hajdu says, the next chapter of research in this area "has just turned absolutely wonderful." Not only is there much room to improve the methodology, there are tremendous new opportunities to explore using the new tools. "The scale of possibilities is huge," he says.

AlphaFold2 does not show, says Thornton, how the path of protein folding occurs, how flexibility shapes protein function or what happens to a structure once it's stabilized with a ligand. At the moment, machine learning struggles with such problems, she says. AlphaFold cannot predict how a mutation affects a protein such that it folds differently or becomes less stable, an effect that lies at the core of many diseases and disorders. "It hasn't seen

Local muscle

AlphaFold was trained on the Protein Data Bank, and the DeepMind team used tensor processing units (TPUs), which are Google's proprietary processors, to do so. Academics can access them through the Google Cloud. As of the end of 2021, AlphaFold could not only be run locally, any TPU constraint was removed, says Burkhard Rost. There is AlphaFold Colab, with which users can predict protein structures using, as the team indicates, a "slightly simplified version of AlphaFold v2.1.0." This sets up an AlphaFold2 Jupyter Notebook in Google Colaboratory, which is a proprietary version of Jupyter Notebook hosted by Google that offers access to powerful GPUs. A user can 'execute' the Python code from a browser on a local computer. AlphaFold2 will run on Google hardware, which might be CPUs, GPUs or TPUs depending on a researcher's needs. Separately, researchers have developed a Colab notebook called ColabFold AlphaFold2 for predicting protein structures with AlphaFold2 or RoseTTAFold.

The developers include Martin Steinegger at Seoul National Laboratory, who is one of the co-authors of the AlphaFold2[1] paper, Sergey Ovchinnikov and his team at Harvard University, and colleagues at other institutions. Graduate student Konstantin Schütze is part of the developer team; he's been a member of the Rost lab and has been working in the Steinegger lab as part of his master's degree research. As the Rost lab's Michael Heinzinger explains, ColabFold speeds up AlphaFold2 protein prediction many times over, mainly by accelerating the way multiple sequence alignments are generated with Steinegger's MMseqs2, which is software for iterative protein sequence searching. Users can install ColabFold locally by following the tips on Konstantin Schütze's section of the ColabFold github page. The ability to run AlphaFold2 on GPUs can remove dependency on Google infrastructure, says Heinzinger, because one can choose to install AlphaFold2 on one's own machine. Multiple sequence alignments can be generated on Steinegger's servers, he says, "so you do not even have to compute your MSAs locally."

all the variants," says Jones, so it cannot extrapolate how changes affect a protein's flexibility or stability. In the wake of AlphaFold, some scientists will likely shift their focus. Thornton has observed that "the crystallographers were the most crushed" by AlphaFold and have privately expressed concern that their skills are no longer needed. In the near future, says Cox, he does not see crystallographers as endangered. "Structural information of whole complex structures still requires experiments," he says. But "the combination of cryo-EM with AlphaFold2 predictions will pose a threat to crystallographers soon."

In 1970, Walter Hamilton, a chemist and crystallographer at Brookhaven National Laboratory, published a paper[11] in which he stated that determining a molecular structure by crystallography is routine and that "we have reached the day when such a determination is an essential part of the arsenal of any chemist interested in molecular configuration—and what chemist is not?" Hamilton worked on the molecular and crystal structure of amino acids. "The professional crystallographers really got on his case," says Berman, for saying it had become routine to experimentally determine the structure of small molecules. They were concerned, she says, that he was putting them out of a job, which didn't happen. And, says Thornton, it's not happening now.

AlphaFold is shifting the research landscape, though, says Thornton, given that protein structures will be available for most any amino acid sequence. Over time, X-ray crystallographers have become electron microscopists, she says. "They're looking at bigger complexes, bigger sets or they're doing electron tomography." As such, they are colleagues needed for the next phase in structural biology.

The research community is now in the same place with protein structures as it was with small-molecule structure, says Berman. Back in the day, Berman and her merry band of like-minded junior scientists petitioned Hamilton and others to set up the PDB[7]. "We were very young, we talked a lot, we were so excited about looking at the structures," she says. Hamilton and others did finally agree, but he unfortunately passed away at age 41.

"Ever since I was a postdoc, I've really started to appreciate how enabling cryo-EM was for structural biology," says Bastian Bräuning, who leads a project group in the lab of Brenda Schulman at the Max Planck Institute of Biochemistry. He completed his PhD research in protein crystallography and dabbled, as he says, in cryo-EM. Now he sees how AlphaFold can help with cryo-ET, which produces lower-confidence data than single-particle cryo-EM, but is leading to ever better predictions for parts of bigger protein complexes. Thus, he says, AlphaFold2 "will really enable cryo-electron tomography, too." Says Bräuning, "I've gone from one revolution to the next between my PhD and my postdoc." Once the big shock and surprise to structural biologists settles in and "you really start looking at the opportunities it gives to you, it becomes less worrying," he says. "There's still so much to be done, and not one method or one revolution is going to solve everything." To a large extent, he says, to characterize proteins bound to small ligands one still needs crystallographic data, which these days are generated at large synchrotrons. This approach is high throughput and is used to screen ligands in a way that cryo-EM cannot yet deliver.

AlphaFold2 is likely to affect a small subset of researchers in negative ways, in that this platform has leapfrogged over their works in progress, says Edwards. He mainly interacts with structural biologists and, to them, solving a structure enables their thinking about a biological problem and guides the design of their next experiment. Traditionally, he says, the "big paper" in the academic world has gone to the scientists who solved the structure, not the person who explained the science of that structure. But he hopes a shift can now take place such that more emphasis will be placed on creative scientific insights about the functions of structures. The academic literature contains fewer than 10 papers on half of the proteins that the human genome generates, says Edwards. Understanding more proteins and more about function is going to help to understand disease. Having structures enables the "'what is life?' question," he says, and the questions about what these proteins do. "The vastness of what we don't know is the coolest thing in biology." ❐

Vivien Marx ✉
Nature Methods.
✉*e-mail: v.marx@us.nature.com*

Published online: 11 January 2022
https://doi.org/10.1038/s41592-021-01359-1

References
1. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
2. Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
3. AlQuraishi, M. *Nature* **596**, 487–488 (2021).
4. Senior, A. W. et al. *Nature* **577**, 706–710 (2020).
5. Evans, R. et al. Preprint at https://doi.org/10.1101/2021.10.04.463034 (2021).
6. Baek, M. et al. *Science* https://doi.org/10.1126/science.abj8754 (2021).
7. Perutz, M. F. et al. *Nature* **185**, 416–422 (1960).
8. Kendrew, J. C. et al. *Nature* **185**, 422–427 (1960).
9. Anonymous *Nat. New Biol.* **233**, 223 (1971).
10. Vaswani, A. et al. Preprint at https://arxiv.org/pdf/1706.03762.pdf (2017).
11. Hamilton, W. *Science* **169**, 133–141 (1970).