DEEPMIND'S AI PREDICTS STRUCTURES FOR A VAST TROVE OF PROTEINS

AlphaFold neural network produced 'transformative' database of more than 350,000 structures.

By Ewen Callaway

The human genome holds the instructions for making more than 20,000 proteins. But only about one-third of those have had their 3D structures determined experimentally. And in many of those cases, structures have been determined only in part.

Now, a transformative artificial intelligence (AI) tool called AlphaFold, developed by Google's sister company DeepMind in London, has predicted the structure of nearly the entire human proteome (the full complement of proteins expressed by an organism). Furthermore, the tool has predicted almost complete proteomes for various other organisms, ranging from mice and maize (corn) to the malaria parasite (see 'Folding options').

The more than 350,000 protein structures, which are available through a public database, vary in their accuracy. But researchers say the resource has the potential to revolutionize the life sciences.

"It's totally transformative from my perspective. Having the shapes of all these proteins really gives you insight into their mechanisms," says Christine Orengo, a computational biologist at University College London (UCL).

But researchers emphasize that the data dump is a beginning, not an end. They will want to validate the predictions and, more importantly, apply them to experiments that were hitherto impossible.

Prizewinning predictions

In the process of preparing AlphaFold's code for public release, DeepMind refined it to make the code run more efficiently. Whereas an earlier version could take days to make a prediction for a protein, the updated one can now compute structures in minutes to hours.

With this added efficiency, the DeepMind team set out to predict the structures of nearly every known protein encoded by the human genome, as well as those of 20 model organisms. The structures are available in a database maintained by EMBL-EBI (the European Molecular Biology Laboratory European Bioinformatics Institute) in Hinxton, UK.

In addition to the predicted structures, which cover 98.5% of known human proteins and a similar percentage for the other organisms, AlphaFold generated a measurement of the confidence of its predictions. "We want to give experimentalists and biologists a really clear signal of which parts of the predictions they should rely on," says Kathryn Tunyasuvunakool, a science engineer at DeepMind and first author of a *Nature* paper describing the human proteome predictions (K. Tunyasuvunakool *et al*. *Nature* https://doi.org/gk9kp7; 2021). For the human proteome, 58% of AlphaFold's predictions for the locations of individual amino acids were good enough to be confident in the shape of the protein's folds, Tunyasuvunakool says. A subset of those predictions — 36% of the total — are potentially precise enough to detail atomic features useful for drug design, such as the active site of an enzyme.

Even the less-accurate predictions might offer insights. Biologists think that a large proportion of human proteins and those of other eukaryotes — organisms with cells that have nuclei — contain regions that are inherently disordered and take on a defined structure only in concert with other molecules. "Many proteins are just wiggly in solution, they don't have a fixed structure," says AlphaFold lead researcher John Jumper. Some of the regions that AlphaFold predicted with low confidence match up with those that biologists suspect are disordered, says Pushmeet Kohli, head of AI for science at DeepMind.

Deluge of data

The approximately 365,000 structure predictions deposited last week should swell to 130 million — nearly half of all known proteins — by the year's end, says Sameer Velankar, a structural bioinformatician at EMBL-EBI.

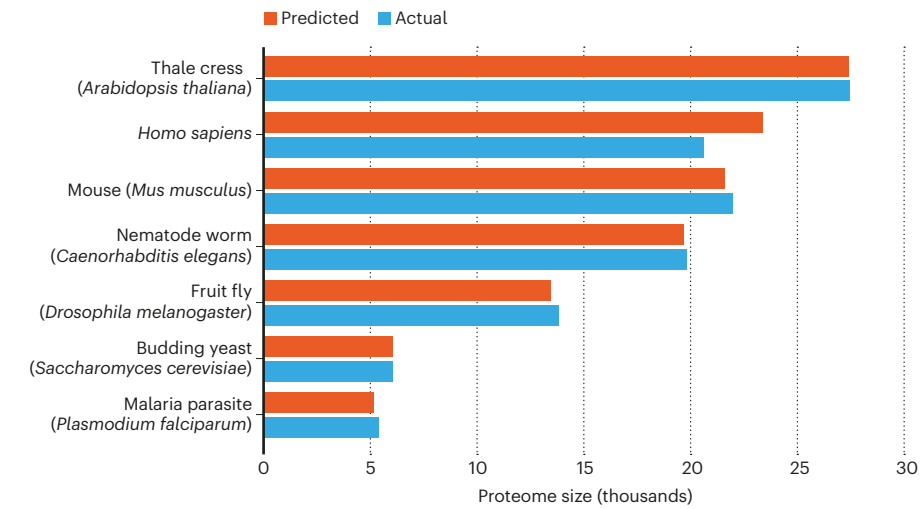
Researchers are already using AlphaFold and related tools to help make sense out of experimental data generated using X-ray crystallography and cryo-electron microscopy. Marcelo Sousa, a biochemist at the University of Colorado Boulder, used AlphaFold to make models from X-ray data of proteins that bacteria use to evade an antibiotic called colistin. The parts of the experimental model that differed from the AlphaFold prediction were typically regions that the software had assigned with low confidence, Sousa notes, a sign that AlphaFold is accurately predicting its limits.

David Jones, a UCL computational biologist who advised DeepMind on an earlier iteration of AlphaFold, is impressed with what the network has achieved. But he says that many of the models predicted by AlphaFold could have been generated with earlier software developed by academics. "For most proteins, those results are probably good enough for quite a lot of the things you want to do."

But the availability of so many protein structures is likely to mark a "paradigm shift" in biology, says Mohammed AlQuraishi, a computational biologist at Columbia University in New York City who works on protein-structure prediction. His field has spent so much time and energy on predicting accurate protein structures on this scale that it hasn't yet worked out what do with such resources. "Everything we do today that relies on a protein sequence, we can now do with protein structure," he says.

FOLDING OPTIONS

AlphaFold has aimed to predict the structure of every protein in humans as well as in 20 model organisms, including those listed here. For some of the proteins, it has provided multiple predictions, which explains why the numbers can be higher than the size of the proteome. In the case of *Homo sapiens*, the predictions include 98.5% of known proteins.



SOURCE: EMBL-EBI AND HTTPS://SWISSMODEL.EXPASY.ORG/REPOSITORY