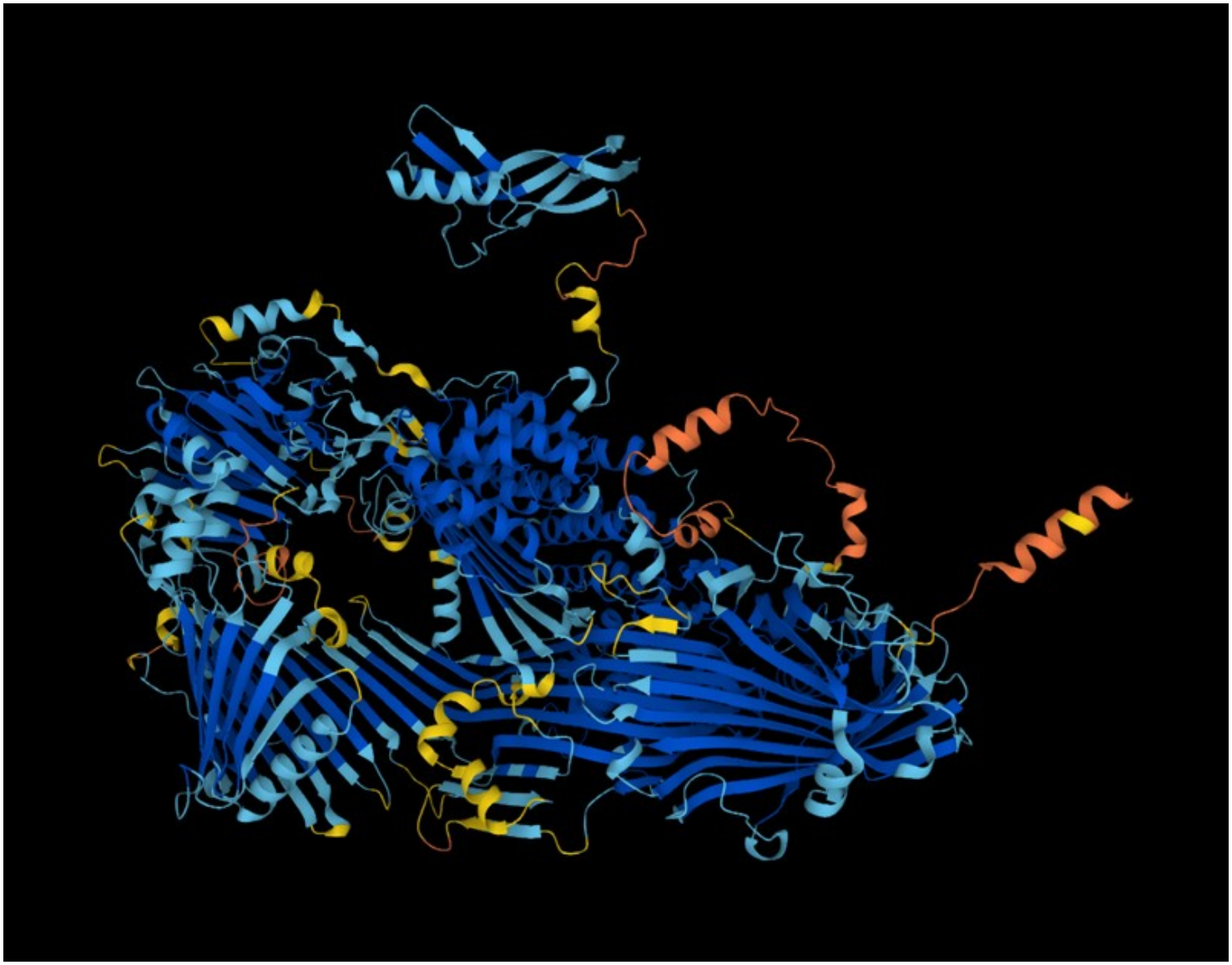
nature

NEWS | 28 July 2022 | Correction 29 July 2022

‘The entire protein universe’: AI predicts shape of nearly every known protein

DeepMind’s AlphaFold tool has determined the structures of around 200 million proteins.

Ewen Callaway



The structure of the vitellogenin protein — a precursor of egg yolk — as predicted by the AlphaFold tool. Credit: DeepMind

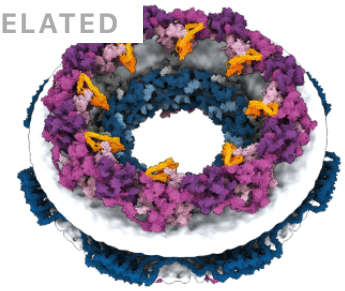
From today, determining the 3D shape of almost any protein known to science will be as simple as typing in a Google search.

Researchers have used AlphaFold — the revolutionary artificial-intelligence (AI) network — to predict the structures of some 200 million proteins from 1 million species, covering nearly every known protein on the planet.

The data dump will be freely available on a database set up by DeepMind, Google's London-based AI company that developed AlphaFold, and the European Molecular

Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), an intergovernmental organization near Cambridge, UK.

RELATED



What's next for AlphaFold and the AI protein-folding revolution

"Essentially you can think of it covering the entire protein universe," DeepMind CEO Demis Hassabis, said at a press briefing. "We're at the beginning of new era of digital biology."

The 3D shape, or structure, of a protein is what determines its function in cells. Most drugs are designed using structural information, and accurate maps are often the first step to discoveries about how proteins work.

DeepMind developed the AlphaFold network using an AI technique called deep learning, and the AlphaFold database was launched one year ago with 350,000 structure predictions covering nearly every protein made by humans, mice and 19 other widely studied organisms. The catalogue has since swelled to around 1 million entries.

"We're bracing ourselves for the release of this huge trove," says Christine Orengo, a computational biologist at University College London, who has used the AlphaFold database to identify new families of proteins. "Having all the data predicted for us is just fantastic."

High-quality structures

The release of AlphaFold last year made a splash in the life-sciences community, which has been scrambling to take advantage of the tool. The network produces highly accurate predictions of the 3D shape, or structure, of proteins. It also provides

information about the accuracy of its predictions, so researchers know which to rely on. Traditionally, scientists have used time consuming and costly experimental methods such as X-ray crystallography and cryo-electron microscopy to solve protein structures.

According to EMBL-EBI, around 35% of the more than 214 million predictions are deemed highly accurate, which means they are as good as experimentally determined structures. Another 45% were deemed confident enough to rely on for many applications.

RELATED

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Many AlphaFold structures are good enough to replace experimental structures for some applications. In other cases, researchers use AlphaFold predictions to validate and make sense of experimental data. Poor predictions are often obvious, and some of them are caused by intrinsic disorder in the protein itself that mean it has no defined shape, at least without other molecules present.

The 200 million predictions released today are based on the sequences in another database, called UNIPROT. It's likely that scientists will have already had an idea about the shape of some of these proteins, because they are covered in databases of experimental structures or resemble other proteins in such repositories, says Eduard Porta Pardo, a computational biologist at Josep Carreras Leukaemia Research Institute (IJC) in Barcelona.

But such entries tend to be skewed toward human, mouse and other mammalian proteins, Porta says, so it's likely that the AlphaFold dump will add significant knowledge because it draws from many more diverse organisms. "It's going to be an

awesome resource. And I'm probably going to download it as soon as it comes out," says Porta.

Because AlphaFold software has been available for a year, researchers have already had the capacity to predict the structure of any protein they wish. But many say that the availability of predictions in a single database will save researchers time, money — and faff. "It's another barrier of entry that you remove," says Porta. "I've used a lot of AlpahFold models. I have not ever run AlphaFold myself."

Jan Kosinski, a structural modeller at EMBL Hamburg in Germany, who has been running the AlphaFold network over the past year, can't wait for the database expansion. His team spent 3 weeks predicting the proteome — the set of all an organism's proteins — of a pathogen. "Now we can just download all the models," he said at the briefing.

One hundred terabytes

Having nearly every known protein in database will also enable new kinds of studies. Orengo's team have used the AlphaFold database to identify new kinds of protein families, and they will now do this on a far grander scale. Her lab will also use the expanded database to understand the evolution of proteins with helpful properties, such as the ability to consume plastic, or worrying ones, like those that can drive cancer. Identifying distant relatives of these proteins in the database can pinpoint the basis for their properties.

Martin Steinegger, a computational biologist at Seoul National University who helped develop a cloud-based version of AlphaFold, is excited to see the database expand. But he says that researchers are likely to still need to run the network themselves. Increasingly, people are using AlphaFold to determine how proteins interact, and such predictions are not in the database. Nor are microbial proteins identified by

sequencing genetic material from soil, ocean water and other ‘metagenomic’ sources.

His hope is that the availability AlphaFold database will have a lasting impact on the life sciences. “It’s going to require quite a big change in thinking.”

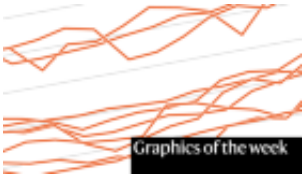
doi: https://doi.org/10.1038/d41586-022-02083-2

UPDATES & CORRECTIONS

Correction 29 July 2022: An earlier version of the standfirst wrongly stated that AlphaFold had determined protein structures from nearly every known species. In fact, it has determined protein structures from nearly every organism with protein sequence data.

Latest on:

Machine learning Computational biology and bioinformatics Proteomics

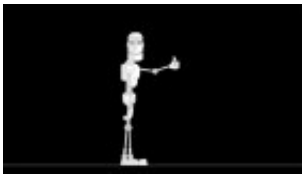


AI's carbon footprint and a DNA nanomotor — the week in infographics

NEWS | 27 JUL 22

Could machine learning fuel a reproducibility crisis in science?

NEWS | 26 JUL 22



Learning over a lifetime

OUTLOOK | 20 JUL 22

Nature (*Nature*) | ISSN 1476-4687 (online) | ISSN 0028-0836 (print)