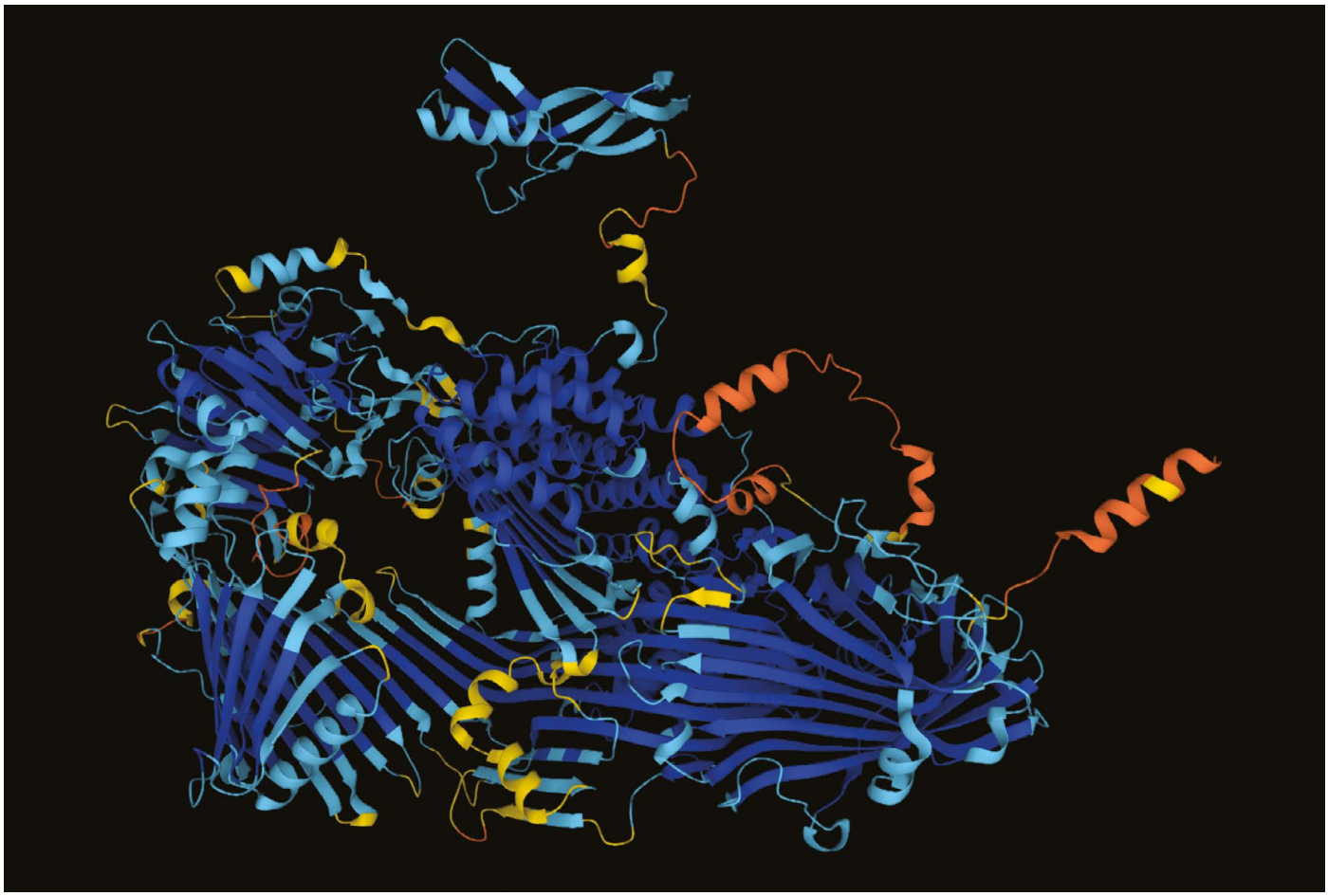
# News in focus



DEEPMIND

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

## 'THE ENTIRE PROTEIN UNIVERSE': AI PREDICTS SHAPE OF NEARLY EVERY KNOWN PROTEIN

DeepMind's AlphaFold tool has determined around 200 million protein structures, which are now available to scientists in a database.

**By Ewen Callaway**

Determining the 3D shape of almost any protein known to science is now as simple as typing in a Google search.

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

On 28 July, the data dump was made available for free in a database set up by DeepMind — the London-based AI company, owned by Google, that developed AlphaFold — and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL–EBI), an intergovernmental organization near Cambridge, UK.

"Essentially you can think of it covering the entire protein universe," DeepMind chief executive Demis Hassabis said at a press briefing. "We're at the beginning of a 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 the creation of accurate maps of proteins' amino-acid arrangement is often the first step to making discoveries about how proteins work.

DeepMind developed the AlphaFold network using an AI technique called deep learning, and the AlphaFold database was launched a year ago with more than 350,000 structure predictions covering nearly every

protein made by humans, mice and 19 other widely studied organisms. Over the months that followed, the catalogue swelled to around 1 million structures.

"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, whose members have since been scrambling to use the tool. The network produces highly accurate predictions of many proteins' structures. It also provides information about the accuracy of its predictions, so researchers know whether they can be relied on. Conventionally, scientists have needed to use 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 to be highly accurate, which means they are as good as experimentally determined structures. Another 45% are considered to be accurate enough for many applications.

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 means it has no defined shape — at least, not without other molecules present.

The 200 million predictions released last week are based on the sequences in another database, called UniProt. It's likely that scientists will have already had an idea about the shapes of some of these proteins, because they are included 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, Spain.

But such entries tend to be skewed towards human, mouse and other mammalian proteins, Porta says. It's likely that the AlphaFold dump will add significant knowledge, because it includes a diverse set of organisms. "It's going to be an awesome resource," says Porta.

Because AlphaFold's 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 AlphaFold 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 once spent three weeks predicting the proteome — the set of all of an organism's proteins — of a pathogen. "Now we can just download all the models," he said at the briefing.

DeepMind chief executive Demis Hassabis.

JUNG YEON-JE/AFP/GETTY

Having almost every known protein in the database will also make new types of study possible. Orengo and her team have used the AlphaFold database to identify new protein families, and they will now do this on a much larger scale. They will also use the expanded repository to help them 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. The identification of 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 to develop a cloud-based version of AlphaFold, is excited about seeing the database expand. But he says that researchers are still likely to need to run the AI network themselves. Increasingly, people are using AlphaFold to determine how proteins interact, and such predictions are not in the database. Other predictions that are not there include microbial proteins identified by sequencing genetic material from soil, ocean water and other 'metagenomic' sources.

Some sophisticated applications of the expanded AlphaFold database might also depend on downloading its entire 23-terabyte contents, which won't be feasible for many teams, Steinegger says. Cloud-based storage could also prove costly. Steinegger has co-developed a software tool called FoldSeek that can quickly find structurally similar proteins and which should also be able to squash the AlphaFold data down.

Even with almost every known protein included, the AlphaFold database will need updating as new organisms are discovered. AlphaFold's predictions can also be improved as new structural information becomes available. Hassabis says DeepMind hopes to update the database annually. His hope is that the repository will have a lasting impact on the life sciences. "It's going to require quite a big change in thinking."

# HOW LONG IS COVID INFECTIOUS? WHAT SCIENTISTS KNOW SO FAR

People with SARS-CoV-2 are told to isolate for a few days. But some can pass on the virus for much longer.

**By David Adam**

When the US Centers for Disease Control and Prevention (CDC) halved its recommended isolation time for people with COVID-19 to five days back in December, it said that the change was motivated by science. Specifically, the CDC said that most SARS-CoV-2 transmission occurs early in the course of the illness, in the one to two days before the onset of symptoms and for two to three days after.

Many scientists disputed that decision then and they continue to do so. Such dissent is bolstered by a series of studies confirming that many people with COVID-19 remain infectious well into the second week after they first experience symptoms. Reductions in the length of the recommended isolation period — now common worldwide — are driven by politics, they say, rather than any reassuring new data.

"The facts of how long people are infectious for have not really changed," says Amy Barczak, an infectious-disease specialist at Massachusetts General Hospital in Boston. "There is not data to support five days or anything