The international journal of science / 29 July 2021

nature

Artificial intelligence in structural biology is here to stay

Machine learning will transform our understanding of protein folding. And it's essential that all data be open.

"I didn't think we would get to this point in my lifetime." That's how one research leader in structural biology responded to last week's publication of research in which artificial intelligence (AI) was used to predict the structure of more than 20,000 human proteins, as well as that of nearly all the known proteins produced by 20 model organisms such as *Escherichia coli*, fruit flies and yeast, but also soya bean and Asian rice. That is a combined total of around 365,000 predictions[1].

The data, publicly accessible for the first time (see https://alphafold.ebi.ac.uk), were released online on 22 July by researchers at DeepMind, a London-based AI company owned by Google's parent company, Alphabet, and the European Bioinformatics Institute, based at the European Molecular Biology Laboratory (EBI-EMBL) near Cambridge, UK.

The DeepMind team developed a machine-learning tool called AlphaFold (see page 635). The team trained this program on DNA sequences, including their evolutionary history, and the already-known shapes of tens of the thousands of proteins contained in a public-access database of proteins hosted by the EBI-EMBL researchers. A week earlier, DeepMind also released the source code for AlphaFold and detailed how it was constructed[2], at the same time that researchers from the University of Washington, Seattle, published details of another protein-structure prediction program – inspired by AlphaFold – called RoseTTAFold (ref. 3).

The unveiling of this catalogue of predicted structures would not be nearly such good news were the data and the methodology not open and freely available. Structural biologists and other researchers are already starting to use AlphaFold to obtain more-accurate models for proteins that have been difficult or impossible to characterize by current experimental methods.

Speeding up structure prediction

Predicting the 3D shape that proteins fold into has been one of biology's unsolved 'grand challenges' since the discovery in 1953 of the structure of DNA itself. Before AI, structure prediction from sequence was an intensely time-consuming, not to say labour-intensive, process with little guarantee of getting an accurate result. The new data will still need to be validated and experimentally verified. But the AI tools can accurately predict protein structures in minutes to hours – compared with the months, or years, that it used to take to determine the structure of just one or two proteins. And that opens up possibilities for applications, for example in the engineering of enzymes to break down environmental pollutants such as microplastics.

> "The AI tools can accurately predict protein structures in minutes to hours."

Last week's breakthrough depended not just on the sharing of open data, but on advances in fundamental science and technology. Since the 1960s, structural biologists have worked on parallel approaches to understanding the science of protein folding. One involves piecing together the structures of proteins by understanding the underlying physical forces. Another attempts to predict the shapes by making comparisons with closely related proteins, using an organism's evolutionary history. And then there's been the all-important role of imaging technologies, starting with X-ray crystallography and now cryo-electron microscopy.

In the basic science of structural biology, key problems remain to be solved. Although AI in science and technology is good at producing accurate results, it doesn't (at least for now) explain how, or why, those results happened. The teams at DeepMind, EBI-EMBL, the University of Washington and elsewhere should be congratulated for crucial breakthroughs. But there is still work to be done to unlock the science – the essential biology, chemistry and physics – of how and why proteins fold.

Public and private

In terms of significance, some are comparing the latest advances to the first draft human genome sequence 20 years ago. And it's true that there are comparisons to be made. Both the Human Genome Project and DeepMind's catalogue of human protein-structure predictions equip their fields with a tool that is set to markedly accelerate discovery.

The human genome's first draft was the result of a race. Solving protein folding has also benefited from a kind of competition – an annual event called the Critical Assessment of Protein Structure Prediction (or CASP), which has been essential to getting a result.

Today's research teams – just like those involved in early genome sequencing – needed open access to data. In making the data and the methodology openly available to all, DeepMind now sets a benchmark that will make it harder for other corporations in this space, such as Facebook and Microsoft, to continue arguing for proprietary data.

And so, what of the future? Over the past week, *Nature* interviewed nearly a dozen researchers in the field. The consensus is that it's too early to predict exactly what impact the application of AI in the life sciences will have, except that any impact will be transformative.

Accurately predicting how AI will change biology needs good training data, which we don't yet have. But in AI, the structural-biology research community – and its collaborators in other fields – have a vast trove of fresh data. In addition to its research and data, AI provides a window into models for research organization and management

nature

that universities should study. For today's researchers, and those in future generations, there is much work to follow up on.

1. Tunyasuvunakool, K. *et al. Nature* https://doi.org/gk9kp7 (2021).
2. Jumper, J. *et al. Nature* https://doi.org/gk7nfp (2021).
3. Baek, M. *et al. Science* https://doi.org/gk7nhq (2021).

The lack of diversity in science images must be fixed

Archives, libraries, photo agencies and publishers need to do better to represent diversity in science.

Last month, *Nature* published a Comment article on how researchers and communities helped each other during a water crisis in Flint, Michigan. While sourcing pictures for the article, *Nature*'s photo editor discovered that there are few images available of the people involved, many of whom are Black.

Recently, we also needed an image of the physicist Elmer Imes, who, in 1918, became only the second African American to be awarded a PhD in physics in the United States. His doctoral work provided early evidence of the quantum behaviour of molecules. But university archives that *Nature* contacted did not have a copy of his photograph. Commercial photography agencies also had nothing. Low-resolution, grainy images do exist, but, shockingly, even the US Library of Congress in Washington DC — which holds images of many important scientists from the nation's history — does not have a photograph. However, such images are available for a number of notable white scientists from Imes's time.

This is far from an isolated case. *Nature* often illustrates articles reporting on communities and countries that are under-represented in science using generic images, in part because universities, national libraries and commercial photo agencies hold relatively few images of people from such communities.

Although we do our best to work with generic images in such situations, they tend to be less compelling than pictures showing real scientists doing real research. When we do use photographs of the researchers themselves, this can boost the impact of the article — attracting greater social media attention, for example — which, in turn, can benefit those individuals and their work.

Systemic racism and science's diversity deficit extend to images, creating a distorted and exclusionary picture of science's past and present. This is an issue that needs attention, and there are several potential ways to rectify it.

> **"Generic images tend to be less compelling than pictures showing real scientists doing real research."**

When it comes to photographs of living scientists from under-represented communities, it can be surprisingly difficult to obtain high-resolution images of a standard that international publishers generally require. The resolution of images on institutional websites is often insufficient. But there's a relatively straightforward fix, at least for some institutions: where resources allow it, universities could ensure that they can provide appropriate access to high-resolution images of their researchers, if individuals have consented to this.

A second and related problem — the lack of high-quality historical images, particularly of people of colour — is also not insurmountable. For example, such images might be available in university records or archives, and, if not, these institutions will often know how to find such images or will have access to ways of improving the quality of the images they do have. National libraries need to work with universities to identify and publish images of notable researchers.

Arguably the most difficult, although no less important, task will be to bring about change in the commercial photography agencies. These agencies are a crucial source of images for media organizations. At *Nature*, we use them all the time, and credit them next to the images. But, more often than not, our searches for photos of particular Black scientists and scientists of other marginalized ethnicities yield negative results, and we are compelled to fall back on generic images of people modelling a generic scene, instead of photos of the scientists themselves. In some cases, photos do exist, but are incorrectly captioned or are not tagged with appropriate keywords, meaning they cannot be found.

Nature approached six large agencies and asked whether they have a dedicated staff member — or an organized process — for improving diversity in their science-related images. Representatives of three agencies responded. None has such a person. One photo repository acknowledged that Black people are not represented in its images of clinical medicine, and that it is actively working to correct this. Tracking diversity needs to be a priority for these agencies.

Community action

Science publishers and media outlets — including *Nature* — also have a responsibility to do more to ensure we are publishing images of the people we feature. And we need to commission more photographers from the communities that we're writing about, something *Nature* has particularly tried to address in our weekly article, Where I Work.

Universities, libraries, publishers and photo agencies — the organizations that hold the keys to so much of the world's photography — must all take steps to diversify our imagery. Science's historical record will remain incomplete while it is missing pictures of people who have contributed to discovery and invention. Such efforts are also essential to make research more welcoming for people from under-represented communities, and to ensure that future generations of researchers reflect those that science has often failed to attract in the past.