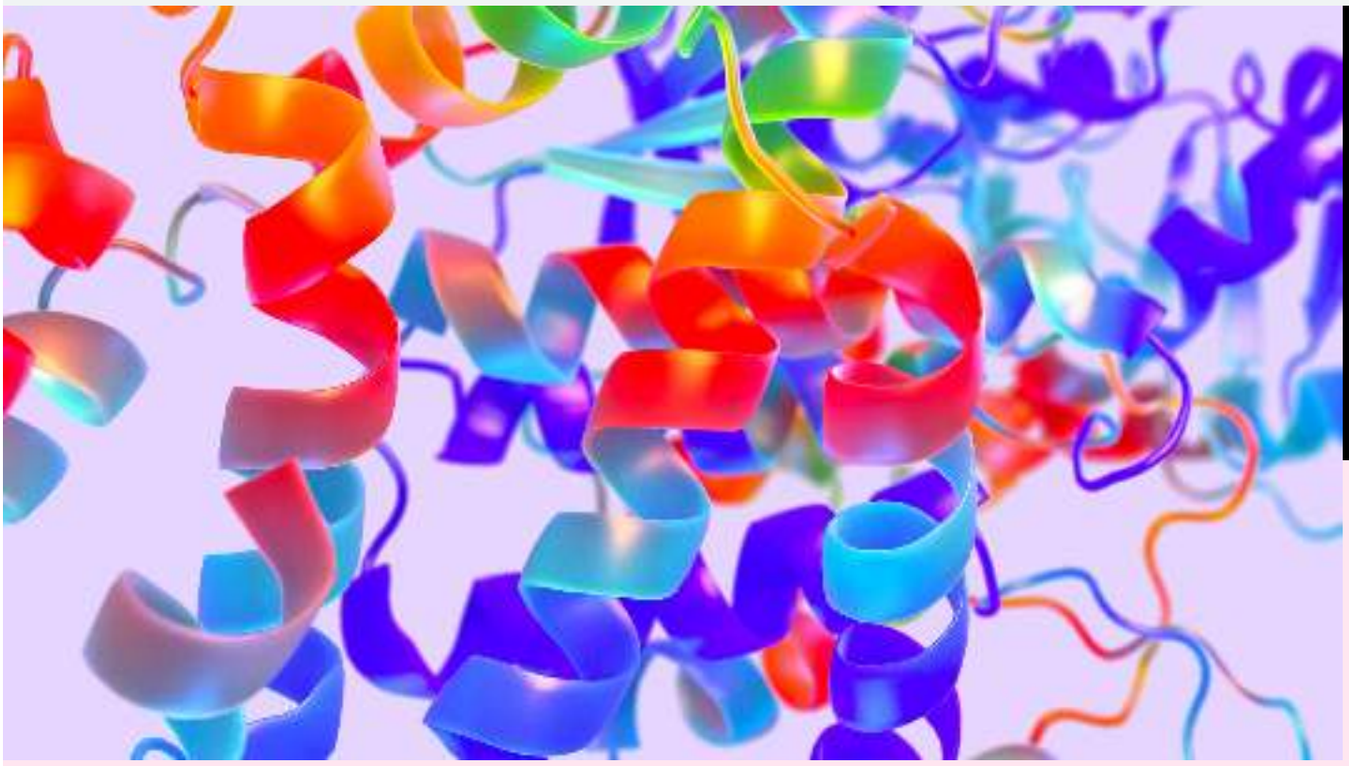
DeepMind

Blog



BLOG POST
RESEARCH

15 JAN 2020

AlphaFold: Using AI for scientific discovery

In our study published today in Nature, we demonstrate how artificial intelligence research can drive and accelerate new scientific discoveries. We've built a dedicated, interdisciplinary team in hopes of using AI to push basic research forward: bringing together experts from the fields of structural biology, physics, and machine learning to apply cutting-edge techniques to predict the 3D structure of a protein based solely on its genetic sequence.

Our system, AlphaFold – described in peer-reviewed papers now published in Nature and PROTEINS – is the culmination of several years of work, and builds on decades of prior research using large genomic datasets to predict protein structure. The 3D models of proteins that AlphaFold generates are far more accurate than any that have come before—marking significant progress on one of the core challenges in biology. The code is available here for anyone interested in learning more or

replicating our results. We're also excited by the fact that this work has already inspired other, independent implementations, including the model described in this paper, and a community-built, open source implementation, described here.

What is the protein folding problem?

Proteins are large, complex molecules essential to all of life. Nearly every function that our body performs—contracting muscles, sensing light, or turning food into energy—relies on proteins, and how they move and change. What any given protein can do depends on its unique 3D structure. For example, antibody proteins utilised by our immune systems are 'Y-shaped', and form unique hooks. By latching on to viruses and bacteria, these antibody proteins are able to detect and tag disease-causing microorganisms for elimination. Collagen proteins are shaped like cords, which transmit tension between cartilage, ligaments, bones, and skin. Other types of proteins include Cas9, which, using CRISPR sequences as a guide, act like scissors to cut and paste sections of DNA; antifreeze proteins, whose 3D structure allows them to bind to ice crystals and prevent organisms from freezing; and ribosomes, which act like a programmed assembly line, helping to build proteins themselves.

The recipes for those proteins—called genes—are encoded in our DNA. An error in the genetic recipe may result in a malformed protein, which could result in disease or death for an organism. Many diseases, therefore, are fundamentally linked to proteins. But just because you know the genetic recipe for a protein doesn't mean you automatically know its shape. Proteins are comprised of chains of amino acids (also referred to as amino acid residues). But DNA only contains information about the *sequence* of amino acids–not how they fold into shape. The bigger the protein, the more difficult it is to model, because there are more interactions between amino acids to take into account. As demonstrated by Levinthal's paradox, it would take longer than the age of the known universe to randomly enumerate all possible configurations of a typical protein before reaching the true 3D structure–yet proteins themselves fold spontaneously, within milliseconds. Predicting how these chains will fold into the intricate 3D structure of a protein is what's known as the "protein folding problem"–a challenge that scientists have worked on for decades. This unsolved problem has already inspired countless developments, from spurring IBM's efforts in supercomputing (BlueGene), to novel citizen science efforts (Folding@Home and FoldIt) to new engineering realms, such as rational protein design.

Why is protein folding important?

Why is protein folding important?

Scientists have long been interested in determining the structures of proteins because a protein's form is thought to dictate its function. Once a protein's shape is understood, its role within the cell can be guessed at, and scientists can develop drugs that work with the protein's unique shape.

Over the past five decades, researchers have been able to determine shapes of proteins in labs using experimental techniques like cryo-electron microscopy, nuclear magnetic resonance and X-ray crystallography, but each method depends on a lot of trial and error, which can take years of work, and cost tens or hundreds of thousands of dollars per protein structure. This is why biologists are turning to AI methods as an alternative to this long and laborious process for difficult proteins. The ability to predict a protein's shape computationally from its genetic code alone – rather than determining it through costly experimentation – could help accelerate research.

Every protein is made up of a sequence of amino acids bonded together

These amino acids interact locally to form shapes like helices and sheets

These shapes fold up on larger scales to form the full three-dimensional protein structure

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA

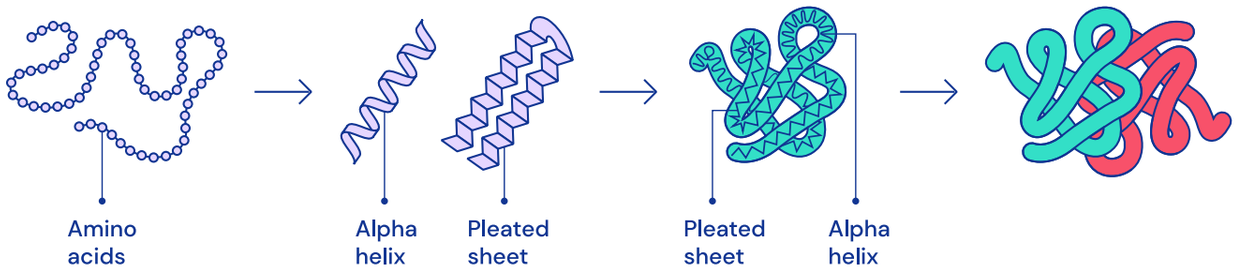


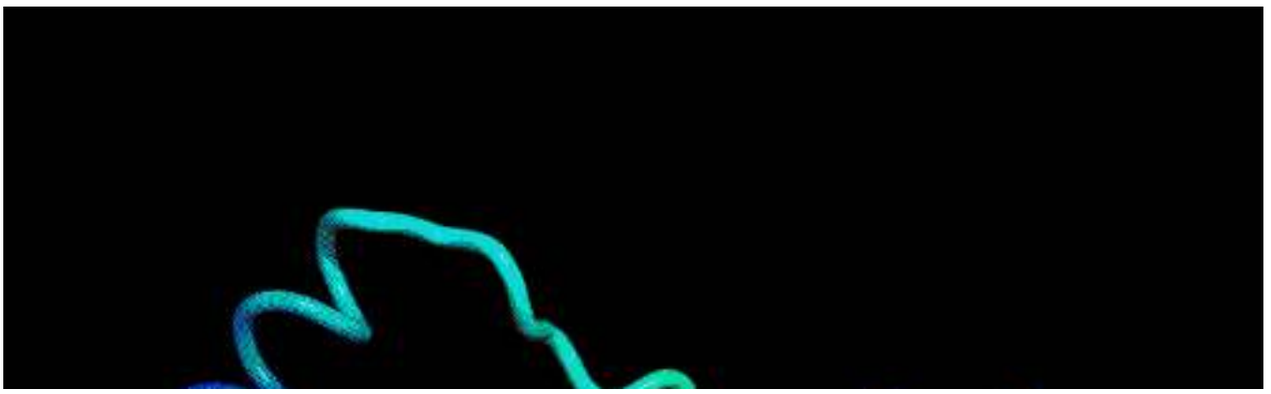
FIGURE 1: COMPLEX 3D SHAPES EMERGE FROM A STRING OF AMINO ACIDS.

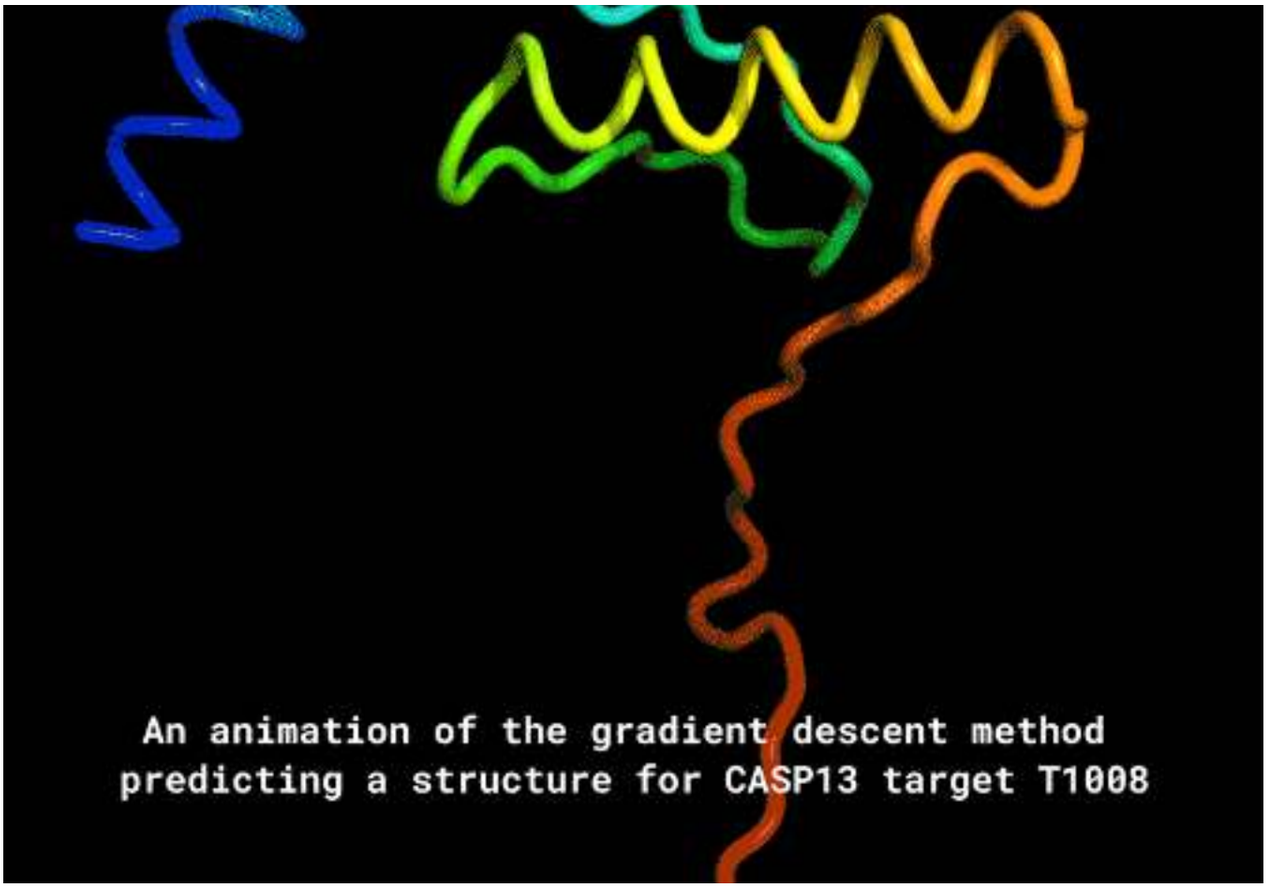
How can AI make a difference?

Fortunately, the field of genomics is quite rich in data thanks to the rapid reduction in the cost of genetic sequencing. As a result, deep learning approaches to the prediction problem that rely on genomic data have become increasingly popular in the last few years. To catalyse research and measure progress on the newest methods for improving the accuracy of predictions, a biennial global competition called the Critical Assessment of Techniques for Protein Structure Prediction (CASP) was established in 1994, and has become the gold standard for assessing predictive techniques. We're indebted to decades of prior work by the CASP organisers, as well as to the thousands of experimentalists whose structures enable this kind of assessment.

DeepMind's work on this problem resulted in AlphaFold, which we submitted to CASP13. We're proud to be part of what the CASP organisers have called "unprecedented progress in the ability of computational methods to predict protein structure," placing first in rankings among the teams that entered (our entry is A7D).

Our team focused specifically on the problem of modelling target shapes from scratch, without using previously solved proteins as templates. We achieved a high degree of accuracy when predicting the physical properties of a protein structure, and then used two distinct methods to construct predictions of full protein structures.

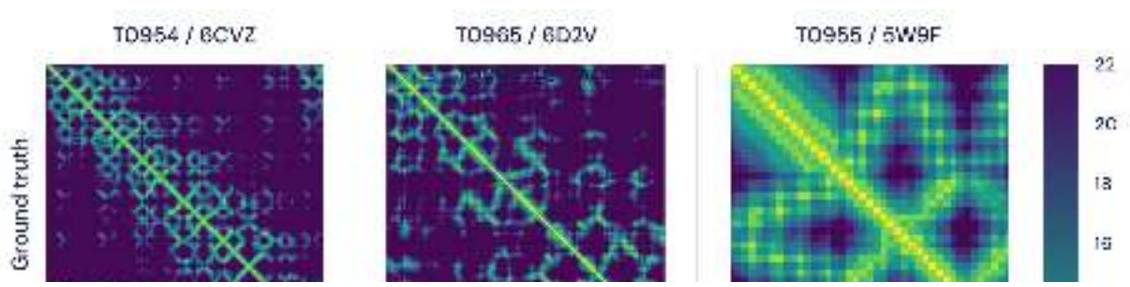




Using neural networks to predict physical properties

Both of these methods relied on deep neural networks that are trained to predict properties of the protein from its genetic sequence. The properties our networks predict are: (a) the distances between pairs of amino acids and (b) the angles between chemical bonds that connect those amino acids. The first development is an advance on commonly used techniques that estimate whether pairs of amino acids are near each other.

We trained a neural network to predict a distribution of distances between every pair of residues in a protein (visualised in Figure 2). These probabilities were then combined into a score that estimates how accurate a proposed protein structure is. We also trained a separate neural network that uses all distances in aggregate to estimate how close the proposed structure is to the right answer.



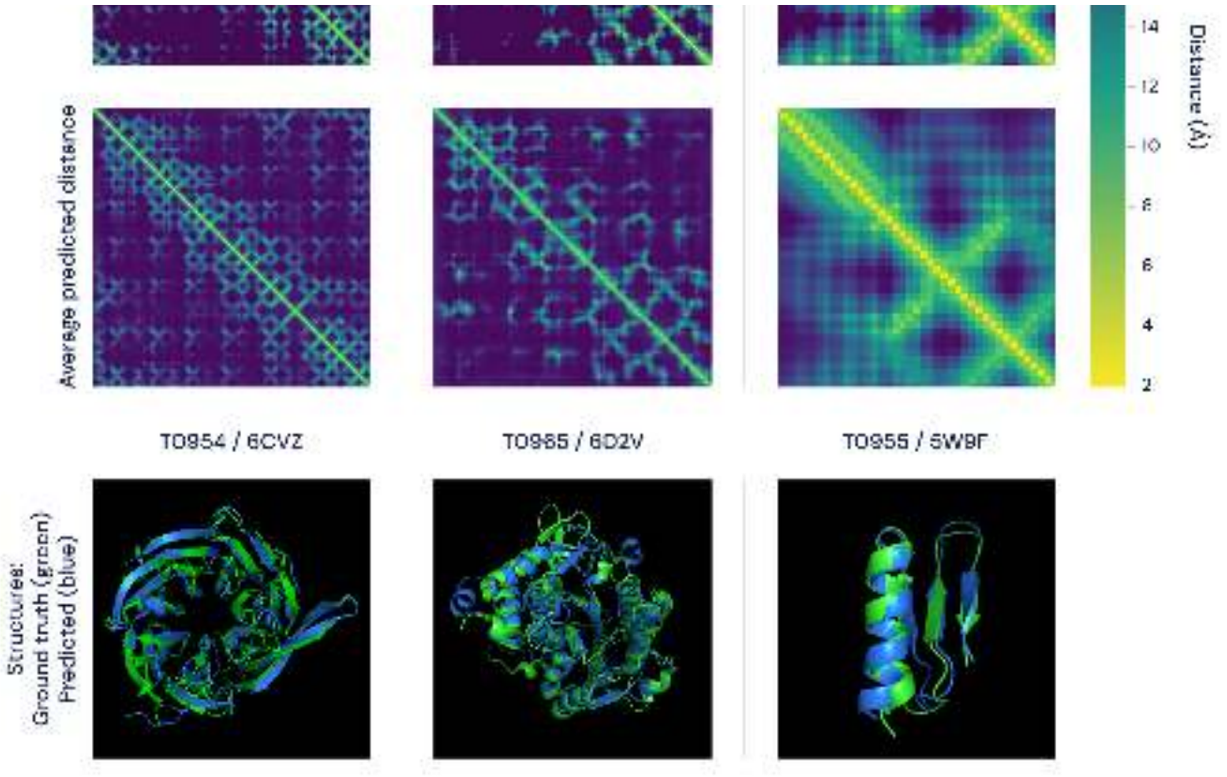
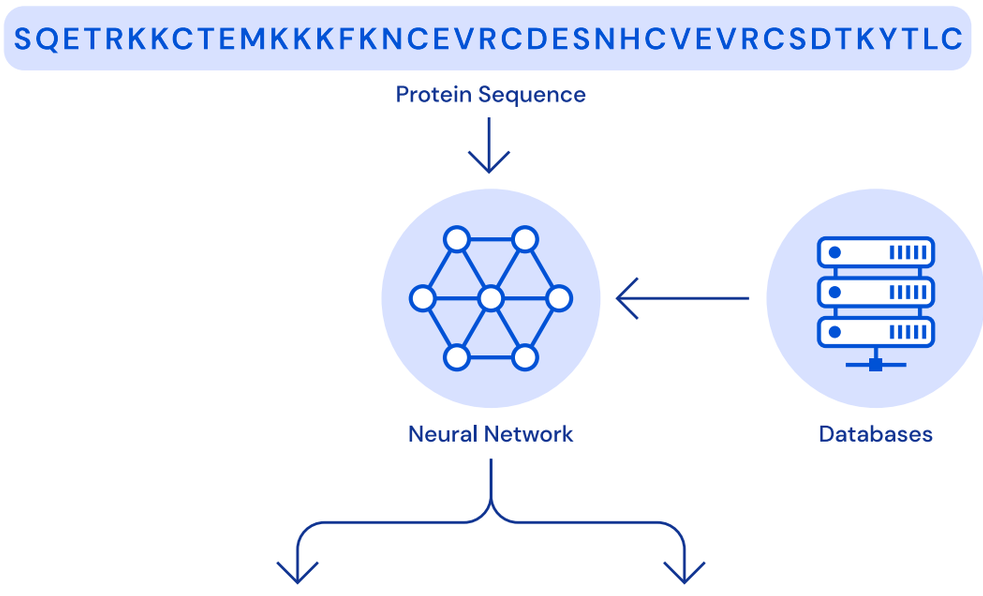


FIGURE 2: TWO WAYS OF VISUALISING THE ACCURACY OF ALPHAFOLD'S PREDICTIONS. THE TOP FIGURE FEATURES THE DISTANCE MATRICES FOR THREE PROTEINS. THE BRIGHTNESS OF EACH PIXEL REPRESENTS THE DISTANCE BETWEEN THE AMINO ACIDS IN THE SEQUENCE COMPRISING THE PROTEIN–THE BRIGHTER THE PIXEL, THE CLOSER THE PAIR. SHOWN IN THE TOP ROW ARE THE REAL, EXPERIMENTALLY DETERMINED DISTANCES AND, IN THE BOTTOM ROW, THE AVERAGE OF ALPHAFOLD'S PREDICTED DISTANCE DISTRIBUTIONS. IMPORTANTLY, THESE MATCH WELL ON BOTH GLOBAL AND LOCAL SCALES. THE BOTTOM PANELS REPRESENT THE SAME COMPARISON USING 3D MODELS, FEATURING ALPHAFOLD'S PREDICTIONS (BLUE) VERSUS GROUND-TRUTH DATA (GREEN) FOR THE SAME THREE PROTEINS.

Using these scoring functions, we were able to search the protein landscape to find structures that matched our predictions. Our first method built on techniques commonly used in structural biology, and repeatedly replaced pieces of a protein structure with new protein fragments. We trained a generative neural network to invent new fragments, which were used to continually improve the score of the proposed protein structure.



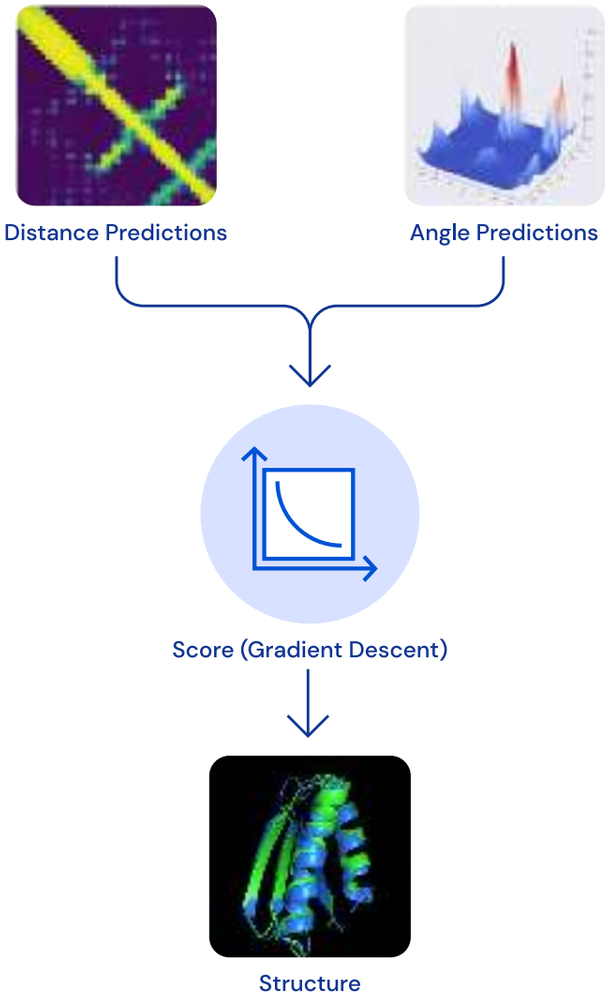


FIGURE 3: A SCHEMATIC OF THE ARCHITECTURE OF THE ALPHAFOLD SYSTEM PREDICTING STRUCTURE FROM PROTEIN SEQUENCE.

The second method optimised scores through gradient descent—a mathematical technique commonly used in machine learning for making small, incremental improvements—which resulted in highly accurate structures. This technique was applied to entire protein chains rather than to pieces that must be folded separately before being assembled into a larger structure, to simplify the prediction process.

The code is available on Github for anyone interested in learning more, or replicating our protein folding results.

What happens next?

While we're thrilled by the success of our protein folding model, there's still much to be done in the realm of protein biology, and we're excited to continue our efforts in this field. We're committed to establishing ways that AI can contribute to basic scientific discovery, with the hope of making real-world impact. This approach might serve to ultimately improve our understanding of the body and how it works, enabling scientists to target and design new, effective cures for diseases more efficiently.

Scientists have only mapped structures for about half of all the proteins made by human cells. Some rare diseases involve mutations in a single gene, resulting in a malformed protein which can have profound effects on the health of an entire organism. A tool like AlphaFold might help rare disease researchers predict the shape of a protein of interest rapidly and economically. As scientists acquire more knowledge about the shapes of proteins and how they operate through simulations and models, this method may eventually help us contribute to efficient drug discovery, while also reducing the costs associated with experimentation. Our hope is that AI will be useful for disease research, and ultimately improve the quality of life for millions of patients around the world.

But potential benefits aren't restricted to health alone–understanding protein folding will assist in protein design, which could unlock a tremendous number of benefits. For example, advances in biodegradable enzymes—which can be enabled by protein design—could help manage pollutants like plastic and oil, helping us break down waste in ways that are more friendly to our environment. In fact, researchers have already begun engineering bacteria to secrete proteins that will make waste biodegradable, and easier to process.

The success of our first foray into protein folding is indicative of how machine learning systems can integrate diverse sources of information to help scientists come up with creative solutions to complex problems at speed. Just as we've seen how AI can help people master complex games through systems like AlphaGo and AlphaZero, we similarly hope that one day, AI breakthroughs will help serve as a platform to advance our understanding of fundamental scientific problems, too.

It's exciting to see these early signs of progress in protein folding, demonstrating the utility of AI for scientific discovery. Even though there's a lot more work to do before we're able to have a quantifiable impact on treating diseases, managing waste, and more, we know the potential is enormous. With a dedicated team focused on delving into how machine learning can advance the world of science, we're looking forward to seeing the many ways our technology can make a difference.

Listen to our podcast featuring the researchers behind this work.

This blog post is based on the following work:

AlphaFold: Improved protein structure prediction using potentials from deep learning (Nature)

Protein structure prediction using multiple deep neural networks in CASP13 (PROTEINS)

Play with the model yourself: see our Github repo

This work was done in collaboration with Andrew Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Sandy Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David Jones, David Silver, Koray Kavukcuoglu and Demis Hassabis

SHARE

AUTHORS

AS Andrew Senior

JJ John Jumper

 Demis Hassabis

PK Pushmeet Kohli

FURTHER READING

Deep Learning

Sciences

Further reading

MODERN SLAVERY STATEMENT

ALPHABET INC