Oct 3, 2021, 07:34pm EDT | 26,852 views

AlphaFold Is The Most Important Achievement In AI—Ever

Rob Toews Contributor ⓘ

AI

I write about the big picture of artificial intelligence.

Follow

Listen to article 19 minutes

DeepMind's AlphaFold represents the first time a significant scientific problem has been solved by ...

[+] IMAGE SOURCE: PROTEINQURE

It can be difficult to distinguish between substance and hype in the field of artificial intelligence. In order to stay grounded, it is important to step back from time to time and ask a simple question: what has AI actually accomplished or enabled that makes a difference in the real world?

This summer, DeepMind delivered the strongest answer yet to that question in the decades-long history of AI research: AlphaFold, a software platform that will revolutionize our understanding of biology.

One of Life’s Great Mysteries

In 1972, in his acceptance speech for the Nobel Prize in Chemistry, Christian Anfinsen made a historic prediction: it should in principle be possible to determine a protein’s three-dimensional shape based solely on the one-dimensional string of molecules that comprise it.

Finding a solution to this puzzle, known as the “protein folding problem,” has stood as a grand challenge in the field of biology for half a century. It has stumped generations of scientists. One commentator in 2007 described it as “one of the most important yet unsolved issues of modern science.”

AI just solved it.

(While use of the word “solved” has generated some disagreement in the community, sometimes devolving into semantics, most experts closest to the topic agree that AlphaFold can indeed be considered a solution to the protein folding problem.)

Why does the protein folding problem matter? And why has it been so hard to solve?

Proteins are at the center of life itself. As prominent biologist Arthur Lesk put it, “In the drama of life at a molecular scale, proteins are where the action is.”

Proteins are involved in basically every important activity that happens inside every living thing, yourself included: digesting food, enabling muscles to contract, moving oxygen throughout the body, attacking foreign viruses and bacteria. Your hormones are made out of proteins; so is your hair.

To put it simply, proteins are so important because they are so versatile. Proteins are able to undertake a vast array of different structures and functions, far more than other types of biomolecules (e.g., lipids or carbohydrates). This incredible versatility is a direct consequence of how proteins are built.

Every protein is comprised of a string of building blocks known as amino acids linked together in a particular order. There are 20 different types of amino acid. In one sense, then, protein structure is elegantly simple: each protein is defined by its one-dimensional sequence of amino acids, with 20 different amino acids to choose from. Proteins can range from a few dozen to several thousand amino acids in length.

But proteins do not stay one-dimensional. In order to become functional, they first fold into complex three-dimensional shapes.

A protein’s shape relates closely to its function. To take one example, antibody proteins fold into shapes that enable them to precisely identify and target particular foreign bodies, like a key fitting into a lock. Understanding the shape that proteins will fold into is thus essential to understanding how organisms function, and ultimately how life itself works.

Here’s the challenge: the number of different configurations that a protein might fold into based on its amino acid sequence is astronomical. Per Levinthal’s paradox, any given protein can theoretically adopt something like 10^300 different configurations. To frame that figure more vividly, it would take longer than the age of the universe for a protein to fold into every configuration available to it, even if it attempted millions of configurations per second. Yet somehow, out of all of these possible configurations, each protein spontaneously folds into one particular shape and carries out its biological purpose accordingly.

Thus, knowing how proteins fold is both ludicrously difficult and absolutely essential to understanding biological processes. Little wonder that the protein folding problem has been something of a holy grail in the field of biology for decades. Enter AlphaFold.

A Triumph of AI

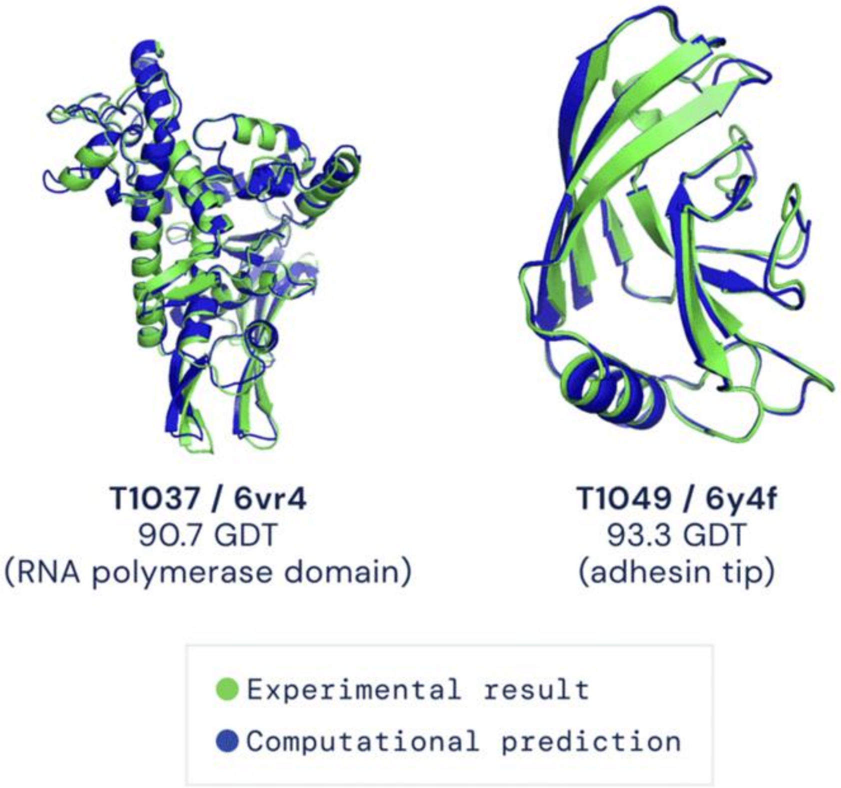
AlphaFold's coming-out party was the Critical Assessment of Protein Structure Prediction (CASP) competition in November 2020. Held every other year, CASP is the most important event in this field—in effect, the Olympics for protein folding.

The competition's format is simple. Contestants are given the one-dimensional amino acid sequences for roughly 100 proteins whose three-dimensional structures have been experimentally determined but are not yet publicly available. Using the amino acid sequences as inputs, the competitors generate predictions for the proteins' structures, which are then compared against the "ground truth" structures to determine their accuracy.

AlphaFold's performance at last year's CASP was historic, far eclipsing any other method to solve the protein folding problem that humans have ever attempted. On average, DeepMind's AI system successfully predicted proteins' three-dimensional shapes to within the width of about one atom. The CASP organizers themselves declared that the protein folding problem had been solved.

Before AlphaFold, we knew the 3-D structures for only about 17% of the roughly 20,000 proteins in the human body. Those protein structures that we did know had been painstakingly worked out in the laboratory over the decades through tedious experimental methods like X-ray crystallography and nuclear magnetic resonance, which require multi-million-dollar equipment and months or even years of trial and error.

Suddenly, thanks to AlphaFold, we now have 3-D structures for virtually all (98.5%) of the human proteome. Of these, 36% are predicted with very high accuracy and another 22% are predicted with high accuracy.



At last year's CASP, AlphaFold accurately predicted the shapes of proteins to within the width of ...
[+] SOURCE: DEEPMIND

CASP co-founder and long-time protein folding expert John Moult put the AlphaFold achievement in historical context: “This is the first time a serious scientific problem has been solved by AI.”

Evolutionary biologist Andrei Lupas was even more effusive: “This will change medicine. It will change research. It will change bioengineering. It will change everything." AlphaFold has already enabled Lupas' lab to determine the structure of a protein that had eluded it for a decade.

How did DeepMind achieve this historic feat?

As with any machine learning effort, the answer starts with training data. AlphaFold was trained on a few different publicly available datasets; helpfully, much important data in this field is open-source. The Protein Data Bank (PDB) is a database containing the three-dimensional structures and associated amino acid sequences for virtually all proteins whose structures have been determined by mankind—around 180,000 in total, spanning human and non-human proteins. Another database, UniProt, contains the amino acid sequences (without structures) for nearly two hundred million more proteins.

The AlphaFold AI model is built with transformers, the same cutting-edge neural network architecture that powers well-known language models like GPT-3 and BERT. Transformers have taken the world of machine learning by storm since being introduced by Google Brain researchers in a seminal 2017 paper. The AlphaFold team created a new type of transformer designed specifically to work with three-dimensional structures, which they call Invariant Point Attention (IPA).

Compared to previous efforts to solve protein folding computationally, one noteworthy characteristic of AlphaFold's design is how massively recursive and iterative it is. The model is architected to maximize information flow at every step; hypotheses pass back and forth prolifically among AlphaFold's many components, enabling the overall system to develop an increasingly accurate prediction of a protein's structure.

A comprehensive overview of AlphaFold's technical details can be found in DeepMind's two recently published *Nature* articles. One more big-picture observation is worth noting here: while DeepMind does have access to far greater computing resources than the typical academic lab, AlphaFold does not merely represent a triumph of brute-force computational power. The amount of compute required to train AlphaFold was in fact modest relative to other high-profile AI models. Building AlphaFold required brilliant software engineering and several significant machine learning innovations. DeepMind is the world's most advanced AI research group, and it showed.

In the words of Columbia University's Mohammed AlQuraishi: "AlphaFold is both a *tour de force* of technical innovation and a beautifully designed learning machine, easily containing the equivalent of six or seven solid ML papers but somehow functioning as a single force of nature."

With all this said, it is important to note that AlphaFold has meaningful limitations.

Its predictions are not always as accurate as more traditional experimental methods. It predicts one stable conformation per protein, but proteins are dynamic and may change shape as they move through the body. Edge cases —like intrinsically disordered proteins and unnatural amino acids—can trip AlphaFold up.

AlphaFold generates predictions about individual protein structures, but it sheds little light on multiprotein complexes, protein-DNA interactions, protein-small molecule interactions, and the like—dynamics that are essential to understand for many biomedical use cases. And because (like any AI system) AlphaFold has learned to make predictions based on its training data, it may struggle to accurately predict the shapes of unusual new proteins, including *de novo* protein designs not found in nature.

Yet there is a broader point to keep in mind here: when the challenges that today remain beyond AlphaFold's grasp get solved, those solutions will themselves almost certainly be powered by deep learning. And the pace of progress will be relentless. For instance, new research out of George Church's lab at Harvard has already improved on the AlphaFold work in important ways. Meanwhile, DeepMind has indicated that it plans to tackle protein complexes next.

The AI genie is out of the bottle; structural biology (and the life sciences more broadly) will never be the same. AlphaFold is just the beginning.

The Power of Open Source

The most important part of the AlphaFold story happened this summer. In July 2021, in a move whose effects will be felt for years to come, DeepMind open-sourced AlphaFold and its associated protein structures.

The London-based AI lab made AlphaFold's source code freely available online and published two peer-reviewed articles in *Nature* detailing its research methodology. Even more important, it launched an online database containing the three-dimensional structures for over 350,000 proteins—again, completely open-source and freely available. This includes structures for nearly every protein in the human body, as well as for 20 other scientifically relevant species like the fruit fly and the mouse.

To put this in perspective, before AlphaFold, humanity had collectively figured out the three-dimensional structures for roughly 180,000 proteins.

And this is just the beginning. DeepMind says that it plans to release structures for *over one hundred million more proteins* in the coming months—that is, nearly every protein whose genetic sequence is known to science.

In the words of EMBL-EBI Director Ewan Birney: "This will be one of the most important datasets since the mapping of the Human Genome."

The implications of DeepMind's decision to open-source AlphaFold are hard to overstate.

As the history of technology makes clear, nothing beats open, permissionless innovation. From the distributed creativity unleashed by the Internet twenty-five years ago to the success and ubiquity of open-source platforms like Kubernetes and Linux, true Cambrian explosions of technology development happen when everyone—not just small closed groups—can freely engage and contribute.

As legendary Sun Microsystems cofounder Bill Joy put it: “No matter who you are, most of the smartest people work for someone else.”

With AlphaFold open-source, an entire ecosystem of biotechnology research and startups will spring up around it in the years ahead. No one can anticipate the many different directions that innovation will flow once millions of high-quality protein structures are freely available at the click of a button. Reflecting the importance and diversity of proteins themselves, the possibilities are limitless.

Early use cases hint at the potential.

Researchers at UCSF have used AlphaFold to uncover previously unknown details about a key SARS-CoV-2 protein, which will advance the development of COVID-19 therapeutics. Using AlphaFold, a team at the University of Colorado Boulder was able to pinpoint a particularly tricky bacterial protein structure, a discovery that will aid their efforts to combat antibiotic resistance, a looming public health crisis. The Boulder team had spent years unsuccessfully trying to determine this protein’s structure; with AlphaFold, they learned it in 15 minutes.

When it comes to commercial opportunity and startup activity, there are two applications in particular for which AlphaFold attracts a lot of attention: drug discovery and protein design.

In a nutshell, designing a new drug entails identifying a compound in the body—most often a protein—that you want to target, and then finding a molecule (the drug) that will successfully bind to that target, producing some beneficial health outcome. Knowing the three-dimensional shape of a prospective protein target is essential to this process because a protein's shape defines which and how other molecules will bind to it. AlphaFold makes available a vast new set of drug target candidates to explore.

One area of drug discovery for which AlphaFold holds particular promise is neglected diseases. Neglected diseases are those for which little research funding is directed to develop treatments, often because the disease affects very few people or because the populations that it affects are low-income and thus represent a less compelling market opportunity.

AlphaFold helps level the playing field in the search for therapeutics for these disease states by, for the first time, making relevant protein structures instantly available without the need for costly laboratory work. DeepMind has already announced a partnership with the nonprofit Drugs for Neglected Diseases initiative (DNDi) to tackle deadly neglected diseases like Chagas disease and Leishmaniasis.

But it is important to keep expectations here tempered. AlphaFold will not transform drug discovery overnight, for a few reasons.

AlphaFold's structures are not always accurate and granular enough for drug discovery purposes, particularly when it comes to a protein's active binding sites. In addition, the fact that AlphaFold predicts structures for individual proteins in isolation is a major limitation: what is most essential to understand for purposes of drug development is the structure of protein-drug interactions. Building a computational system to predict protein-drug structures is an even more daunting puzzle than individual protein folding (due above all to training data requirements); it remains out of reach today.

And ultimately, target discovery is just the first step in the very long and expensive process of creating a new drug. While AlphaFold may help accelerate this initial phase, it can do little about the many less tractable downstream bottlenecks (namely, human clinical trials) in the years-long journey to bring a new drug to market.

Protein researchers and entrepreneurs seem even more excited about AlphaFold's potential to boost the burgeoning field of *de novo* protein design.

The basic insight motivating protein design is that there is a vast universe of proteins that could theoretically be constructed, of which only an infinitesimal fraction have actually ended up in the world as a result of natural evolution. By exploring this uncharted universe of possible structures not found in nature, researchers seek to bring novel proteins into the world that are tailor-made for particular applications, from fighting disease to slowing climate change.

AlphaFold may prove to be a powerful tool in these efforts. For instance, verifying that a particular *de novo* protein candidate will actually fold up in a structurally viable way is a major gating function; before AlphaFold, this was costly and time-consuming and therefore could only be done for a small handful of candidates. With AlphaFold, it is trivial to map an amino acid sequence to a hypothesized three-dimensional structure, enabling much more rapid experimentation.

DeepMind and the University of Portsmouth in the U.K. recently announced a partnership to use AlphaFold to help design new types of proteins that more efficiently break down plastic waste, in order to combat pollution.

"Structure-informed enzyme engineering is a core aspect of our work but obtaining those structures has, until now, always been the bottleneck in our research pipeline," said University of Portsmouth professor Andy Pickford. "Being given access to AlphaFold has transformed our research strategy."

Conclusion

AlphaFold is a scientific achievement of the first order. It represents the first time that AI has significantly advanced the frontiers of humanity's scientific knowledge. Credible industry observers have speculated that it might one day win the researchers at DeepMind a Nobel Prize.

"This is a history book moment," said protein folding researcher Carlos Outeiral.

At the same time, AlphaFold is no silver bullet for real-world challenges like drug discovery. Figuring out the most viable and impactful ways to translate AlphaFold's fundamental insights into products that create value in the real world will entail years of hard work from researchers and entrepreneurs. But make no mistake: the long-term impact will be transformative.

The European Molecular Biology Laboratory (EMBL), the non-profit research organization in charge of stewarding AlphaFold, summed it up well: "AlphaFold will provide new insights and understanding of fundamental processes related to health and disease, with applications in biotechnology, medicine, agriculture, food science and bioengineering. It will probably take one or two decades until the full impact of this development can be properly assessed."

Follow me on Twitter.

Rob Toews

Follow

Rob Toews is a venture capitalist at Radical Ventures.