Deepmind AI Predicts 214 Million Protein Structures

Find out more about the machine learning algorithms that have helped the AlphaFold project predict more than 214 million protein structures.

AUSTIN, TEXAS, UNITED STATES, December 13, 2022 / EINPresswire.com/ -- In this Formaspace laboratory report, we take a look behind the headlines at the machine learning algorithms that have helped the AlphaFold project predict more than 214 million protein structures.

Why Understanding The 3D Structure Of Proteins Is Difficult Yet Vital For Biological Research

No matter what your lab requirements, Formaspace can build a custom lab solution that fits your needs.

Understanding the 3D structures of a protein can be enormously useful in uncovering the function and evolution of specific metabolic processes, such as identifying the cause (and hopefully cure) of a disease or evaluating whether an animal protein can be used as a human disease research model.

> " In July of this year, Deep Mind announced they had identified 214 million 3D protein structure predictions, making them freely available for public use."
>
> *Formaspace*

Yet visualizing the complex 3D structure of proteins accurately has remained a challenge, requiring time-consuming, painstaking direct observations using X-ray crystallography or (increasingly) cryo-EM techniques to create density maps of the protein structure in 3D space.

Researchers have long been looking for a shortcut, asking whether it would be possible to use our knowledge of the sequence of amino acids found in each protein to predict what its 3D structure would look like.

Why Are Protein Structures So Complicated?

To borrow the words of Winston Churchill (speaking about Russia in 1939):

"It's a riddle, wrapped in a mystery, inside an enigma."

Calculating 3D protein structures is hard. To illustrate this, here is a breakdown of the first three ways they can fold in 3D space:

1) Primary Protein Structure

2) Secondary Protein Structure

3) Tertiary Protein Structure

The primary structure of a protein is a long backbone of amino acids held together with peptide bonds.

α-helix   β-Sheet (3 strands)

Long chains of amino acids are folded or twisted into smaller secondary structures; the two most common are the alpha helix (twisted like a coil spring) and the beta sheet (folded back and forth like a string of paper dolls).

CASP, The Bi-Annual Scientific Challenge For Predicting 3D Protein Structures

As you can imagine, it's challenging to derive the 3D version of a protein if all you have to work with is the original sequence of amino acids in the chain.

But where there is a daunting scientific challenge, there is often a contest designed to spur innovation.

In 1994, Professor John Moult, Institute for Bioscience and Biotechnology Research at the University of Maryland, co-founded CASP, the Critical Assessment of Techniques for Protein Structure Prediction.

Every two years, CASP holds an international protein folding prediction contest that has drawn participation from over 100 top research groups from around the world.

What's Behind AlphaFold's Prowess In Predicting 3D Protein Structures?

The AlphaFold project is one of several emerging AI-based tools that can predict the complex 3D structure of proteins.

AlphaFold is a project of DeepMind; the London-based company co-founded in 2010 by Demis Hassabis, which was subsequently acquired by Google's parent company, Alphabet, in 2014.
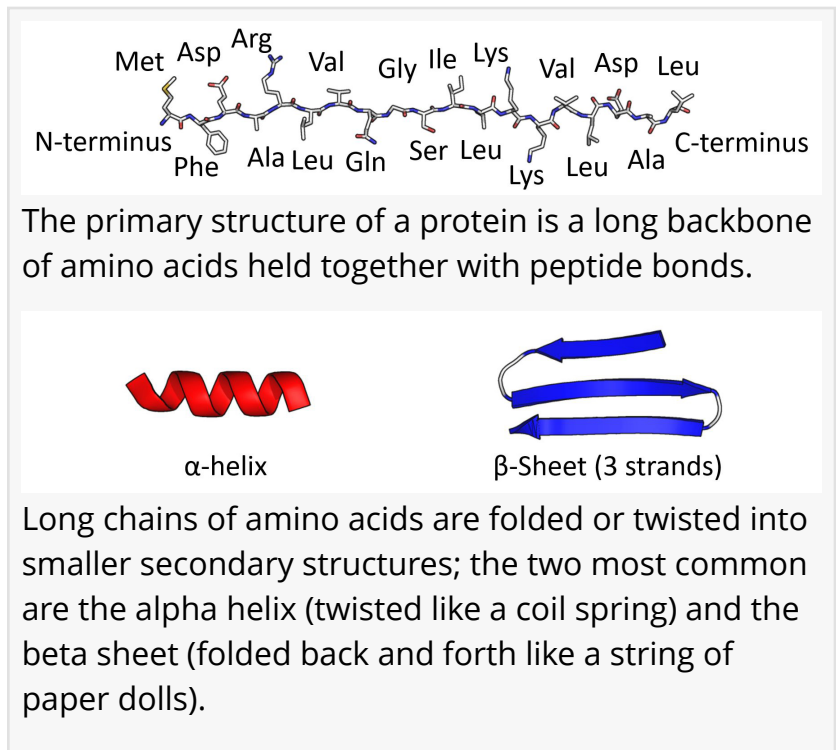
The DeepMind team first achieved widespread fame when its artificial intelligence-based AlphaGo program beat world champions playing the world's oldest board game, Go.

In 2016, the organization turned its expertise to the problem of predicting protein folding.

They entered the CASP competition for the first time in 2018 (at CASP 12), where the AlphaFold team took top honors in predicting moderately difficult protein targets.

The AlphaFold team returned in 2020, where they once again swept the board at CASP 12, achieving a near 90% accuracy score in predicting 3D protein structures.

In July of this year, Deep Mind announced they had identified 214 million 3D protein structure predictions, making them freely available for public use.

It's already having a dramatic effect on medical research.

For example, at the University of Oxford, AlphaFold protein predictions have sped up laboratory research on the parasites that cause malaria, helping lab researchers quickly identify target sites where antibodies could attach to proteins and block disease transmission.

How Does AlphaFold Make 3D Protein Structure Predictions?

The secret behind AlphaFold's dramatic achievements is machine learning, which is proving to be a powerful tool for solving problems that have a defined set of boundary conditions that could result in several possible outcomes.

Examples of where machine learning has met with success include recognizing text on images or video, rapid two-way translation of spoken or written human languages, interpreting medical scan imagery for signs of disease, creating sophisticated artwork based on keywords, and even powering autonomous self-driving cars (this last point, of course, remains a work-in-progress).

Training The Machine Learning Data In AlphaFold, A Brief Overview

Let's look at a schematic diagram from Kathryn Tunyasuvunakool, Research Scientist at DeepMind. (For easier identification, we have added labels A through H.)

· Input Sequence A
This is the known amino acid sequence of the protein that we are trying to predict its corresponding 3D structure.

· Training Sets B, C, And D
Many machine-learning programs rely on training sets to "teach" the algorithm about the real-

world problems it seeks to solve.

In the case of AlphaFold (and machine-learning operations from other competing research teams), the training set information is sourced from the international Protein Data Bank (PDB), which catalogs protein structures verified by X-ray crystallography, NMR spectroscopy, Cryo-EM, or other advanced methods.

This PDB information is reformulated into three training sets to be used by the machine-learning transformer networks:

B) Multiple Sequence Alignments (MSA), which are short amino acid chain sequences shared (e.g. "aligned") among related proteins, such as chymotrypsin (an enzyme that breaks down other proteins during digestion) which is found in humans and other species. Identifying related proteins with significant alignments can often help "guide" the machine learning algorithm toward a more accurate solution.

Read more...

Julia Solodovnikova
Formaspace
+1 800-251-1505
email us here
Visit us on social media:
Facebook
Twitter
LinkedIn

This press release can be viewed online at: https://www.einpresswire.com/article/606149517