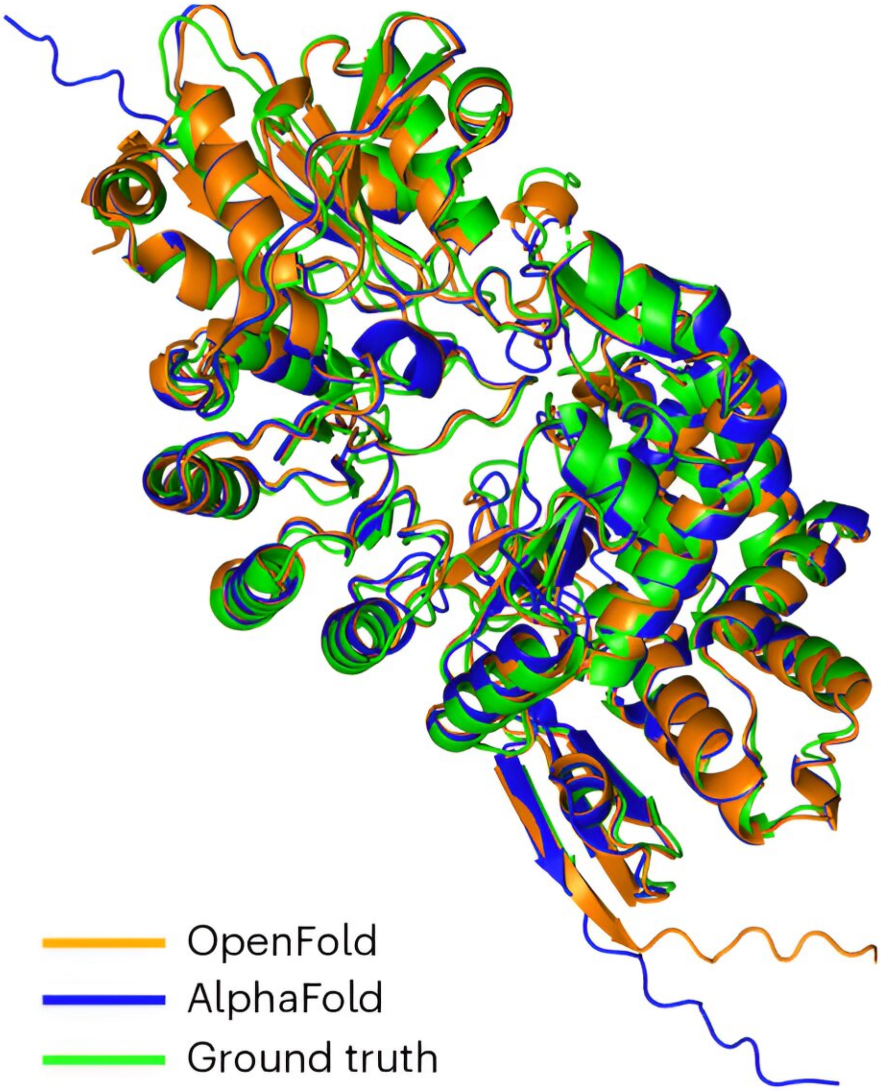
PHYS.ORG

AI, computation, and the folds of life: Supercomputers help train a software tool for the protein modeling community

August 13 2024, by Jorge Salazar



A new, open source software tool called OpenFold has been developed by scientists that uses artificial intelligence and harnesses the power of supercomputers to predict protein structures. Image illustrates OpenFold matching the accuracy of AlphaFold2, using predictions by OpenFold and AlphaFold2 overlaid with an experimental structure of the Streptomyces tokunonesis TokK protein. Credit: *Nature Methods* (2024). DOI: 10.1038/s41592-024-02272-z

Form follows function, and this is especially true for life's building blocks—proteins. The folds and shape of molecular proteins reveal their function in supporting life.

A new, open source software tool called OpenFold has been developed by scientists that uses artificial intelligence (AI) and harnesses the power of supercomputers to predict protein structures.

The research could help develop new medicines and better understand misshapen proteins such as those linked to neurodegenerative diseases like Parkinson's and Alzheimer's disease.

OpenFold builds on the success of AlphaFold2, developed by Google DeepMind and used since 2021 by over two million researchers for protein predictions in vaccine development, cancer treatments, and more.

"AlphaFold2 was a breakthrough for science," said Nazim Bouatta, a senior research fellow at Harvard Medical School who works at the interface of AI and biology. "We built a fully open source version—OpenFold—that is now helping academia and industry to move the field forward."

Bouatta co-authored a study in the journal *Nature Methods* announcing OpenFold, a fast, memory efficient, and trainable implementation of AlphaFold2.

He started the project with his colleague Mohammed AlQuraishi, formerly at Harvard but now at Columbia University. The project grew into the OpenFold Consortium, a syndicate of startup companies working in collaboration with academia.

"Extremely talented students from Harvard and Columbia also contributed to the work, with Gustaf Ahdritz doing a remarkable job. They all did an amazing job implementing the code," Bouatta said.

A core facet of AI is the large language models (LLMs), which take vast quantities of text and generate new and meaningful text from it, such as the human-like ability of ChatGPT to answer queries based on substantial amounts of text data.

"We need about 100 graphic processing units (GPUs) to train a system like OpenFold. To put things into perspective, to train the latest ChatGPT, you need thousands and thousands of GPUs," Bouatta said.

One of the very first applications of OpenFold came from Meta AI, previously Facebook. Meta AI recently released an atlas of more than 600 million proteins from bacteria, viruses, and other microorganisms that had not yet been characterized.

"They used OpenFold to integrate a 'protein language model,' very similar to ChatGPT, but where the language is the amino acids that make up proteins," Bouatta said.

"In a way, the information in living organisms is organized in a language," Bouatta explained, referring to the example of the letters A-C-

G-T that represent the four bases of DNA—adenine, cytosine, guanine, and thymine. "This is the language that nature picked to build these sophisticated living organisms."

Going even further, there is a second layer of language for proteins, the letters that represent the 20 amino acids that make up all proteins in the human body and characterize what the protein can do.

Genome sequencing has generated large data on the letters of life, but missing until now is a 'dictionary' that can take those letters and yield the shape of a protein in three dimensions and model the sites to bind small molecules to it.

"Machine learning allows us to take a string of letters, the amino acids that describe any kind of protein that you can think of, run a sophisticated algorithm, and return an exquisite three-dimensional structure that is close to what we get using experiments. The OpenFold algorithm is very sophisticated and uses new developments that we're familiar with from ChatGPT and others," Bouatta said, referring to the concepts developed by Google transformers and elements of the main ChatGPT algorithm.

A key advantage of OpenFold lies in its ability to train the model with a scientist's own data, something that is not possible with the publicly available version of AlphaFold2. "Having the ability to train a system with OpenFold is opening major avenues for research both in academia and industry," Bouatta said.

In the coming months, Bouatta expects to release a modality of OpenFold with the ability to characterize a protein-ligand complex, the complicated orientation of small molecules that bind to a protein.

"That's how drugs achieve their mechanism of action. Understanding this is particularly important," he explained.

TACC awarded the OpenFold team allocations on the Frontera and Lonestar6 supercomputers, in particular the GPU nodes that have been instrumental in powering AI applications worldwide.

"TACC has been an extremely good collaborator," Bouatta said. "I would like to thank TACC for allowing us to access these resources, which allowed us to deploy machine learning and AI at the scales we needed."

"Supercomputers in combination with AI are radically changing how we approach biology. The power of a supercomputer is that it allows us to predict 100 million structures in just a few months. Once the system is trained, we can get structures in seconds. They will not replace experiments, however, because we need to go back to the lab to test our ideas."

The integration of AI systems like OpenFold with more traditional physics-based systems is helping scientists understand life at the most fundamental level and opening avenues for treating neurodegenerative disease.

"Supercomputers are the microscope of the modern era for biology and drug discovery," Bouatta concluded. "If we keep putting more resources into using the AI/computational approach with supercomputers, we can bootstrap our abilities to understand life and cure diseases."

More information: Gustaf Ahdritz et al, OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization, *Nature Methods* (2024). DOI: 10.1038/s41592-024-02272-z

Provided by University of Texas at Austin

Citation: AI, computation, and the folds of life: Supercomputers help train a software tool for the protein modeling community (2024, August 13) retrieved 13 August 2024 from https://phys.org/news/2024-08-ai-life-supercomputers-software-tool.html