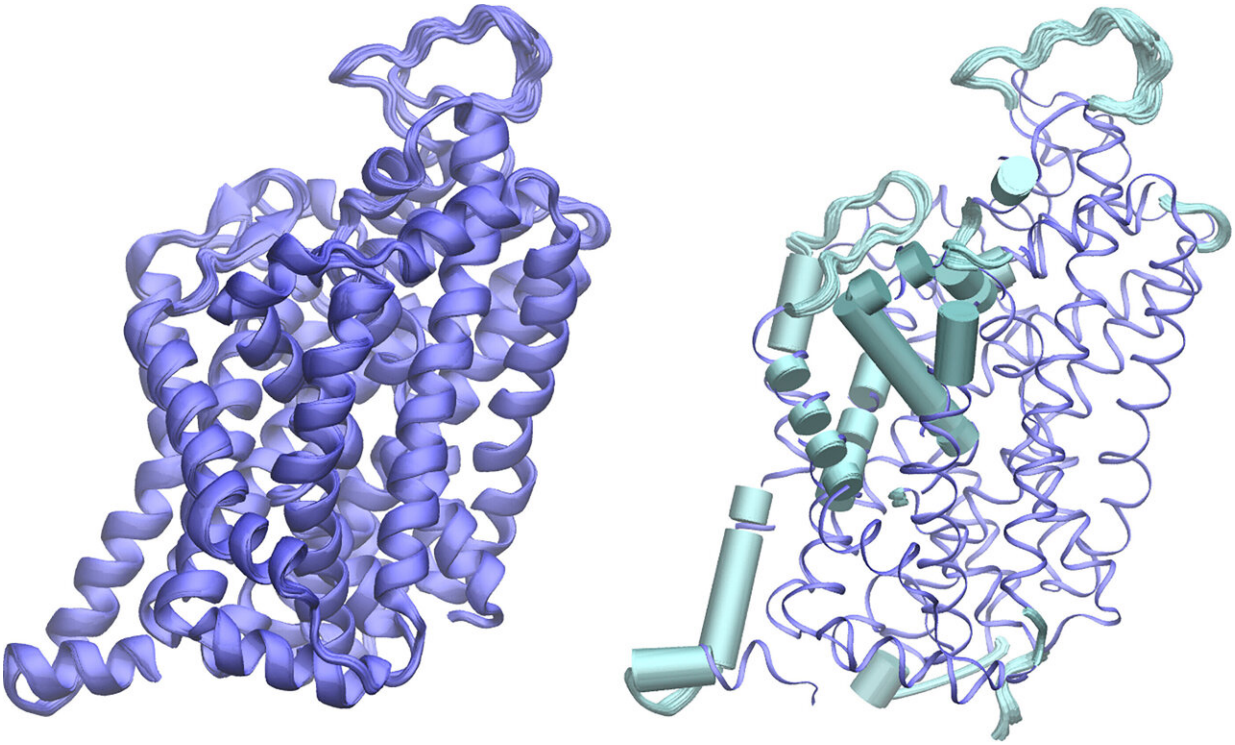
Supercomputing in the age of AI to accelerate protein structure prediction

June 28 2024



Protein structure used to test APACE: serotonin transporter (PDB accession: 6AWO; shorthand SERT). The Left panel is 100 SERT predicted conformational ensemble overlaid, which has good agreement with ground truth SERT. The Right panel is high variant transmembrane domains, shown in cyan, and computed with root mean square fluctuations overlayed. Credit: *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2311888121

For researchers, using high-performance computers can be a little intimidating. Understanding the best interface to use, how to make software scale, and working with huge datasets requires its own expertise.

Fortunately, NCSA does more than deploy and operate these powerful systems. The center is home to the Scientific and Engineering Applications Support (SEAS) team, which helps researchers make efficient use of hardware and software resources available at NCSA.

Working with SEAS, researchers can get help installing Python packages, learn to select the best parallel computation engines for their project or—thanks to breakthrough work published in the journal *PNAS*—learn to successfully deploy artificial intelligence models. The paper is titled "APACE: AlphaFold2 and advanced computing as a service for accelerated discovery in biophysics."

The *PNAS* research paper, authored by Roland Haas, a senior research programmer in the SEAS group, Eliu Huerta, lead for translational AI at the U.S. Department of Energy's (DOE) Argonne National Laboratory and CASE senior scientist at the University of Chicago, Hyun Park, then an Illinois Ph.D. student in biophysics, and Parth Patel, an NCSA graduate research assistant, describes a novel computational framework that simplifies and speeds up the process of using AI tools and algorithms to understand three-dimensional protein structure.

The framework also predicts conformational diversity of proteins, an important property since proteins are malleable structures that can flip between different conformations to do their job.

The team developed APACE, a computational tool that effectively handles AlphaFold2, an AI program used to predict protein structure on high-performance computing systems. They deployed APACE on the

Delta supercomputer at NCSA to measure how well it performed predicting the structures of four exemplar proteins.

Using up to 300 ensembles distributed across 300 NVIDIA A100 GPUs, they found that APACE is up to two orders of magnitude faster than off-the-shelf AlphaFold2 implementations.

Moreover, the same approach could be used in a variety of scientific disciplines and could be linked with robotics laboratories to automate and accelerate scientific discovery. The team later reproduced the work on the Polaris supercomputer at the Argonne Leadership Computing Facility, a DOE Office of Science user facility.

"Foundation AI models have the potential to transform the practice of science if they are findable, accessible and ready to use by the broader scientific community," said Huerta. "This project demonstrates how to create and share the required scientific data infrastructure to truly democratize cutting-edge AI and leverage modern computing environments to maximize its science reach."

Biomedical researchers study proteins to understand a wide range of biological functions. Proteins are chains of amino acids and their ordering into 3D structures determines biological functions.

Understanding how proteins are formed—a process often called protein folding, in which amino acids come together in structured chains capable of carrying out specific functions—is crucial to understanding normal biological functions as well as how folding mistakes can lead to serious diseases.

Predicting protein folding is extremely computationally intensive since a typical protein can have hundreds of amino acids and thousands of cells that can combine in different ways.

The usual methods for studying protein structure are X-ray crystallography, a tool for determining the atomic and molecular structure of a crystal, and cryo-EM, which involves flash-freezing molecules in liquid nitrogen and bombarding them with electrons to capture their images with a special camera.

AlphaFold and AlphaFold2 showed that AI software can accurately and quickly predict protein structure from amino acid sequences, and the development of APACE builds on this breakthrough.

APACE optimizes AlphaFold2 to run at scale on high-performance computing platforms and effectively handles its multiple-terabyte protein database. The work shows that large AI models can be combined with the power of high-performance computing to allow scientists to study multi-protein complexes and obtain results quickly, accurately and at higher resolution—all factors that could lead to a fuller understanding of protein structure and kickstart the development of new drugs that can treat many diseases.

"Research in new drugs is extremely time-consuming and bottlenecked by the need to synthesize different candidate compounds to test their medical effectiveness in a laboratory," said Haas.

APACE allows drug researchers to drastically reduce the time required to screen out potential candidate compounds and thus focus on the most promising substances. This way, more compounds can be tested and the time to develop a new drug, for example, one tailored towards a specific viral strain, can be reduced.

A key feature of APACE is better data management, which is achieved by hosting AlphaFold2's multi-terabyte model and database on the supercomputer, from which the framework's neural networks can readily access data. Other improvements include CPU optimization and GPU

optimization to parallelize GPU-intensive neural network protein structure prediction steps.

"The first problem with using an AI model is the storage of the data," said Park, who, like Patel, was at Argonne for an internship when the work on APACE was done.

"We need to pass 2.6 terabytes (the size of the AlphaFold2 database) as well as the computation from sequence to structure prediction. Some university labs may be able to do that, but what matters is that you scale it up so that scientists around the world can use it."

Patel added, "That's why HPC utilization is important, especially for AI models. Anyone who can get into an HPC system can have access to both data and also the computational capability to do the actual AI model calculation. Not to mention, there's a huge speed increase."

Huerta said the team chose to work with AlphaFold2 because it is used extensively in different research communities, including biophysics, chemistry, and drug design and discovery.

"APACE provides all the capabilities of the original AlphaFold2 model and empowers researchers with the ability to leverage supercomputers to reduce time-to-solution and to connect this tool with self-driving laboratories to automate and accelerate discovery," he said.

Huerta said the team will continue to build a community of APACE users to maximize the usability of AI models with HPC platforms. Haas said the team is now focused on attacking the remaining bottlenecks in the system to further improve speed. He'd also like to make APACE available on more compute clusters so that more scientists can take advantage of it.

"We'd also like to explore using the methods we've developed to speed up Alphafold2 with other foundational machine learning models that are too complex to easily use on common desktop workstations," said Haas. "It's all about making the best tools available and as easy to use as possible."

More information: Hyun Park et al, APACE: AlphaFold2 and advanced computing as a service for accelerated discovery in biophysics, *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2311888121

Provided by National Center for Supercomputing Applications