Cerebras Raises $1.1B Series G at $8.1B Valuation: Read the Press Release

DEC 14 2021

A Big Chip for Big Science: Watching the COVID-19 Virus in Action - Cerebras

VISHAL SUBBIAH

It's hard to imagine a better example of "AI for good" than figuring out how the virus that causes COVID-19 works. If we know how it works, we can develop ways to prevent the virus replicating and end a global scourge. I was fortunate to co-author, along with my colleagues Jessica Liu and Tanveer Raza, a massive scientific study to figure out the "how". The study was nominated for a Gordon Bell Special Prize and presented at this year's SC21 supercomputing conference.

This is big science. Researchers from 12 national labs, universities, and companies like Cerebras developed a host of new computational techniques to create a simulation of the virus' replication mechanism that runs across 4 supercomputing sites!

The idea was to create a fully functional model of the SARS-CoV-2 virus "replication-transcription machinery". The word "machinery" is apt: this is an intricate biological mechanism made up of millions of atoms moving in three dimensions as it hijacks the host's own replication mechanism to make copies of itself.

The process starts with three-dimensional images of the virus captured using cryo-electron microscopy. This technique can achieve near-atomic resolution, but the images are still not good enough, or dynamic enough, to show us how the mechanism really works. To fill in the missing data, the research team layered on top two completely different, but complementary techniques, working at different scales. First, we can treat biomolecules the way we treat any materials problem. We can use a type of the finite element analysis tools we routinely use to design continuum-scale objects such as engine parts. And second, we can simulate molecules atom-by-atom like a much more sophisticated version of the ball-and-stick els we all remember from chemistry class.

TSMC-1082
TSMC Ltd.
IPR2025-01212

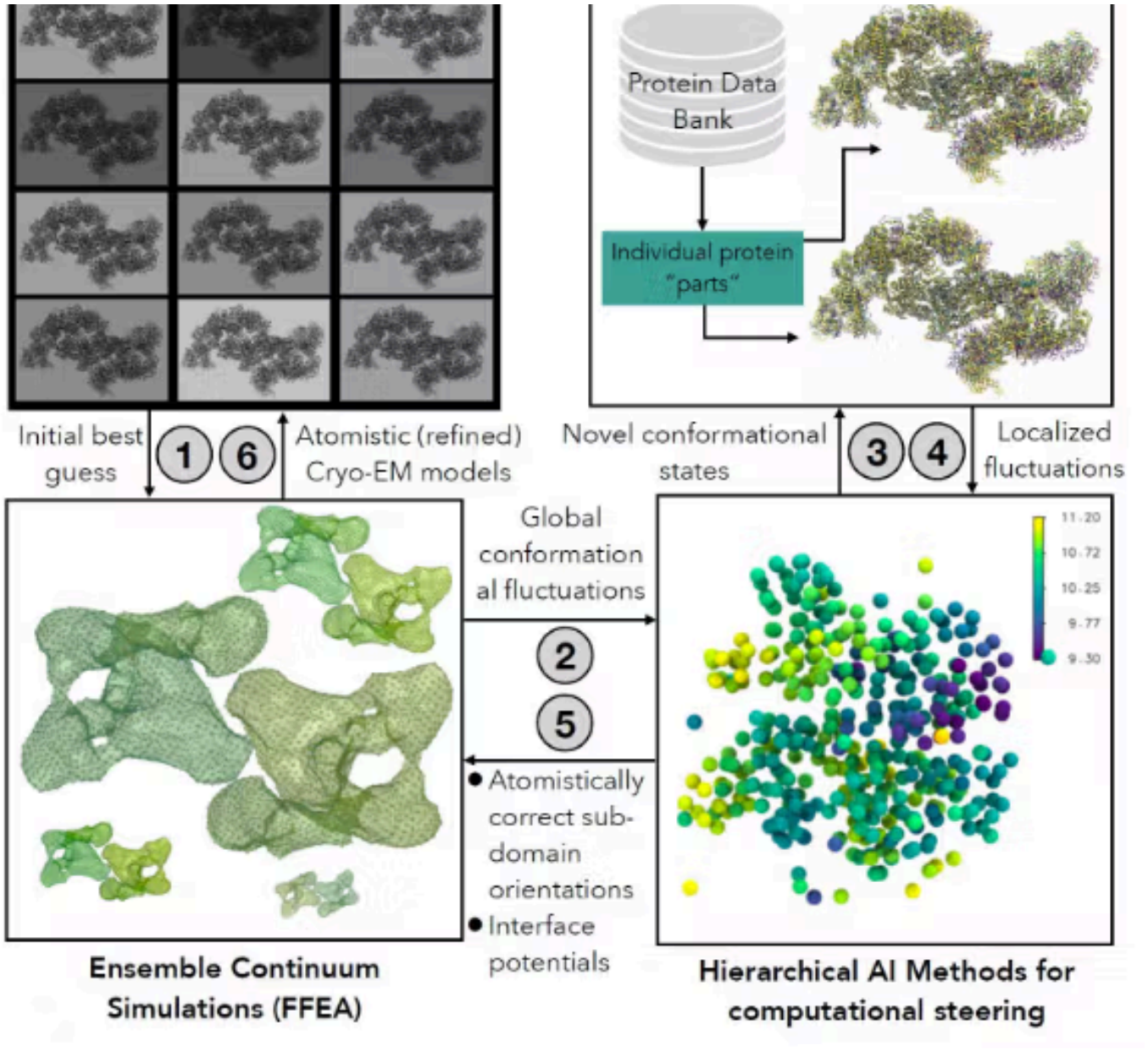


Diagram from the paper showing how the components of the study fit together. Our work is in the computational steering part.

Putting all this together is a mammoth task. Innovation was needed at, as it were, every scale, from a novel workflow architecture that allows widely-distribute computing resources to mesh seamlessly and automatically, to improving the computational efficiency of the individual models.

That last part – improving computational efficiency – is where Cerebras comes in. In the past, `e simulations took so long to create that it was only possible to study a few tens of ɔseconds of motion at one time. However, to reach a broader understanding, they

What role, exactly, does ML play here, I hear you ask? Each simulation experiment ties up a supercomputer with thousands of processing nodes for a long time. To avoid wasted time, it's vitally important to "steer" these experiments, recognizing and halting simulations that are going down dead ends, and encouraging, so to speak, those that may prove fruitful. This is easier said than done. It's very difficult to specify the characteristics of a "bad" simulation beforehand. But it's easy after the fact to recognize that a bad thing happened. This is a classic ML opportunity: you know what the answer looks like, but you don't know how to define rules to describe it.

We address this with a machine learning model called a "convolutional variational autoencoder", or CVAE. Oversimplifying, a CVAE takes a complex "high-dimensional" input and transforms or "encodes" it into a smaller form. You can think of this as a kind of figure of merit. We train the model by letting it observe snapshots of the simulations. We then run the reverse transformation – or decode it. If the decoded version is a good match for the original, we know the CVAE is working. That trained model can then be used during "real" experiments by another algorithm that does the actual steering. However, as the paper points out: "CVAE is quadratic in time and space complexity and can be prohibitive to train."

Cerebras comes into the picture here because this bit of the problem was being explored at Oak Ridge National Laboratory on the Summit supercomputer and on the Argonne AI-Testbed at Argonne National Laboratory, which just happens to feature a Cerebras accelerator. The ANL researchers compared training their CVAE model on 256 nodes of Summit, for a total of 1,536 GPUs, and on a single Cerebras CS-2 system.

And how did we do? In terms of pure performance, rather well. Quoting the paper again: "the CS-2 delivers out-of-the-box performance of 24,000 samples/s, or about the equivalent of 110-120 GPUs."

As impressive as this number is, perhaps even more impressive is the "out of the box" comment. Distributing a promising algorithm across a large cluster of compute nodes is difficult and time-consuming even for experts in the field. The CS-2 system, by contrast, is intentionally architected as a single, ultra-powerful node with cluster-scale performance. Our software makes it easy to get a neural network running by changing just a couple of lines of code.

Many organizations have problems that could be solved with some serious AI horsepower, but the sad fact is that few of us have the funds to construct or run supercomputers to run on. And moreover, few of us have the specialized developers capable of rewriting and

To quote the paper again: “Because a single CS-2 here delivers the performance of over 100 GPUs, it is a practical alternative for organizations interested in this workflow who do not have extremely large GPU clusters.” We couldn’t agree more.

Finally, it’s important to bear in mind that while this study has direct benefits in the treatment of COVID-19, the new tools and workflow may ultimately prove much more significant. This methodology can be applied to any kind of molecular machinery, paving the way for more rapid and better understanding of molecular interactions across a wide range of use cases, including treatment discovery for a range of diseases. It’s hugely satisfying to know that I was able to play a part.



To learn more about the study, read the paper “Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action” which will appear in International Journal of High Performance Computing Applications, 2021.

ıer image by Argonne National Laboratory/University of Illinois at Urbana-Champaign.