TRACKING VIRUS VARIANTS WITH AI

ARGONNE NATIONAL LABORATORY RESEARCHERS WIN GORDON BELL SPECIAL PRIZE

About the Customer

Argonne National Laboratory is a U.S. Department of Energy multidisciplinary science and engineering research center, where talented researchers work together to answer the biggest questions facing humanity. The Aurora exascale computer, scheduled to be operational at the Argonne Leadership Computing Facility (ALCF) in 2023, will leverage several technological innovations to support cutting-edge machine learning and data science workloads alongside more traditional modeling and simulation. In the run-up to exascale with Aurora, Argonne's Polaris system is already enabling advances in a variety of scientific and research projects.

"

We're excited about the collaboration we have with Altair, and we can't wait to see what the future holds on Polaris as we address future challenges… not to mention the increased impact it will have once PBS Professional is running on our forthcoming, next-generation Aurora supercomputer.

Bill Allcock, Director of Operations, ALCF at Argonne National Laboratory

Argonne National Laboratory researchers, with university and industry collaborators, won the 2022 Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research for their work using AI to track virus variants.

Their Challenge

The COVID-19 pandemic has impacted the entire planet, and researchers continue to investigate its catalyst: the SARS-CoV-2 virus and its variants. "It's a process of evolution of the virus inside of the human cell," said Arvind Ramanathan, computational biologist, Argonne. Because viruses can evolve rapidly, sometimes becoming deadlier and more transmissible in the process, it's important to quickly determine which variants are variants of concern (VOCs). Discovering VOCs quickly can save lives by giving scientists time to develop effective vaccines and treatments — but existing methods of tracking variants can be slow.

Our Solution

A team of researchers at Argonne National Laboratory, along with university and industry collaborators, tackled the problem of tracking virus variants by using artificial intelligence (AI). The ALCF's powerful Polaris supercomputer, along with Cerebras' AI-hardware accelerator and NVIDIA's GPU-accelerated Selene system, enabled the research. Polaris is also equipped with GPUs and with workload orchestration by Altair® PBS Professional®. Efficient workload management is critical for large, complex workloads. "A key challenge in this problem is dealing with long sequence lengths and tackling these foundation models at the scale of the viral genome," Ramanathan said. Polaris can handle the load, including the year's worth of genome data the team used for the project.

The project team's work training large language models (LLMs) for the job has implications beyond COVID-19, and for their work modeling how pandemic-causing viruses — especially SARS-CoV-2 — are identified and classified, they won the ACM's prestigious 2022 Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research. Machine learning played an important role in the research, and the researchers analyzed 1.5 million complete, high-quality SARS-CoV-2 genome sequences — a process that would have previously been time- and labor-intensive, individually examining every protein and mapping mutations. Instead, the team streamlined the process by developing the first genome-scale language model (GenSLM).

Results

The prize-winning research team was recognized at SC22 in Dallas, and their paper will be published in the International Journal of High-Performance Computing Applications (IJHPCA). The results the Argonne researchers and their collaborators have achieved paves the way for faster, more detailed insight into the virus mutation process, enabling scientists all over the globe to act on emergent variants and develop ammunition to reduce severity and slow the spread, ultimately saving lives. There's much more work to be done, and it's applicable well beyond COVID-19. Excitingly, according to the team, their effort's "full potential on large biological data is yet to be realized."