Department: Leadership Computing
Editors: James J. Hack, jhack@ornl.gov
Michael E. Papka, papka@anl.gov

# The U.S. High-Performance Computing Consortium in the Fight Against COVID-19

**James J. Hack**
Oak Ridge National Laboratory

**Michael E. Papka**
Argonne National Laboratory

***Abstract*—U.S. computing leaders, including Department of Energy National Laboratories, have partnered with universities, government agencies, and the private sector to research responses to COVID-19, providing an unprecedented collection of resources that include some of the fastest computers in the world. For HPC users, these leadership machines will drive the AI to accelerate the discovery of promising treatments, enable at-scale simulations to understand the virus's protein structure and attack mechanisms, and help inform policymakers to deploy resources effectively.**

**Beginning at the** time the World Health Organization declared COVID-19 a pandemic in mid-March 2020, the research community has undertaken an immense global collaboration to suppress SARS-CoV-2, the novel coronavirus that causes the disease. The two biggest objectives in COVID-19 research: development and evaluation of vaccines and therapeutic treatments, and predicting possible trajectories of future outbreaks by developing a clear understanding of how this disease is transmitted.

Governments and public health officials worldwide are struggling to control a virus that is already widespread and easily transmissible, including by asymptomatic carriers. Countries with the advanced research facilities and infrastructure, ready experimental pipelines, and the participation of partnerships between entities such as biotechnology companies, academic

*Digital Object Identifier 10.1109/MCSE.2020.3019744*
*Date of current version 9 October 2020.*

research laboratories, and government national laboratories—all of which the U.S. has rapidly put in place—have a distinct advantage in shaping that response.

Early in the battle to understand, treat, and control the COVID-19 pathogen, a unique private-public effort was established, led by the White House Office of Science and Technology Policy, the U.S. Department of Energy (DOE), and IBM to bring together federal government, industry, and academic leaders who would make computational resources available to the COVID research community. The goal of the *COVID-19 High Performance Computing Consortium* ("The HPC Consortium") was to help accelerate the pace of scientific discoveries in COVID-19 research projects.

The consortium brings together more than 30 partners from 7 national labs; NASA; 8 National Science Foundation-funded organizations, notably XSEDE, which is processing the requests for consortium resources, and the University of Texas at Austin's Texas Advanced Computing Center, one of the consortium's leading providers of computing cycles; 11 technology companies and 14 academic institutions, and provides resources ranging from small clusters to cloud and web services to high-end supercomputers.

The DOE's Office of Science operates many of the world's largest user facilities enabling the most cutting-edge coronavirus research studies today, such as efforts to understand its structure and underlying biochemistry, and evaluating possible therapeutics. Among those user facilities are DOE's Leadership Computing Facilities, located at Oak Ridge National Laboratory and Argonne National Laboratory. Oak Ridge's Summit IBM supercomputer and Argonne's Theta Intel Cray supercomputer are currently enabling the most cutting-edge coronavirus research studies today, powering research in molecular modeling, bioinformatics, and epidemiology to help accelerate the development of treatments and strategies to combat the COVID-19 pandemic. Access to these unique computational ecosystems has been made available through proposal mechanisms established by the HPC Consortium where the capabilities include access to state-of-the-art modeling algorithms, data analysis tools, and huge numbers of CPUs and GPUs. While relatively few consortium members are able to exploit all of these capabilities, the inclusion of leadership computing machines in this arsenal represents a significant commitment of technologies that are suited to run simulations at the scale necessary to significantly accelerate knowledge on how to control a global epidemic, and to address rapidly evolving advances in our understanding of the virus. Consortium projects include running complex biophysics investigations at scale to help understand what treatment protocols are and are not working; rapidly screening for possible therapeutics using machine learning and deep learning techniques; and running agent-based models to run scenarios for how the virus will spread given evolving guidance on social distancing and other hygienic behavior.

This article highlights several HPC Consortium projects, and other related activities, actively being supported at DOE's Leadership Computing Facilities and how these machines are being used to accelerate the science needed to develop treatments and strategies to combat COVID-19.

## COVID-19 PUTS HIGH-END COMPUTING ARCHITECTURES TO THE TEST

Summit's large array of very powerful hybrid CPU-GPU nodes—more than 4600 nodes, each containing 44 Power 9 cores and 6 NVIDIA Volta GPUs—make the platform especially well-suited to some specific but important types of investigations related to the pandemic. First, molecular dynamics (MD) simulations of the virus and its interactions with human cells can be carried out on Summit at scales and speeds that are unobtainable on other resources. In addition, the use of machine learning techniques to study how potential therapeutics interact with one another, and with other drug compounds, can take advantage of the unique AI features of Summit's GPU architecture.

Theta is a Cray XC40 manycore system built on the Intel KNL 7230 processor with 64 cores each. The system is comprised of 4392 nodes for a total core count of 281 088. The system is well suited for efforts that rely on the x86 instruction set. In addition, Argonne has enabled near-real-time response queues for Theta by exploiting drain time of the system—this allows for coupling with other unique instruments like Argonne's Advanced Photon Source (APS) for

on-demand analysis. This capability, along with Theta's traditional support for modeling and simulation, make it an important resource in supporting research for the pandemic.

## COMPUTATIONAL ADVANCES IN EPIDEMIOLOGICAL MODELING

- Argonne computational scientist *Jonathan Ozik* and his team have developed CityCOVID, an agent-based model capable of tracking detailed COVID-19 transmission. Agent-based modeling is an approach for capturing the dynamics of heterogeneous, interacting, adaptive agents at an individual, granular level of detail. When applied to a city such as Chicago, CityCOVID includes a synthetic population representing the 2.7 million residents of Chicago and the 1.2 million geolocated places where they can colocate. Throughout a simulated day, each individual, or agent, moves from place-to-place, hour-by-hour, engaging in social activities and interactions with colocated agents, where COVID-19 exposure events can occur. The COVID-19 disease progression is modeled within each individual, including differing symptom severities, hospitalizations, and age-dependent probabilities of transitions between disease stages. Using Argonne computing resources, CityCOVID is being used to calibrate unobserved model parameters, such as the time-varying degree of individual self-protective behaviors across the population, and to simulate a variety of interventions and future scenarios. While the Argonne team has been using Chicago as a testbed for developing these capabilities, CityCOVID is being extended to other regions as well.
- University of West Florida computer science professor *Ashok Srinivasan* is using Argonne resources to analyze the spread of COVID-19 during air travel through pedestrian dynamics simulation. Srinivasan's work attempts to better understand the impact of new boarding processes on exposure to the virus. To do that, Srinivasan developed a computationally efficient constrained linear movement model and code and carried out a large parameter space sweep to simulate realistic boarding scenarios. The simulation results show that back-to-front boarding roughly doubles the infection exposure compared with random boarding. It also increases exposure by around 50% compared to a typical boarding process prior to the outbreak of COVID-19.
- Purdue University professor *Yung-Hsiang Lu* and a team of engineers built a website that pools together live footage and images from approximately 30 000 network cameras in more than 100 countries to study social distancing, making data easier to analyze. The site has documented footage since March 2020 that could help evaluate the effectiveness of lockdowns and restrictions. The project uses Argonne computational resources and storage.

## ADVANCED MODELING AND SIMULATION

- *G. Andrés Cisneros*, at the University of North Texas, is using Summit to investigate inhibitor mechanisms of existing drugs for the RNA dependent RNA polymerase (RDRP) protein to provide insights that could serve to improve treatment options for COVID-19. He and his team are using classical MD techniques to investigate the structure and dynamics of the protein with and without inhibitors. He is also looking at Remdesivir, a broad-spectrum antiviral medication, and several other drugs, to understand how these protein inhibitors are useful in battling COVID-19 infections. Since protein inhibitors can be used to prevent viruses and bacteria from reproducing, their plan is to study the interactions and reaction mechanisms at the atomic level, something that cannot be done in a traditional lab.

## RAPID ANALYSIS COUPLED WITH EXPERIMENTAL REFINEMENT

- A team of researchers at Oak Ridge and the University of Tennessee, Knoxville (UTK) led by *Jeremy Smith*, has already used Summit to build a model of the coronavirus' spike protein, also called the S-protein, based on early studies of the protein structure. Using MD

simulations, they explored approximately 800 000 different compounds' docking properties to the spike to determine if any might prevent it from binding to human cells. They ultimately identified 77 small-molecule drug compounds—medications, natural compounds, etc.—shown by the simulations to bind to regions of the spike that are important for entry into the human cell and, therefore, might interfere with the infection process. The team then took this work further using more complete descriptions of the virus structure, one of which was determined using x-ray crystallography by a team from the Spallation Neutron Source at Oak Ridge and a team at Argonne's APS. In a first-of-its-kind measurement, the protease was determined from room-temperature crystallized protein samples to achieve a more accurate three-dimensional model of the protein (see Figure 1). This provided the opportunity to computationally explore compounds that can be shown to bind to the protein, which might block the virus's replication mechanism. Using the improved protein structures, the Smith team performed MD simulations to computationally screen 1.5 billion chemical compounds for COVID-19 in only 24 h, providing a wealth of data for SARS-CoV-2 drug discovery efforts. Because the virus's main protease is an important drug target, the simulations might lead to an understanding of which drugs could be successfully repurposed for COVID-19.

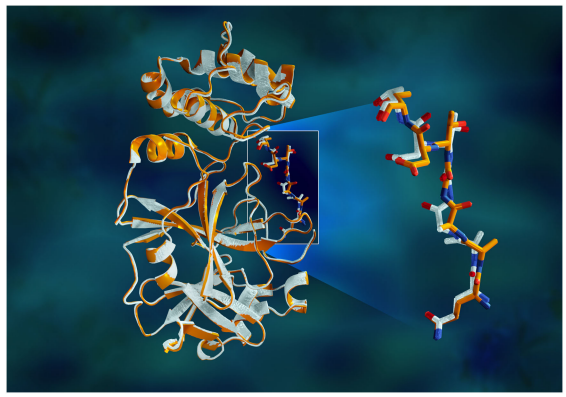

**Figure 1.** Overlapping x-ray data of the SARS-CoV-2 main protease shows structural differences between the protein at room temperature (orange) and the cryogenically frozen structure (white). This work is being used by Jeremy Smith and his team who are conducting drug docking simulations using Summit. Credit: Jill Hemman/ORNL, U.S. Dept. of Energy.

- Scientists from Argonne's Structural Biology Center at the APS, including Argonne Distinguished Fellow *Andrzej Joachimiak*, and working with *Mateusz Wilamowski* of the Center for Structural Genomics of Infectious Diseases (CSGID), performed experiments on the Nsp10+Nsp16 protein complex of SARS-CoV-2 created at CSGID. The experiment provided the first low-dose, room-temperature insight into the structure of this protein complex. Its results will give the community greater biological insight into the complex than is possible with traditional crystallography techniques. The experiment is designed to determine the metal activation of the complex and later lead to time-resolved experiments, which will be the first dynamic structural experiment of a SARS-CoV-2 related protein. Time-resolved structural dynamics will help elucidate the electrochemistry of this protein function and give insights into the virus. To support the rapid processing requirements, a team led by computer scientists *Ryan Chard* and beamline scientist *Darren Sherrell* deployed an automated data acquisition, analysis, curation, and visualization pipeline, leveraging Theta for high-speed on-demand analysis. The pipeline reactively analyzes data as it is collected, moving images of the sample from the APS to the Argonne Leadership Computing Facility where they are rapidly analyzed and visualized. The same automated pipeline then moved results to a repository and extracted metadata for publication in a data portal, which scientists can monitor during an experiment.
- A team led by *Albert Lau*, assistant professor of biophysics and biophysical chemistry at the Johns Hopkins School of Medicine, is using a joint computational and experimental approach to identify FDA-approved drugs that possess antiviral activity against SARS-CoV-2. Theta is being used to screen a library of compounds that includes existing FDA-approved drugs against ten intracellular catalytic SARS-CoV-2 protein targets for

binding, with the goal of identifying drugs that disrupt viral protein function and diminish viral viability. Reducing the number of compounds and protein targets to an experimentally manageable number will help the team establish priorities for the subsequent workflow, in which target proteins will be expressed and purified, and activity assays will be developed. Finally, prioritized compounds in the Johns Hopkins Drug Library will be screened against the purified target proteins to identify those that exhibit antiviral activity.

- A multi-institutional team from the UTK, the Yale School of Medicine, the U.S. Department of Veterans Affairs, the Versiti Blood Research Institute, the University of Kentucky and the Cincinnati Children's Hospital Medical Center, led by Oak Ridge computational biologist *Dan Jacobson*, has performed data analyses on Summit to analyze samples of lung fluid cells from COVID-19 patients. Analysis of COVID and control patients is aimed at finding gene expression and coexpression patterns that may explain the runaway symptoms produced by the body's response to SARS-CoV-2. The team required the power of Summit to run 2.5 billion correlation calculations that helped them understand the normal regulatory circuits and relationships for the genes of interest. With Summit, the team completed the calculations in one week rather than spending months doing them on a desktop computer. Early results have found that genes related to one of the body's systems responsible for lowering blood pressure—the bradykinin system—appear to be excessively "turned on" in the lung fluid cells of those with the virus. A bradykinin storm could explain the wide variety of symptoms experienced by COVID-19 patients, such as muscle pain, fatigue, nausea, vomiting, diarrhea, headaches, and decreased cognitive function. At least ten existing drugs are known to act on the specific pathways the team studied, but large-scale clinical trials are needed to determine whether they might be effective at treating COVID-19.

A bradykinin storm could explain the wide variety of symptoms experienced by COVID-19 patients.

## RAPID SCREENING USING AI TECHNIQUES

- The many clinical manifestations of COVID-19 have so far proven to be resistant to existing single drug therapies. A team led by *Jennifer Diaz* of the Icahn School of Medicine at Mount Sinai has developed a machine learning classifier to predict whether certain drug combinations might be effective against the virus. The team is using the classifier on Summit to predict gene expression patterns for more than 700 000 combinations of existing drugs. In a preliminary analysis, Diaz's team identified U.S. Food and Drug Administration (FDA)-approved drugs that are already being studied to treat COVID-19 as well as novel drugs that may have a synergistic effect when combined. Studies have suggested these synergistic gene expression predictions can be used to identify biological pathways and processes that will be altered for each combination. The team plans to expand the database, improve the classifier, and use gene set enrichment analysis to make predictions of drug pairs that may be a synergetic COVID-19 treatment. Predictions from this unique analysis must be validated in wet labs or clinical trials to determine their viability and efficacy. Successful laboratory validation will allow a broadening of the project scope to explore highly focused drug combinations in the millions.
- An Argonne-based team is seeking to address both the fundamental biological mechanisms of the virus and the disease, while simultaneously targeting the entire viral proteome to identify potential therapeutics. Argonne computational biologist *Arvind Ramanathan* and Argonne Associate Laboratory Director *Rick Stevens*, are leveraging Argonne and Oak Ridge leadership machines to design novel

therapeutics against SARS-CoV-2 using AI approaches that integrate information from experimental observations, and rigorous, physics-based virtual screening and molecular simulations, to identify viable drugs that can inhibit viral proteins. These AI approaches, based on advances in deep learning and reinforcement learning, are capable of predicting how strongly a small molecule will bind to a protein as well as explore the structural space of compounds that are predicted to bind to find more suitable variants.

The database of potential drug candidates for COVID-19 is immense, including millions to billions of potential compounds. So far, computational screening of small molecules has resulted in identifying molecules that can potentially inhibit viral function in wet lab experiments. These experiments, performed at the Argonne-located BSL-3 facility, involve live human lung cell cultures being exposed to small molecules followed by subsequent measurements that monitor viral replication. These molecules are being further refined to optimize them for binding to specific viral target proteins. Using AI techniques, the team has screened over 6 million small molecules and are validating them at Argonne for activity against the virus and rapidly expanding it to screen billions of compounds. The potential impact of this work is the design of new generative models based on reinforcement learning for both small molecules and antibodies; and development of large-scale, AI-driven simulations of the entire viral particle and drugs bound to the various viral targets, as a better pathway to an antiviral drug.

## ACKNOWLEDGMENTS

This article has been authored by UT-Battelle, LLC, operator of Oak Ridge National Laboratory, a U.S. Department of Energy Office of Science Laboratory, operated under contract number DE-AC05-00OR22725 and by UChicago Argonne, LLC, operator of Argonne National Laboratory, a U.S. Department of Energy Office of Science Laboratory, operated under contract number DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. http://energy.gov/downloads/doe-public-access-plan. The authors wish to acknowledge the contributions of Laura Wolf, Argonne National Laboratory, in the preparation of this article.

**James J. Hack** is currently the ORNL Computing and Computational Science Director of HPC Strategies, and was the former Director of the National Center for Computational Sciences at Oak Ridge National Laboratory, which houses the Oak Ridge Leadership Computing Facility, from 2007 to 2019. After earning a Ph.D. degree in atmospheric dynamics from Colorado State University, he was a research staff member at the IBM T. J. Watson Research Center working on HPC architectures, later moving to the National Center for Atmospheric Research, where he was a leader in the development of NCAR's global atmospheric model. His primary scientific interests include physical parameterization techniques, numerical methods and their implementation on HPC platforms, and diagnostic methods for evaluating simulation quality. A former editor of the *Journal of Climate*, he has testified before Congress on the topic of climate change, and has been involved in National Research Council studies on climate modeling. Contact him at jhack@ornl.gov.

**Michael E. Papka** is a senior scientist and member of the senior leadership at Argonne National Laboratory. He joined Argonne in 1992 and has served as Deputy Associate Laboratory Director for Computing, Environment and Life Sciences since 2006, and also as Director of the Argonne Leadership Computing Facility since 2010. He received the master's and the Ph.D. degrees in computer science from the University of Chicago. His research interests span visualization, large-scale data analysis, and the development and deployment of research infrastructure in support of science. He is a Presidential Research, Scholarship and Artistry Professor in the Department of Computer Science at Northern Illinois University and leads the Data, Devices, and Interaction Laboratory (ddiLab). Contact him at papka@anl.gov.