R&D

The Laboratory's Habit of INNOVATION

HIGH-PERFORMANCE computing (HPC) has joined theory and experiment to become the third pillar of scientific research at Lawrence Livermore. The Laboratory leads HPC advancements to meet the needs of the U.S. Department of Energy (DOE) and the National Nuclear Security Administration (NNSA) and to enable groundbreaking discoveries. In 2024, El Capitan—NNSA's first exascale-class supercomputer—will come online at Livermore, further cementing the Laboratory's leadership in HPC. This same leadership—from computational math and science to cybersecurity and software engineering—has enabled Laboratory researchers to win 10 R&D 100 awards in the Software–Services category in the past decade. "The large number of R&D 100 awards for Livermore's computing innovations is a nice testimonial to the quality and impact of our researchers and their work," says Bruce Hendrickson, principal associate director for Lawrence Livermore's Computing organization. "Our three winning projects this year are emblematic of our drive to anticipate and shape the future of computing and its impacts."

Building on Success

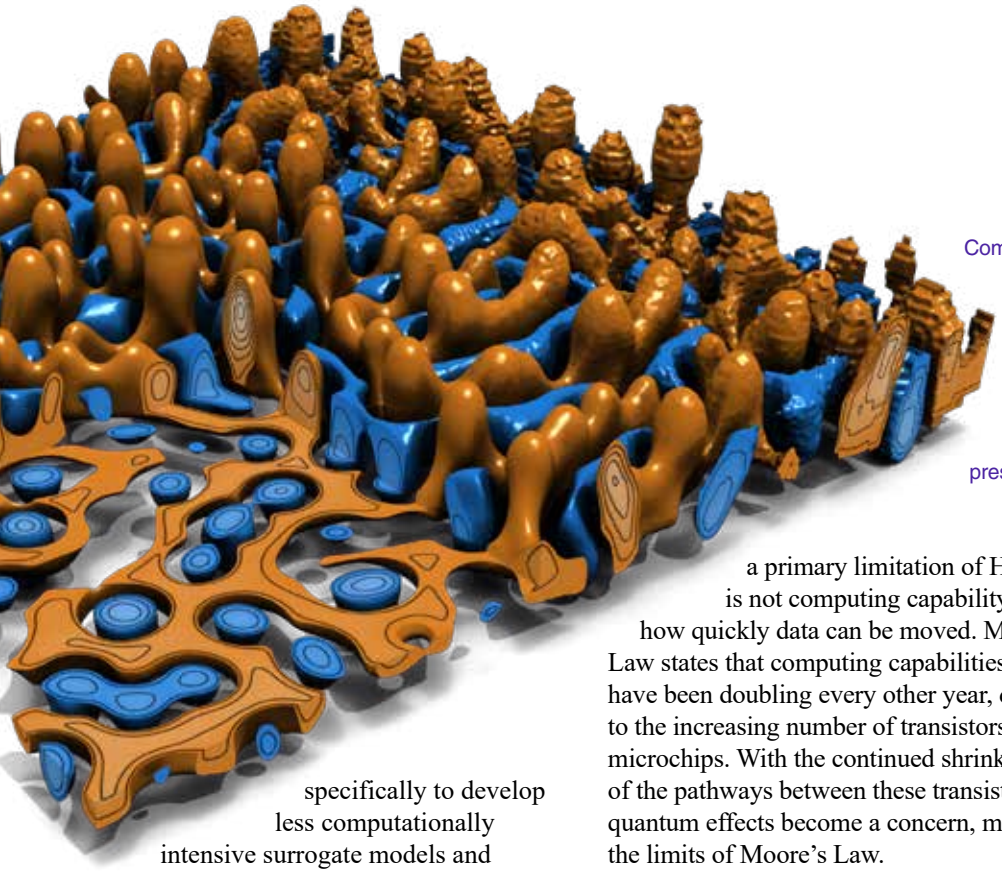
One example of Lawrence Livermore's past R&D 100 award success includes the Carbon Capture Simulation Initiative (CCSI), which has continued to mature since its 2016 win. CCSI, a collaboration among five national laboratories and five universities, developed a suite of tools and capabilities to curb atmospheric carbon

100

Lawrence Livermore's high-performance computing capabilities play a significant role in international science research and innovation.

dioxide emissions and reduce impacts from industrial processes. The CCSI toolset addresses key challenges in scaling up development and reducing the costs of preparing carbon capture technologies. As part of CCSI, Livermore team members developed FOQUS, a software to assess uncertainty quantification in carbon capture models. Charles Tong, one of the FOQUS developers, says, "With simulations, we don't know that what we predicted will happen because the simulation itself is imperfect. Uncertainty quantification recognizes that each prediction is a distribution instead of a single value, and it takes into account the range of options when developing a system with specific performance requirements." As one of the CCSI toolset's most critical products, FOQUS identifies the best carbon capture processes and rapidly determines the associated levels of uncertainty, as well as other related statistical analyses. Livermore's expertise in uncertainty quantification in weapons systems and HPC capabilities enabled the development of the framework for optimization and quantification of uncertainty and sensitivity—FOQUS.

CCSI has evolved into its next stage, the Carbon Capture Simulation for Industry Impact ($CCSI^2$) project to engage with stakeholder companies interested in developing effective carbon capture technologies, assist them with research and development, and help them apply what they have learned. Livermore has integrated machine-learning (ML) capabilities into uncertainty quantification,



Compressed floating point (ZFP) allows researchers to define error tolerance and control required accuracy. This simulation shows a Rayleigh–Taylor instability—the interface between two fluids with different densities, one in blue and the other in brown. In the simulation, the compression ratios—defined by the desired error tolerances—vary from 50 times on the left to 600 on the right. This results in varying amounts of detail preserved in the simulation.

specifically to develop less computationally intensive surrogate models and investigate different carbon capture system configurations. New ML capabilities have further optimized FOQUS's process to realize new uncertainty quantification applications, such as informing experimental design and guiding stakeholders to the next best set of experiments to validate and improve models.

Building on the success of such innovative past awardees as CCSI^2 and FOQUS, Lawrence Livermore's 2023 R&D 100 award winners exemplify the critical role the Laboratory plays in advancing HPC. Three of the 100 prizes in 2023 were awarded to Livermore-led or Livermore-supported HPC projects, all in the Software–Services category. Each winning software technology has its own unique, long-lasting impacts on the future of scientific research, from helping speed scientific calculations to making HPC systems more accessible to nontechnical users to accelerating cancer drug studies.

Compressing Compression Time

In recent years, scientific data sets have dramatically increased in size, and to keep pace, computing capabilities have had to advance in tandem. Despite this increase, a primary limitation of HPC is not computing capability but how quickly data can be moved. Moore's Law states that computing capabilities have been doubling every other year, due to the increasing number of transistors on microchips. With the continued shrinking of the pathways between these transistors, quantum effects become a concern, marking the limits of Moore's Law.

To meet Moore's Law's conditions, the solution has been to add more compute cores. However, additional cores are not automatically put to use upon installation. Developers must first determine how to implement parallelization, or the simultaneous running of multiple processes. Parallelization is not automatically built into the computing codes, and the process can be labor-intensive, especially when legacy code is in use. Unless the additional cores are used in parallel, adding the cores does not yield the expected benefits.

"If we don't have data to feed the cores, they're just going to sit there and wait," says Peter Lindstrom, who leads the compressed floating point (ZFP) data compression project at Lawrence Livermore. An immense amount of data movement is required to complete a computational task—between memory and registers, CPU and GPU, compute node and disk, and among multiple compute nodes. Reducing the volume of data that must be moved helps accelerate computations. If the data can be compressed quickly by taking advantage of the many compute nodes, the transfer will be much faster.

This need for faster computations was the initial motivation for ZFP: to reduce the amount of data that needs to be moved. "To accomplish this goal, we had to invent an entirely new compression capability that provides some level of random access," says Lindstrom. "Otherwise, the situation is similar to a cassette tape, in which the listener must play the tape through to find a specific track, versus a CD in which the listener can jump directly to the desired track."

ZFP introduces a new method of compressing large data sets while maintaining high-speed, on-demand access to the compressed data for both reading and writing applications—a capability not found in any other compressor. Researchers can continue to work with the data in real time while it remains compressed, whether they are analyzing it or producing visualizations. "Normally, it's just compress, compress, compress, and write the result out to disk. If researchers want to get back a piece of that data, they essentially have to start from the beginning and decompress the data until they find that piece they are looking for," says Lindstrom.

This ability to compress data while allowing researchers to continue working exists because of the way ZFP separates the data. "ZFP takes a large array of data and divides it into tiny chunks. Each chunk is compressed and decompressed entirely independently of all the other chunks," says Lindstrom. Accessing this data is normally a sequential process, but with ZFP, millions of chunks can be individually compressed and decompressed

at the same time, enabling quick access to a specific data point or range.

In addition to allowing random access to compressed data, ZFP offers a powerful method for achieving the compression itself. The software exploits redundancies in data using a principle similar to JPEG image compression, which breaks images up into blocks of pixels that appear to have shared properties. Most scientific data are similar between neighboring values, especially in their leading digits. In physics-based problems, for example, properties such as pressure, density, energy, and temperature often vary little from one point in space to another a short time and distance away. One of ZFP's unique capabilities is a decorrelation step, in which ZFP extracts the redundancy between spatially correlated data to avoid representing the same information multiple times, making it easier to store in fewer bits and more compressible. Furthermore, the software enables users to select the amount of accuracy required or the amount of compressed storage to allocate to an array, a capability unique to ZFP.

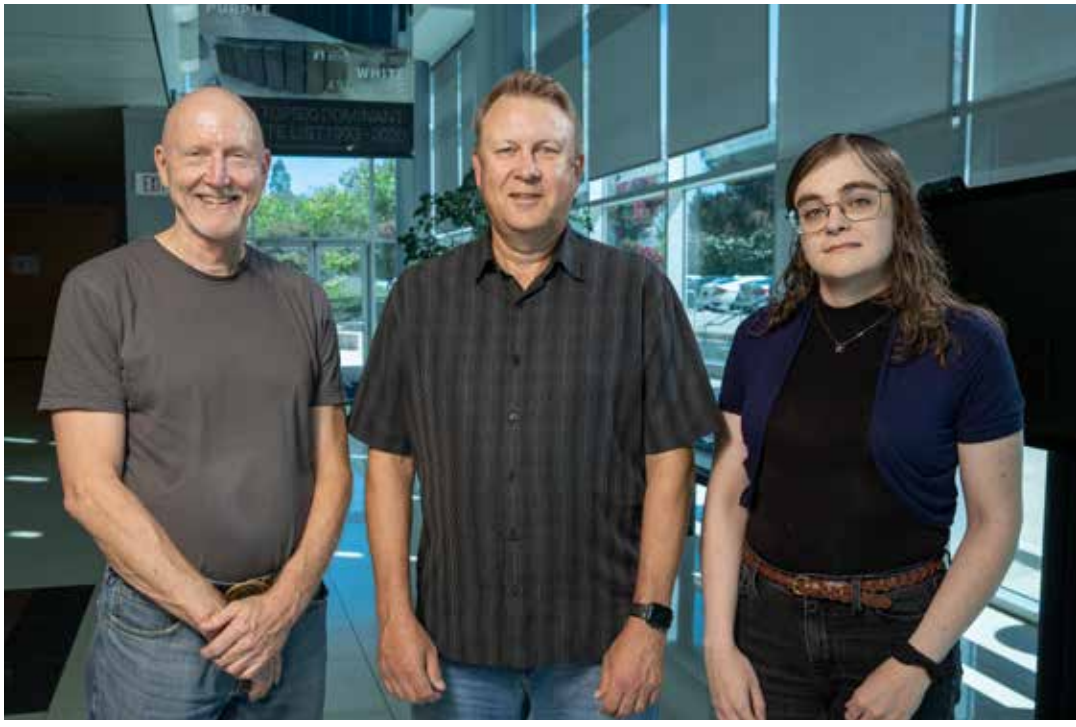
With these advances, ZFP provides solutions for time-sensitive use cases, while saving at least an order of magnitude in memory storage compared to conventional, uncompressed floating-point storage, and more than two orders of magnitude in storage for visualizations. ZFP is downloaded more than 1.5 million times per year by users from across the DOE and other government and nongovernment agencies, and its scientific applications include geographic information systems, climate science, seismology, and tornado simulations, among others. The software has also demonstrated value for nonscientific applications, particularly in photography and making high-resolution maps of the seafloor for sailors. "What ZFP buys is the ability to store data at much higher spatial and temporal resolution," says Lindstrom.

A Volume of Software Libraries

Supercomputers depend on a complex coordination of application codes, diverse hardware components, and system software to enable scientific discovery. Balancing the power, energy usage (up to 30 megawatts in some cases), and performance capability of these sophisticated systems is an essential aspect of improving their overall efficiency. Each hardware component has a unique set of dials that scientists, software engineers, and system administrators use to monitor and manage the parameters for each hardware component. However, these dials are often difficult to access, nonintuitive to the user, and are different among vendors and products. In an effort to better control these physical processes, Tapasya Patki, a computer scientist in Livermore's Center for Applied Scientific Computing (CASC), leads development of Variorum, a vendor-neutral software library for exposing and monitoring the power, energy, and performance of low-level dials across diverse architectures in a user-friendly manner. Variorum helps optimize energy usage and performance capabilities of applications, saving resources and time.

Named for a volume of work that collates variations among editions, Variorum combines the different interfaces into one simple interface, eliminating the need for users to understand the intricacies of each vendor and device in their system. Variorum's dials create an interface with the lowest level of programming bits accessible, which change the performance of an application.

Improving this manual process requires painstaking effort from the Variorum team through analyzing documentation; communicating with vendor partners; and determining the details, use cases, and limits for every processor and accelerator with which they work. Each time the documentation is changed, the team must thoroughly understand the revisions and determine how the Variorum code must be adjusted to maintain its compatibility



Development team for the ZFP data compression library: (from left) Mark Miller, Peter Lindstrom, and Danielle Asher.

across vendors. By doing this heavy lifting for the HPC community, Variorum helps prevent errors in the long run, so users do not have to discover vendor or device-specific subtleties and technicalities on their own. Patki says, "Many of these vendor-dependent factors are difficult details that are challenging to explain and explore, but the interface has made optimizing computing systems extremely easy for users to figure out."

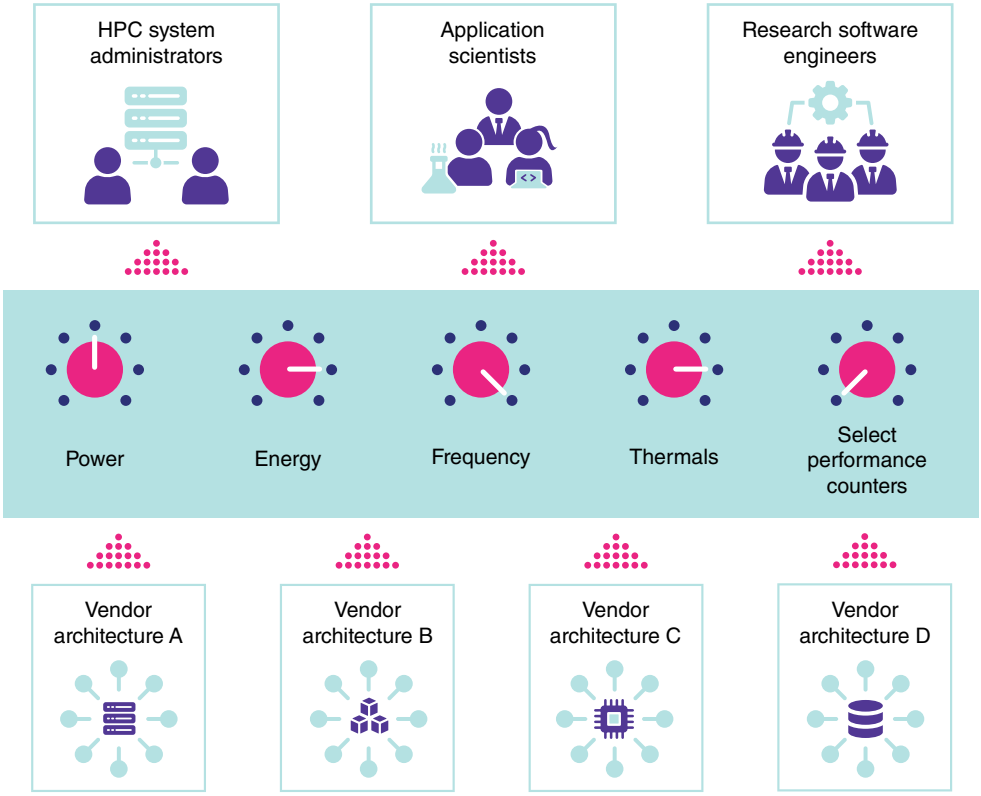
Integrating Variorum into an HPC workflow is simple. Patki says, "Users have to tell the interface what architecture they're building on, but they only have to do it in one place—the rest of their code stays the same." Regardless of the processors and accelerators applied within a given system, the user only needs to integrate a single interface into their own application. Moreover, since Variorum does not require users to know the details of their specific vendor implementation or understand how to interface with each vendor and each hardware device, the code they use to implement Variorum is easily portable between systems.

By offering monitoring solutions, Variorum can also achieve software optimization goals. Users can interface with Variorum during software execution to reduce the burden for scientific applications and workflows. "We designed Variorum for 95 percent of the users: people who want to do high-level power and thermal monitoring," says Aniruddha Marathe, another computer scientist on the Variorum team.

Each member of the team encountered the need to debug their physical monitoring code with every new feature in their own work, and this circumstance motivated them to work through the problem with one of their regular vendors, Intel. "Our vendor has many different processors and different processor generations. Each one of them has a different set of dials, so just within the same vendor, we saw tremendous variation," says Patki. The team then organized a large community of developers and researchers in scientific computing and discovered that these stakeholders were running into similar problems. The Livermore team realized the issue needed to be addressed more effectively, and decided to come up with a solution, which launched their efforts to develop Variorum.

In addition to simplifying the computing interface, Variorum has impacts in sustainable computing. By pushing HPC to its limits and ensuring the best possible use of computing resources, Variorum enables better utilization, overall energy savings, and improvements to the longevity of HPC systems. The software provides users with power management tools that are otherwise difficult to access, allowing them to monitor their energy efficiency and set usage caps to avoid massive fluctuations and grid exhaustion.

Open source and portable, Variorum can be applied not only by HPC systems but also for cloud computing and personal laptops. The system has provided support for all three DOE exascale computers: the upcoming El Capitan system at Lawrence Livermore, Aurora at Argonne National Laboratory, and Frontier at Oak Ridge National Laboratory. Patki says, "Software power and energy utilization



When different users such as high-performance computing (HPC) system administrators, application scientists, and research software engineers input their system architecture into Variorum, the technology extracts the details of the hardware's vendor-specific implementations and makes dials that enable measurement and control of various physical features on processors and accelerators.

Development team for Variorum: (from left) Stephanie Brink, Tapasya Patki, Barry Rountree, Elena Green, Kathleen Shoga, and Aniruddha Marathe.

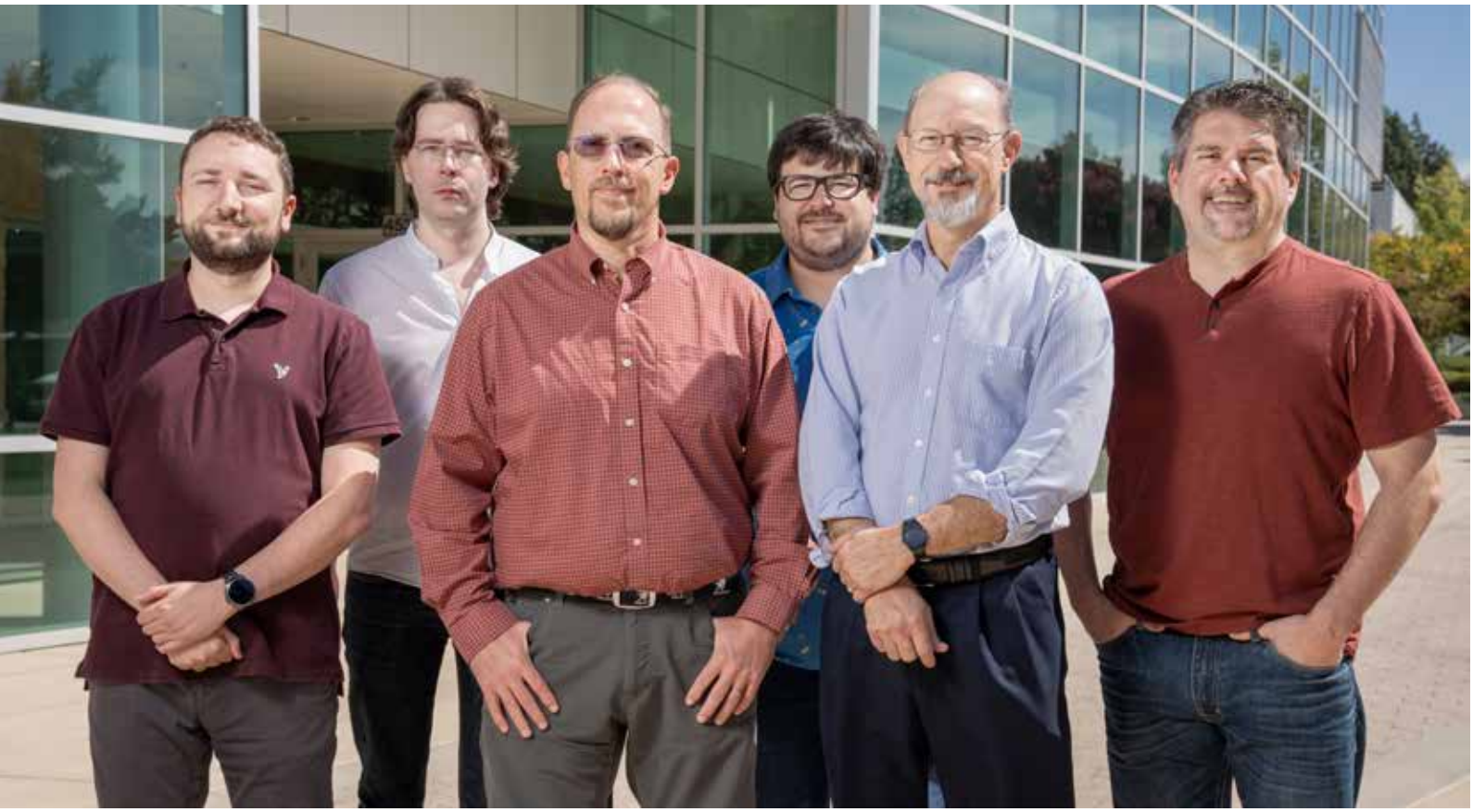
remains a relatively niche area, but the Variorum team will continue to maintain their relationships with vendors to ensure more accessible hardware monitoring."

Lighting a CANDLE to Cancer Research

An early adopter of using ML for scientific applications, the Cancer Distributed Learning Environment (CANDLE) provides ML capabilities for applications related to cancer research. In particular, CANDLE enables capabilities for extracting key information and finding relationships within large, disconnected data sets to help solve cancer-specific drug challenges. CANDLE is a collaboration among Lawrence Livermore, Los Alamos, Oak Ridge, and Argonne national laboratories; the Frederick National Laboratory for Cancer Research; the National Institutes of Health (NIH); and the National Cancer Institute (NCI). Lawrence Livermore's contribution to CANDLE, led by Brian Van Essen, a computer scientist in CASC, who is focused on developing artificial intelligence (AI) models for molecular dynamics.

At the time of CANDLE's inception in 2016, large-scale scientific ML techniques were in their infancy. CANDLE was initially the only AI project within the DOE's Exascale Computing Project and was designed to bring a deep learning lens to cancer research. Van Essen says, "We were thinking outside of what industry was doing, about how AI can be applied to precision medicine, and how we could leverage advanced computing systems. This research was an early demonstration of that."

The participating laboratories split the CANDLE project to support the three pilot projects of the ongoing Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) research program. The first pilot project focused on understanding patients' responses to drug treatments with the aim of developing predictive clinical drug response models to drive precision medicine. The second pilot project sought to understand a protein interaction commonly present in cancers by studying the RAS (rat sarcoma) protein pathway. RAS is the basis for many cancer studies. The final pilot project focused on better understanding treatment outcomes for patients to automate the manual process of extracting and comparing millions



Livermore development team for the Cancer Distributed Learning Environment (CANDLE): (from left) Tal Ben-Nun, Nikoli Dryden, Brian Van Essen, Pier Fiedorowicz, Fred Streitz, and Adam Moody.

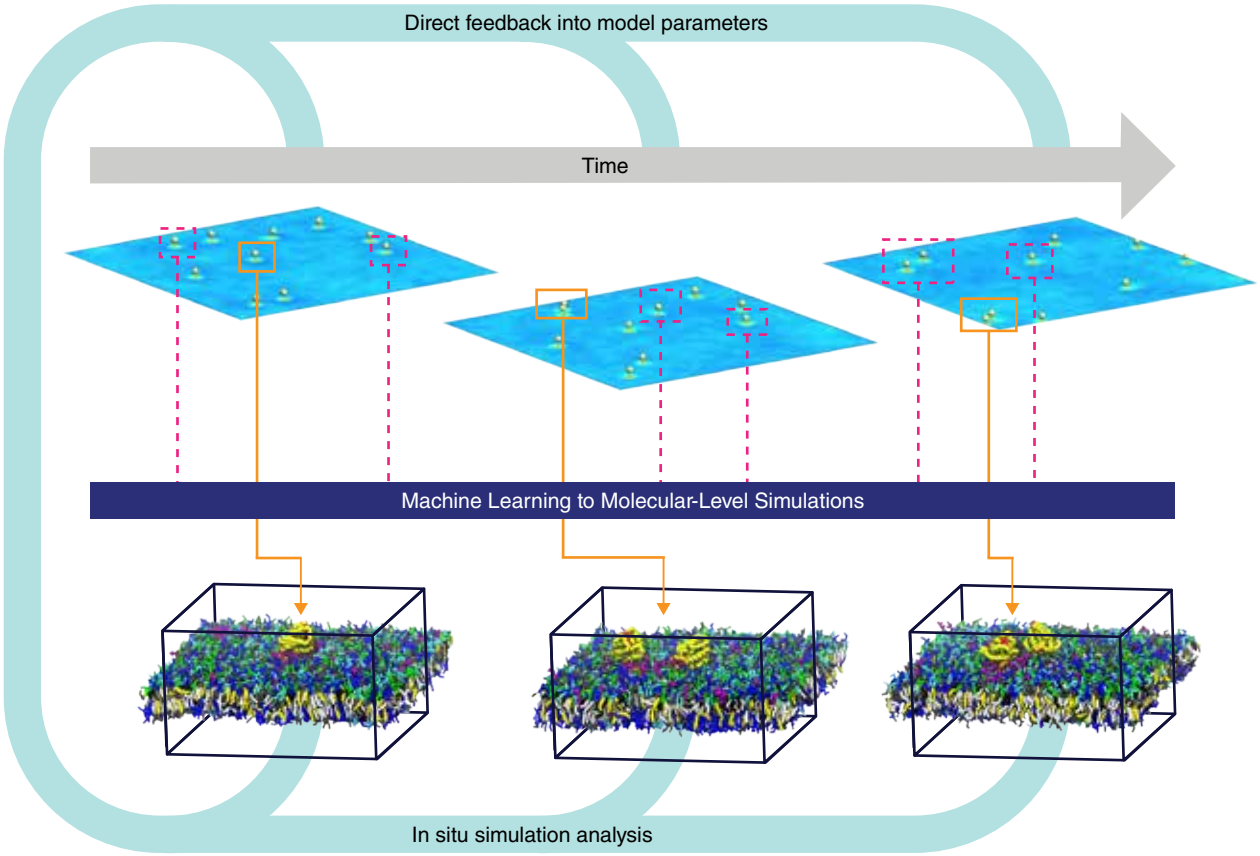
of patient records to find meaningful patterns and optimize treatment strategies.

Livermore supported the second objective to determine how to use AI for molecular dynamics applied to a complex biological system. Van Essen's team focused on developing AI models that identified transitions in the behavior of the RAS–RAF (rapidly accelerated fibrosarcoma) protein complex as it engages with a lipid membrane. Dysregulation of this interaction pathway occurs in 30 percent of cancers and regulates the proliferation and survival of cells.

Understanding transitions of the RAS–RAF interaction requires processing immense volumes of data. The Lawrence Livermore CANDLE team determined a way to scale the training of AI models' cancer research to enable ML capabilities for complex, biological molecular dynamics. They developed an unsupervised learning model that analyzes molecular dynamics runs to identify when transitions were occurring to better focus the computational efforts of a multiscale molecular dynamics simulation workflow.

The CANDLE team's research efforts helped develop new advanced computing techniques that can be applied to scientific questions not limited to precision medicine. These developments include improvements in ML, large novel algorithms, large-scale AI methods—all important components in CANDLE's molecular dynamics simulations—and parallelism among multiple computing nodes to expedite these complex calculations. As a result, CANDLE has introduced groundbreaking ideas into scientific computing, not only in terms of drug prediction and toolkit innovation, but particularly in the use of AI in model prediction tasks. "What the CANDLE project spearheaded has contributed to the AI community, and hints, shadows, and flavors of this can be seen in new applications of scientific machine learning," says Van Essen.

Key challenges in cancer drug research include managing and uncovering trends across multiple large, uncurated data sets. The advancements in ML and large-scale AI enabled by CANDLE have achieved deep-learning benchmarks that allow researchers to select and analyze specific parameters within these data sets. In



CANDLE machine-learning models identify novel transitions of the RAS (rat sarcoma)–RAF (rapidly accelerated fibrosarcoma) protein complex. Yellow dots outlined by solid orange lines represent areas of interest for further analysis at the molecular level (bottom boxes). As data is fed back into the model and a new protein–lipid patch is analyzed, over time the model yields a workflow that precisely investigates features of the RAF–RAS protein complex.

this way, the technology developed by CANDLE has accelerated cancer drug research and the process of studying molecular interactions to better understand patient treatment outcomes.

In addition to these cutting-edge cancer research solutions, CANDLE has demonstrated potential application beyond oncology. NIH and NCI have applied CANDLE's techniques for rapid, novel COVID-19 drug discovery, and the DOE and NCI's joint Accelerating Therapeutics for Opportunities in Medicine venture have applied CANDLE to develop effective drug therapies more efficiently. CANDLE has also helped provide benchmark tests for DOE supercomputers, including Lawrence Livermore's upcoming El Capitan supercomputer.

Following Lawrence Livermore's 2023 R&D 100 successes, the Laboratory will undoubtedly continue to advance HPC, make computing applications easier and more reliable to use, and stretch the boundaries of what AI, ML, and other computing technologies can do in applied research. As El Capitan comes online and Livermore's HPC capabilities increase, HPC will play a correspondingly important role in scientific research and advancements across national security, climate change, biosecurity, advanced manufacturing, and materials engineering programs. The Laboratory's habit of innovation positions Livermore to continue to lead and advance the field of computing for years to come. "Behind Livermore's innovations in computing is an extraordinarily talented and dedicated workforce," says Jeffrey Hittinger, director of CASC. "Their expertise and ingenuity, driven by the immediate challenges of our mission, sustains our leadership in HPC."

—Anashe Bandari

For further information contact Bruce Hendrickson (925) 422-8673 (hendrickson6@llnl.gov).