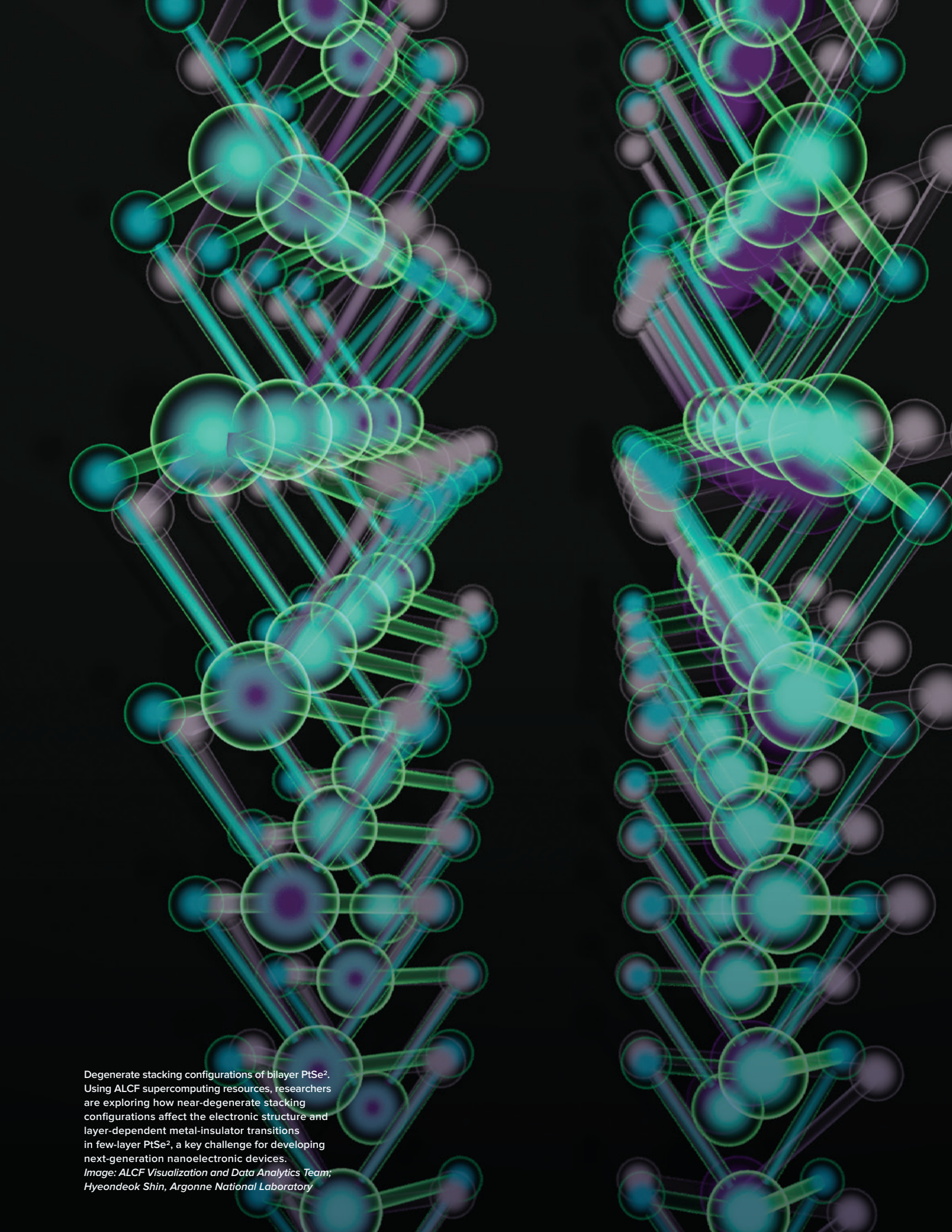
Argonne NATIONAL LABORATORY

Argonne Leadership Computing Facility

Science

2025



Degenerate stacking configurations of bilayer $PtSe_2$. Using ALCF supercomputing resources, researchers are exploring how near-degenerate stacking configurations affect the electronic structure and layer-dependent metal-insulator transitions in few-layer $PtSe_2$, a key challenge for developing next-generation nanoelectronic devices.
Image: ALCF Visualization and Data Analytics Team; Hyeondeok Shin, Argonne National Laboratory

CONTENTS

MICHAEL E. PAPKA
ALCF Director

BILL ALLCOCK
RICK WAGNER
ALCF Director of Operations

SUSAN COGHLAN
ALCF-3 Project Director

KEVIN HARMS
ALCF-4 Technical Director

KALYAN KUMARAN
ALCF Director of Technology

JINI RAMPRAKASH
ALCF Deputy Director
ALCF-4 Project Director

KATHERINE RILEY
ALCF Director of Science

LETTER FROM ALCF LEADERSHIP

The Argonne Leadership Computing Facility (ALCF) offers powerful computing resources to support open scientific research. Over its history, ALCF has developed and deployed some of the most advanced computing platforms, networking infrastructure, software tools, and data analysis capabilities the world has ever seen.

Recent achievements include Aurora, one of the nation's first exascale systems; Polaris, a powerful precursor machine for complex workloads; and a cutting-edge AI Testbed. The facility also offers Sophia, an AI-accelerator system for AI model training and inference, plus visualization clusters, advanced storage systems, and high-speed networking. Together, these resources support traditional modeling, advanced data science, and evolving AI workloads, extending ALCF's research impact into the exascale era.

Argonne's Nexus effort is supporting the Department of Energy's Integrated Research Infrastructure (IRI) initiative, which aims to link DOE experimental research facilities, data, and computing resources to accelerate discovery. ALCF staff and resources are helping lead this ambitious effort.

The year 2025 concludes two decades of remarkable innovation and signals a new era of breakthroughs. ALCF works closely with the DOE to proactively plan and execute systems that evolve with user needs, leveraging software development, strategic planning, engineering, and community training. Our team advances the facility by co-designing systems, monitoring industry trends, building pipelines for future services, and collaborating on major scientific projects, ensuring relevance to our user community from the terascale to the exascale eras.

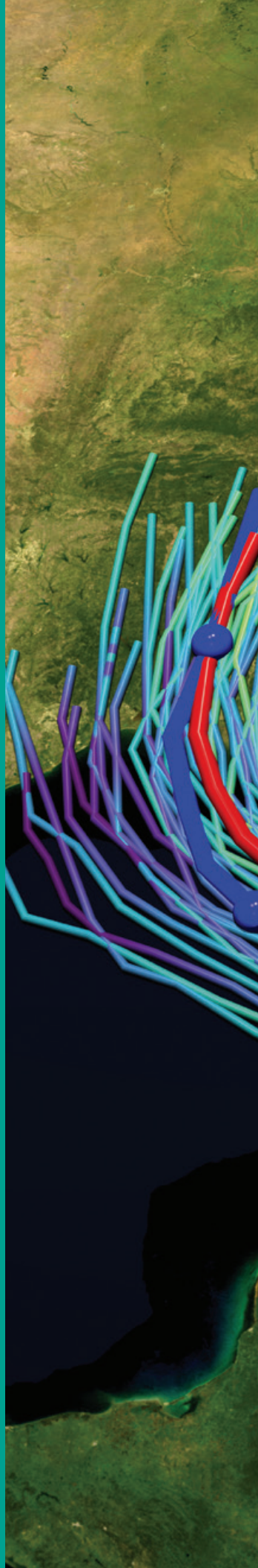
Aurora launched in 2025 as a fully operational exascale computing platform, advancing discovery in quantum materials, nuclear technologies, high-energy physics, and AI-driven science supporting national priorities. The advanced accelerators in ALCF's AI Testbed further expanded the reach of AI and machine learning, boosting scientific workflows across more fields. ALCF also updated its training series for next-gen AI practitioners, adding advanced topics to encourage wider participation and greater impact, and introduced a new training series on service-enabled science.

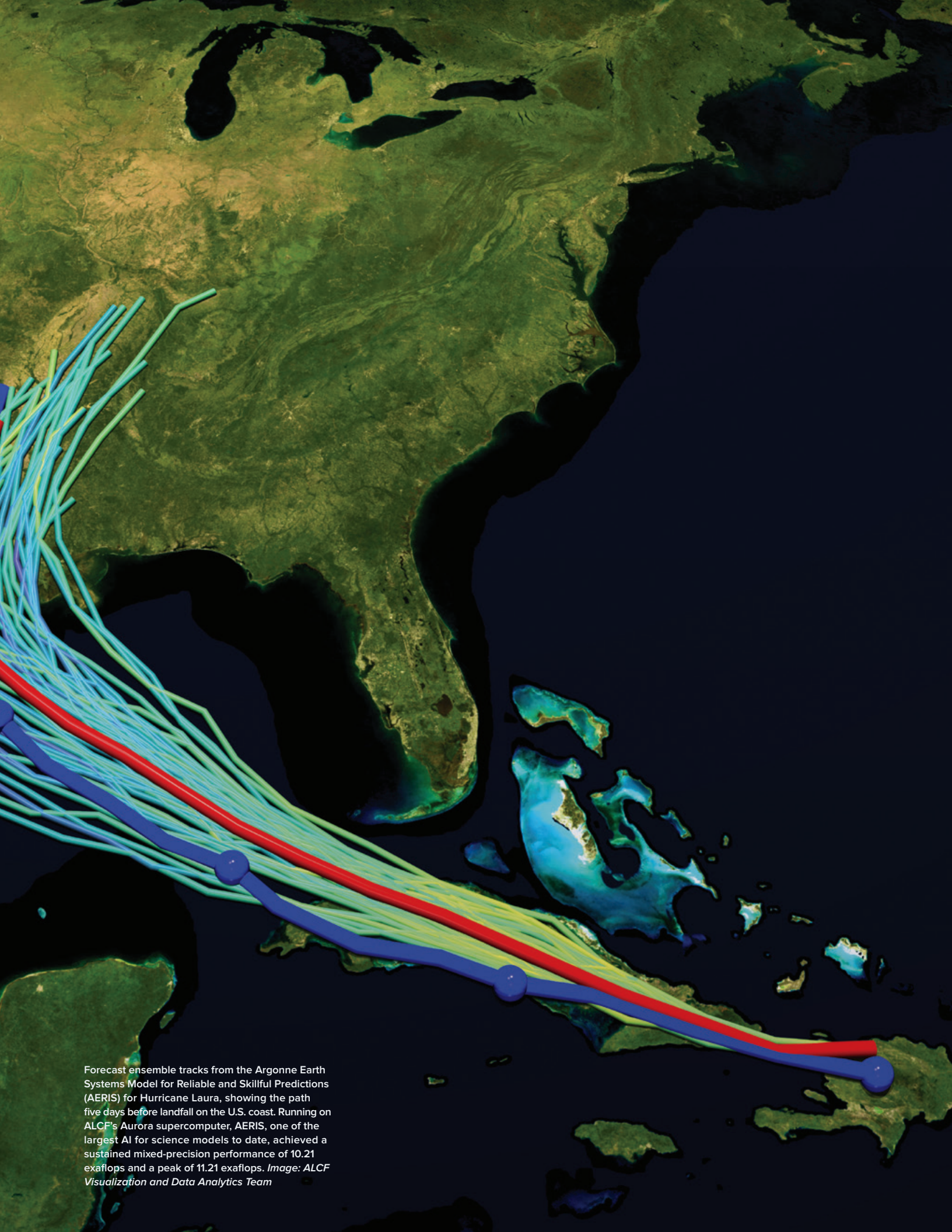
The ALCF Science Report showcases major research successes, illustrating the impact of user research in areas like materials science, biology, chemistry, and computer science. These pages highlight progress in cancer drug discovery, battery innovation, hypersonic flight engineering, and more, enabled by ALCF supercomputing and AI resources.

Through leadership in high-performance computing, ALCF empowers discoveries that benefit the nation and the world. We thank you for your continued support over the past two decades and into the future.

ARGONNE LEADERSHIP COMPUTING FACILITY

The ALCF provides world-class computing resources and expertise, enabling the research community to achieve breakthroughs in science and engineering.





Forecast ensemble tracks from the Argonne Earth Systems Model for Reliable and Skillful Predictions (AERIS) for Hurricane Laura, showing the path five days before landfall on the U.S. coast. Running on ALCF's Aurora supercomputer, AERIS, one of the largest AI for science models to date, achieved a sustained mixed-precision performance of 10.21 exaflops and a peak of 11.21 exaflops. *Image: ALCF Visualization and Data Analytics Team*

About ALCF

The ALCF is a national scientific user facility located at Argonne National Laboratory.

The Argonne Leadership Computing Facility (ALCF), a U.S. Department of Energy (DOE) Office of Science user facility at Argonne National Laboratory, enables breakthroughs in science and engineering by providing supercomputing and AI resources to the research community.

ALCF computing systems, services, and expertise—available to researchers from academia, industry, and government agencies—support large-scale computing projects aimed at solving some of the world's most complex and challenging scientific problems. Through awards of computing time, the ALCF enables researchers to accelerate the pace of discovery and innovation across a broad range of disciplines.

As a key player in the nation's efforts to provide the most advanced computing resources for science, the ALCF is helping to chart new directions in scientific computing through a convergence of simulation, data science, and AI methods and capabilities.

Supported by the DOE's Advanced Scientific Computing Research (ASCR) program, the ALCF and its partner organization, the Oak Ridge Leadership Computing Facility, operate leadership-class supercomputing resources that are orders of magnitude more powerful than the systems typically used for open scientific research.

ALCF Team

The ALCF team provides HPC expertise and support to enable the research community to achieve breakthroughs in science and engineering.

Operations
The ALCF's HPC systems administrators manage and support all ALCF computing systems, ensuring users have stable, secure, and highly available resources to achieve their scientific goals. This includes the ALCF's production supercomputers, AI accelerators, supporting system environments, storage systems, and network infrastructure. The team's software developers create tools to support the ALCF computing environment, including software for user account and project management, job failure analysis, and job scheduling.

User support specialists provide technical assitance to ALCF users and manage the workflows for user accounts and projects. In the business intelligence space, staff data architects assimilate and verify ALCF data to ensure accurate reporting of facility information.

Science
Computational scientists with multidisciplinary domain expertise work directly with ALCF users to maximize and accelerate their research efforts. In addition, the ALCF team brings extensive experience in data science, machine learning, data visualization and analysis, and mathematics to help application teams leverage ALCF resources to pursue data-driven discoveries.

With a deep knowledge of the ALCF computing environment and expertise with a wide range of numerical methods, programming models, and computational approaches, staff experts help researchers optimize the performance and productivity of simulation, data, and learning applications on ALCF systems.

Technology
ALCF computer scientists, performance engineers, and software engineers develop and optimize new tools and capabilities to enable innovative science on the facility's current and future computing resources. This includes the deployment of scalable machine learning frameworks, in situ visualization and analysis capabilities, data management services, workflow packages, and container technologies. In addition, ALCF staff are actively involved in programming language standardization efforts and contribute to cross-platform libraries to further enable the portability of HPC applications.

The ALCF team plays a key role in designing and validating the facility's next-generation supercomputers. By collaborating with compute vendors and the performance tools community, staff members ensure the requisite programming models, tools, debuggers, and libraries are available on ALCF platforms. The team also helps manage Argonne's Joint Laboratory for System Evaluation, which houses next-generation testbeds that enable researchers to explore and prepare for emerging computing technologies.

Outreach
ALCF staff members organize and participate in training events that prepare researchers for efficient use of leadership computing systems. They also contribute to a wide variety of educational activities aimed at cultivating a skilled HPC community and workforce in the future. In addition, staff outreach efforts include facilitating partnerships with industry and academia, and communicating the impactful research enabled by ALCF resources to external audiences.

Computing Resources

SUPERCOMPUTING

ALCF supercomputing resources support large-scale, computationally intensive projects aimed at solving some of the world's most complex and challenging scientific problems.

SYSTEM NAME	AURORA	POLARIS
Purpose	Science Campaigns	Science Campaigns
Architecture	HPE Cray EX	HPE Apollo 6500 Gen10+
Peak Performance	2 EF	34 PF 44 PF of Tensor Core
Processors per Node	2 Intel Xeon CPU Max Series	1 3rd Gen AMD EPYC (Milan)
GPU per Node	6 Intel Data Center GPU Max Series	4 NVIDIA A100 Tensor Core
Nodes	10,624	560 CPUs: 21,248 GPUs: 63,744
Cores	9,264,128	17,920
Memory	20.4 PB	280 TB (DDR4) 87.5 TB (HBM)
Interconnect	HPE Slingshot 11 with Dragonfly Configuration	HPE Slingshot 11 with Dragonfly Configuration
Racks	166	40

ALCF AI TESTBED

The ALCF AI Testbed provides an infrastructure of next-generation AI-accelerator machines for research campaigns at the intersection of AI and science. AI testbeds include:

SYSTEM NAME	CEREBRAS CS-2	CEREBRAS CS-3	SAMBANOVA CARDINAL SN30
System Size	2 Nodes (Each with a Wafer-Scale Engine) Including MemoryX and SwarmX	4 Nodes (Each with a Wafer-Scale Engine) Including MemoryX and SwarmX	64 Accelerators (8 Nodes and 8 Accelerators per Node)
Compute Units per Accelerator	850,000 Cores	900,000 Cores	1,280 Programmable Compute Units
Single Accelerator Performance (TFlops)	>5,780 (FP16)	125,000 (FP16)	>660 (BF16)
Software Stack Support	Cerebras SDK, TensorFlow, PyTorch	Cerebras Model Zoo, PyTorch	SambaFlow, PyTorch
Interconnect	Ethernet-based	Ethernet-based	Ethernet-based

SOPHIA	CRUX	MINERVA
Science Campaigns	Science Campaigns	AI Training & Inference
NVIDIA DGX A100	HPE Cray Ex	NVIDIA DGX B200
3.9 PF (FP64)	1.18 PF	(per node) 72 PF (FP8) 144 PF (FP4)
2 AMD EPYC 7742 (Rome)	2 AMD EPYC 7742 (Rome)	2 Intel Xeon Platinum
8 NVIDIA A100 Tensor Core	—	8 NVIDIA B200 Tensor Core
24	256	8
3,072	16,384	1,024
26 TB (DDR4) 8.32 TB (GPU)	64 TB (DDR4)	16 TB (DDR5) 11.5 TB (HBM)
NVIDIA HDR with InfiniBand	HPE Slingshot 11	InfiniBand
7	1	5

SAMBANOVA METIS SN40L	GROQRACK	GRAPHCORE BOW POD-64
32 Accelerators (16 Nodes and 2 Accelerators per Node)	72 Accelerators (9 Nodes and 8 Accelerators per Node)	64 Accelerators (4 Nodes and 16 Accelerators per Node)
1,040	5,120 Vector ALUs	1,472 Independent Processing Units
637.5 (BF16)	>188 (FP16) >750 (INT8)	>250 (FP16)
SambaStudio, SambaStack	GroqWare SDK, ONNX	PopART, TensorFlow, PyTorch, ONNX
Ethernet-based	RealScale	IPU Link

DATA STORAGE SYSTEMS

ALCF disk storage systems provide intermediate-term storage for users to access, analyze, and share computational and experimental data. Tape storage is used to archive data from completed projects.

SYSTEM NAME	Aurora DAOS (Preproduction)	EAGLE	GRAND	SWIFT	TAPE STORAGE
File System	—	Lustre	Lustre	Lustre	—
Storage System	HPE Distributed Asynchronous Object Storage	HPE ClusterStor E1000	HPE ClusterStor E1000	All NVMe Flash Storage Array	LTO6 and LTO8 Tape Technology
Usable Capacity	220 PB	100 PB	100 PB	123 TB	300 PB
Sustained Data Transfer Rate	25 TB/s (not validated)	650 GB/s	650 GB/s	48 GB/s	—
Disk Drives	16,384 SSD	8,480	8,480	24	—

NETWORKING

Networking is the fabric that ties all of the ALCF's computing systems together. InfiniBand enables communication between system I/O nodes and the ALCF's various storage systems. The production HPC SAN is built upon NVIDIA Mellanox High Data Rate (HDR) InfiniBand hardware. Two 800-port core switches provide the backbone links between 80 edge switches, yielding 1600 total available host ports, each at 200 Gbps, in a non-blocking fat-tree topology. The full bisection bandwidth of this fabric is 320 Tbps. The HPC SAN is maintained by the NVIDIA Mellanox Unified Fabric Manager (UFM), providing adaptive routing to avoid congestion, as well as the NVIDIA Mellanox Self-Healing Interconnect Enhancement for Intelligent Datacenters (SHIELD) resiliency system for link fault detection and recovery.

When external communications are required, Ethernet is the interconnect of choice. Remote user access, systems maintenance and management, and high-performance data transfers are all enabled by the local area network (LAN) and wide area network (WAN) Ethernet infrastructure. This connectivity is built upon a combination of Extreme Networks SLX and MLXe routers and NVIDIA Mellanox Ethernet switches.

ALCF systems connect to other research institutions over multiple 100 Gbps connections that link to many high-performance research networks, including regional networks like the Metropolitan Research and Education Network (MREN), as well as national and international networks like the Energy Sciences Network (ESnet) and Internet2.

JOINT LABORATORY FOR SYSTEM EVALUATION

Argonne's Joint Laboratory for System Evaluation (JLSE) provides access to leading-edge testbeds for research aimed at evaluating future extreme-scale computing systems, technologies, and capabilities. Here is a partial listing of the novel technology that makes up the JLSE.

Arm Ecosystem: Apollo 80 Fujitsu A64FX Arm system, NVIDIA Ampere Arm and A100 test kits, and an HPE Comanche with Marvell ARM64 CPU platform provide an ecosystem for porting applications and measuring performance on next-generation systems

Edge Testbed: NVIDIA Jetson Xavier and Jetson Nano platforms provide a resource for testing and developing edge computing applications

NVIDIA GPUs: Clusters of NVIDIA GH200, H100, V100, A100, and A40 GPUs for preparing applications for heterogeneous computing architectures

AMD GPUs: Clusters of MD MI300A, MI300x, MI250, MI50 and MI100 GPUs for preparing applications for heterogeneous computing architectures

Intel GPUs: Intel Data Center GPU Max 1550 (PVC)

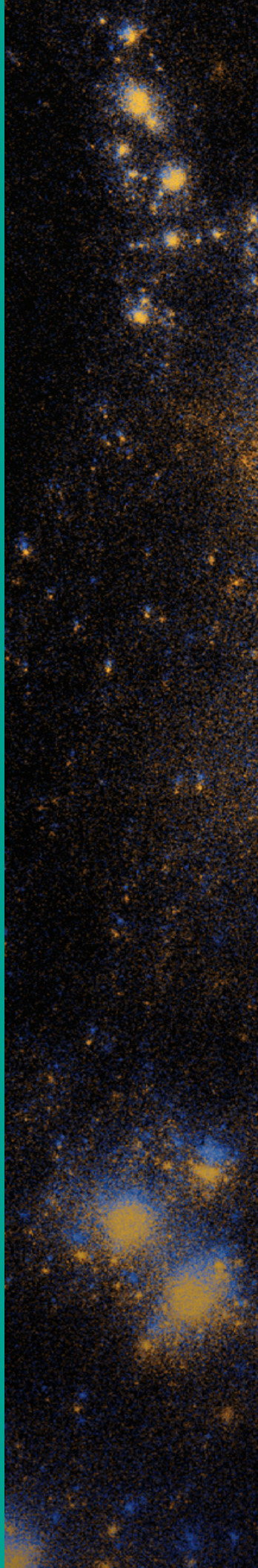
NVIDIA Bluefield-2 DPU SmartNICs: Platform used for confidential computing, MPICH offloading, and APS data transfer acceleration

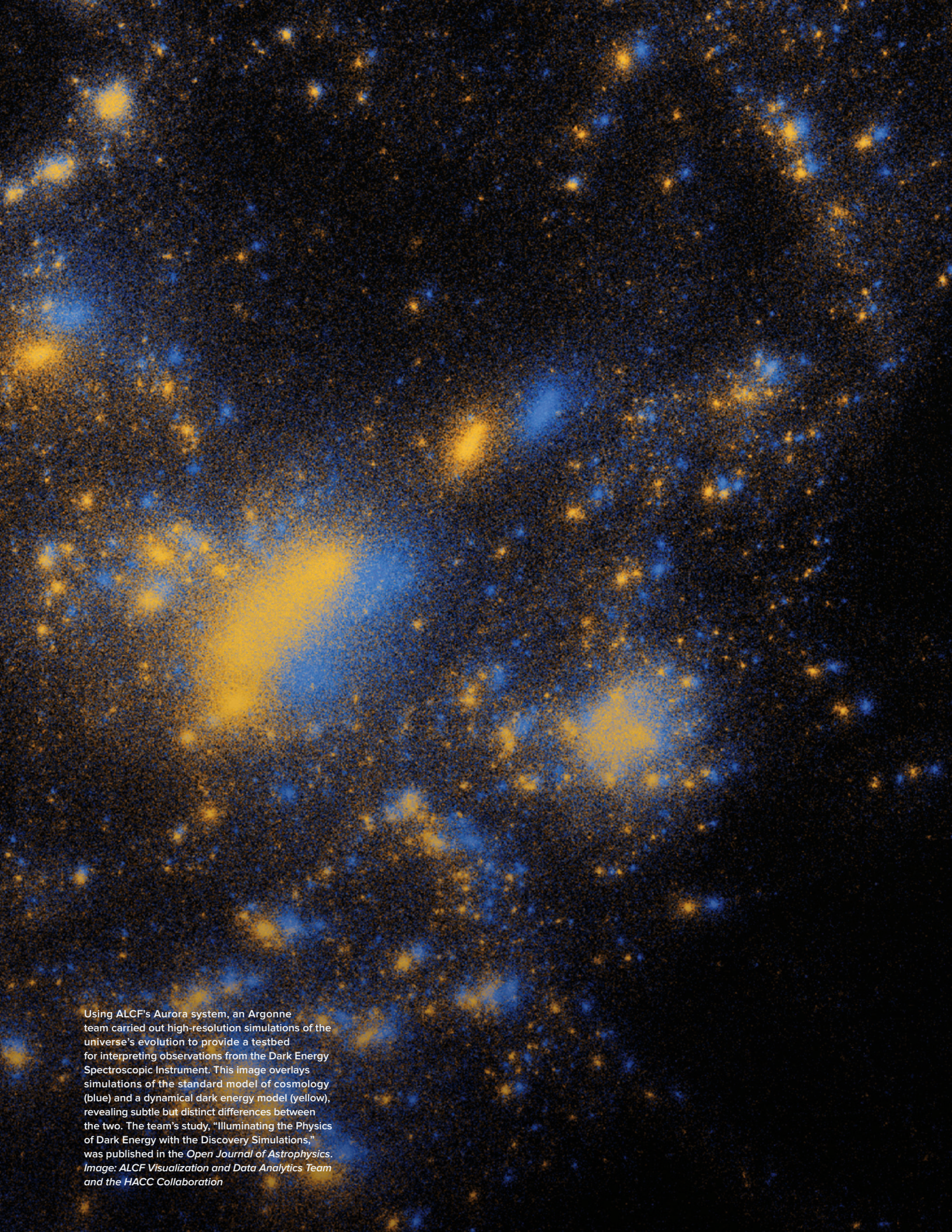
NextSilicon Maverick: First-generation product being tested by Argonne researchers

Atos Quantum Learning Machine: Platform for testing and developing quantum algorithms and applications

ADVANCING HPC AND AI FOR SCIENCE

With world-class computing resources and expertise, the ALCF is advancing exascale science, AI-driven research, and the integration of supercomputers with data-intensive experiments.





Using ALCF's Aurora system, an Argonne team carried out high-resolution simulations of the universe's evolution to provide a testbed for interpreting observations from the Dark Energy Spectroscopic Instrument. This image overlays simulations of the standard model of cosmology (blue) and a dynamical dark energy model (yellow), revealing subtle but distinct differences between the two. The team's study, "Illuminating the Physics of Dark Energy with the Discovery Simulations," was published in the *Open Journal of Astrophysics*. *Image: ALCF Visualization and Data Analytics Team and the HACC Collaboration*

Aurora Supercomputer Begins Driving Discoveries

With exascale computing power and advanced AI capabilities, Aurora is transforming how researchers tackle complex problems in science and engineering.

In January 2025, the ALCF deployed the Aurora exascale supercomputer, making it available to researchers worldwide. The milestone was marked later in the year with a dedication event attended by DOE leadership and partners from Intel and HPE, recognizing the collaborative effort behind one of the most powerful computing systems ever built.

Capable of performing more than 1 quintillion calculations per second, Aurora joins El Capitan at DOE's Lawrence Livermore National Laboratory and Frontier at DOE's Oak Ridge National Laboratory as the world's first exascale systems. Together, they occupy the top three spots on the Top500 and HPL-MxP benchmarks, which measure processing speed and AI performance, respectively.

Spanning 10,000 square feet and powered by more than 60,000 GPUs, Aurora is one of the largest and most technically advanced computing installations to date. Its deployment reflects years of collaborative development under DOE's Exascale Computing Project and ALCF's Aurora Early Science Program, ensuring that critical applications, software, and tools were ready at launch. Aurora is now fully embedded in the DOE computing ecosystem, supporting projects awarded time through the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program and the ASCR Leadership Computing Challenge (ALCC).

Aurora is being used for a wide range of applications, including high-resolution simulations of complex systems like the human circulatory system, nuclear reactors, and supernovae; large-scale data analysis from facilities like Argonne's Advanced Photon Source and CERN's Large Hadron Collider; and the training of large AI models for protein design, drug discovery, battery materials, and other research areas. Teams are also applying these AI and simulation capabilities to challenges such as fusion energy science and quantum algorithm development, integrating machine learning with physics-based methods to explore new avenues for discovery across disciplines.

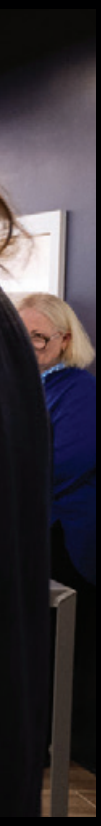
With its entry into production, Aurora marks the beginning of a new era of science powered by exascale computing, AI, and large-scale data analysis. The breakthroughs it enables are expected to advance understanding in fundamental science, accelerate technology development, and help address some of the most pressing challenges in energy, health, and security.

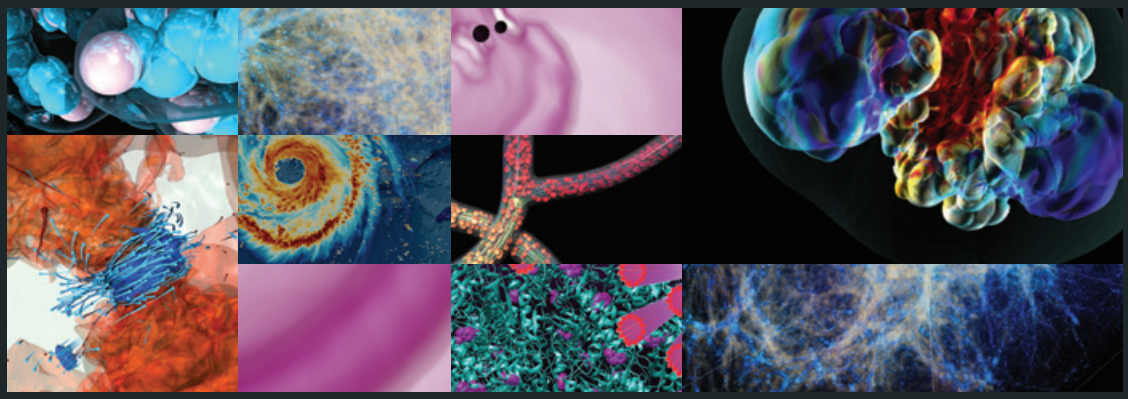
Looking ahead, the ALCF is advancing plans for its next-generation leadership system under the ALCF-4 project, which aims to deliver a major leap in performance over Aurora and further integrate HPC, AI, and data-driven approaches for future scientific discovery.

U.S. Secretary of Energy Chris Wright delivers remarks at the Aurora dedication ceremony on July 16, 2025, at Argonne National Laboratory.

Argonne's Rick Stevens and U.S. Secretary of Energy Chris Wright visit the Aurora data center at the dedication event.

Leaders from DOE, Argonne, University of Chicago, Intel, and HPE cut the ribbon at the dedication ceremony to celebrate the launch of Aurora.





With Aurora now in production, researchers are using the system to pursue breakthroughs in materials science, biology, cosmology, and beyond.

Science in Action

Aurora is already propelling research across multiple domains, helping scientists tackle complex problems with unprecedented speed and scale. In aerospace engineering, researchers are running high-resolution simulations that capture airflow around airplanes during flight. These studies examine turbulence, shock waves, and airflow through jet engine fans to better understand their effects on drag and noise, improving predictive models and accelerating the design of quieter, more efficient aircraft.

In the life sciences, Aurora is enabling the creation of large-scale AI models that address challenging questions in biology and medicine. Teams are using the system to predict virus evolution, advance cancer research, map neural connections in the brain, and generate detailed models of human physiology. By combining AI with advanced simulations, researchers are working to accelerate drug discovery, enhance preparedness for diseases, and deepen understanding of human biology.

Aurora is also supporting fusion energy research by enabling scientists to simulate the extreme conditions inside experimental reactors like ITER. Research teams are modeling plasma behavior, heat and particle transport, and the effects of impurities on reactor performance while using AI tools to predict and control energetic particle dynamics. This combination of high-fidelity simulation and AI is speeding up research that aims to bring sustained fusion power closer to reality.

In addition, Aurora's exascale capabilities are helping advance quantum computing, including validating the security and reliability of quantum-generated randomness and simulating quantum algorithms for complex molecular design problems. These efforts are supporting the development of hybrid quantum-classical methods and laying the groundwork for practical applications of quantum computing.

Taken together, these projects illustrate Aurora's versatility as a platform that combines exascale power and AI to accelerate discovery in engineering, life sciences, energy, and emerging computing technologies.

Supercomputing Meets AI

Researchers are using ALCF computing resources to harness AI in new ways, driving discoveries in medicine, energy, and engineering.

At the ALCF, researchers are combining advanced AI techniques with some of the world's most powerful supercomputers to develop innovative approaches that address scientific challenges faster and more efficiently. From molecular simulations and drug design to materials discovery and turbulence modeling, the facility's computing tools and services are changing the way complex research problems are solved.

With the arrival of the Aurora exascale system and the ALCF AI Testbed, the facility offers researchers across the nation a unique platform for developing, training, and deploying large-scale AI models. Supported by dedicated allocation programs and staff expertise, this ecosystem is enabling pioneering research campaigns in fields that push the limits of computing and scientific discovery.

A Computing Ecosystem for AI Research

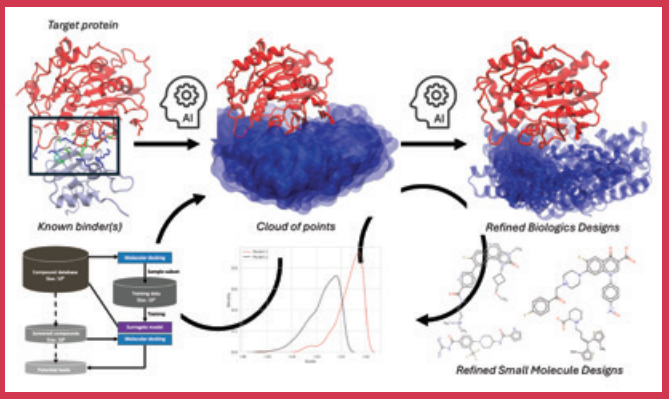
The ALCF's resources provide the scale and flexibility required to explore AI techniques that go beyond the reach of conventional computing systems. Aurora, Polaris, and Sophia all feature extensive GPU capabilities for large-scale AI and data analysis, while the ALCF AI Testbed gives researchers access to cutting-edge AI accelerators from multiple vendors. Together, these resources enable workflows that integrate model training, simulation, inference, and experimental data analysis.

Aurora's performance in leading benchmarks underscores its capacity for large-scale AI workloads. The system reached 11.6 exaflops on the HPL-MxP mixed-precision benchmark, placing it as one of the top systems globally and highlighting its readiness for the most computationally intensive AI applications. Aurora's storage systems also excel in high-throughput data handling, as shown with DAOS holding the top spot on the IO500 production list since Aurora's debut. DAOS achieved more than 10 TiB/s bandwidth in aggregate and over 100 million metadata operations per second across 642 server nodes, demonstrating strong performance for both large sequential and highly concurrent I/O workloads in production environments. Furthermore, DAOS achieved nearly 1 TB/s write and 600 GB/s read throughput on about one-third of the system in the MLPerf Storage v2.0 benchmark, supporting efficient checkpointing and data management during large-scale AI training. These results highlight Aurora's ability to support complex AI workloads reliably, giving researchers the performance needed to employ AI for science at unprecedented scale.

In addition, ALCF allocation programs, such as DOE's INCITE, have been expanded to support a growing number of AI and machine learning projects. Complemented by multiple avenues of user support, including training initiatives, hackathons, and community workshops, the ALCF is empowering scientists to effectively scale and optimize their AI workloads on the facility's powerful computing resources.

Hundreds of scientists gathered at Argonne to explore the latest AI models as part of the 1,000 Scientist AI Jam held at nine DOE national labs in March.

Driving Breakthroughs Across Scientific Domains



A schematic showing how small molecules and biologics are designed using scalable agentic workflows on Aurora. The method uses large training datasets to screen small molecules and peptides, while continuously optimizing the designs successively using diffusion models and molecular simulations.
Image: Archit Vasan and Arvind Ramanathan, Argonne National Laboratory

ALCF's HPC and AI capabilities have established the facility as a hub for AI-enabled science, supporting innovation in areas ranging from the discovery of new drugs and battery materials to the development of improved turbulence models for aircraft design. The following efforts illustrate how these technologies are revolutionizing the use of AI for scientific research.

Accelerating Cancer Drug Discovery

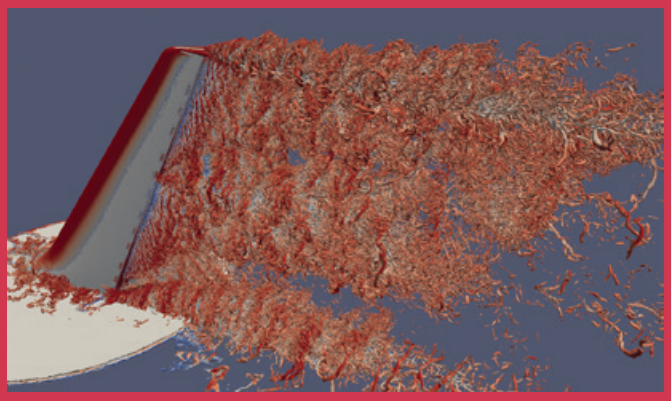
Researchers at Argonne are harnessing Aurora to transform drug discovery for cancer. By combining AI methods with large-scale molecular simulations, scientists can rapidly screen vast chemical libraries and identify compounds that may inhibit tumor growth. As part of the Aurora Early Science Program, researchers used the system to screen 50 billion small molecules in about 20 minutes, a task that was previously infeasible.

Building on this foundation, a new effort supported by the Advanced Research Projects Agency for Health is targeting proteins that are considered undruggable in cancer progression. By integrating structural biology data from the Advanced Photon Source with Aurora-powered simulations, researchers are mapping protein structures and identifying binding sites where small molecules could disrupt tumor activity. This convergence of AI, simulation, and experiment is advancing the search for therapies against some of the most challenging cancer targets.

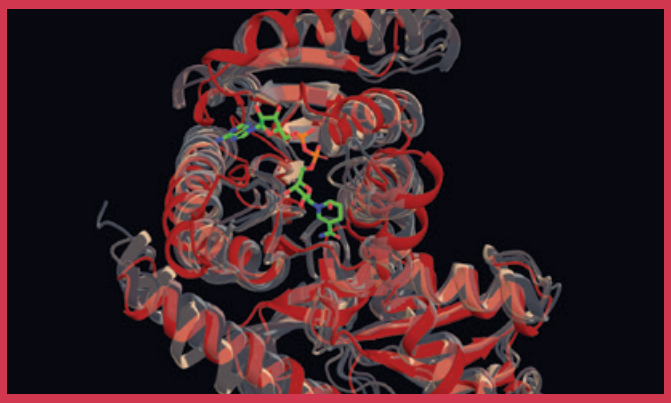
Foundation Models for Battery Materials

A University of Michigan-led team is using ALCF systems to develop AI foundation models aimed at accelerating battery materials discovery. Trained on data from billions of molecules, these models can predict key properties of battery electrolytes and electrodes with far greater speed and accuracy than conventional methods.

After using Polaris to train one of the largest molecular foundation models to date, with a focus on electrolytes, the team is now leveraging Aurora to develop a model for molecular crystals, key components of battery electrodes. Validated against experimental data, these foundation models are opening new pathways for designing safer, longer-lasting, and higher-performance batteries.



Researchers from the University of Colorado Boulder are using Aurora to combine large-scale simulations and machine learning to improve turbulence modeling for aircraft design. *Image: Kenneth Jansen, University of Colorado Boulder*



Using the AI-driven MProt-DPO framework, scientists identified synthetic versions of malate dehydrogenase that preserve the protein's critical structure and key binding areas. *Image: Arvind Ramanathan, Argonne National Laboratory*

This project also highlights how ALCF training and expertise play a key role in advancing research at the facility. Collaborations at ALCF hackathons and cross-disciplinary exchanges with teams working on AI for biology and chemistry have helped researchers scale their workloads and adapt language-modeling methods for battery materials research.

Smarter Turbulence Models for Aircraft Design

A team led by the University of Colorado Boulder is combining AI and fluid dynamics simulations on Aurora to improve turbulence modeling for aircraft design. The researchers run high-fidelity simulations of airflow around aircraft structures and use the resulting data to train machine learning models, which in turn enhance turbulence models used in lower-resolution simulations that guide real-world engineering decisions.

By integrating online machine learning with large-scale fluid dynamics simulations, the team is bypassing traditional data bottlenecks and developing predictive models that capture the complex behavior of turbulent flows. This approach has the potential to reduce the need for costly wind tunnel experiments while informing the design of more efficient, low-drag aircraft.

Breaking New Ground in AI-Driven Protein Design

An Argonne-led team developed MProt-DPO, a multimodal protein design framework that combines language models with experimental data, molecular simulations, and text-based descriptions of protein properties. Using direct preference optimization, the framework incorporates feedback from experiments and simulations to guide designs toward proteins that are stable, functional, and catalytically active. Aurora powered the large-scale training and simulation workflow, achieving 5.57 exaflops of mixed-precision performance. The team demonstrated its effectiveness by generating synthetic variants of enzymes such as HIS7 and malate dehydrogenase, with early laboratory tests confirming improved catalytic behavior. This work not only validated the framework but also provided key insights for Argonne's AuroraGPT initiative, which aims to advance autonomous discovery across disciplines.

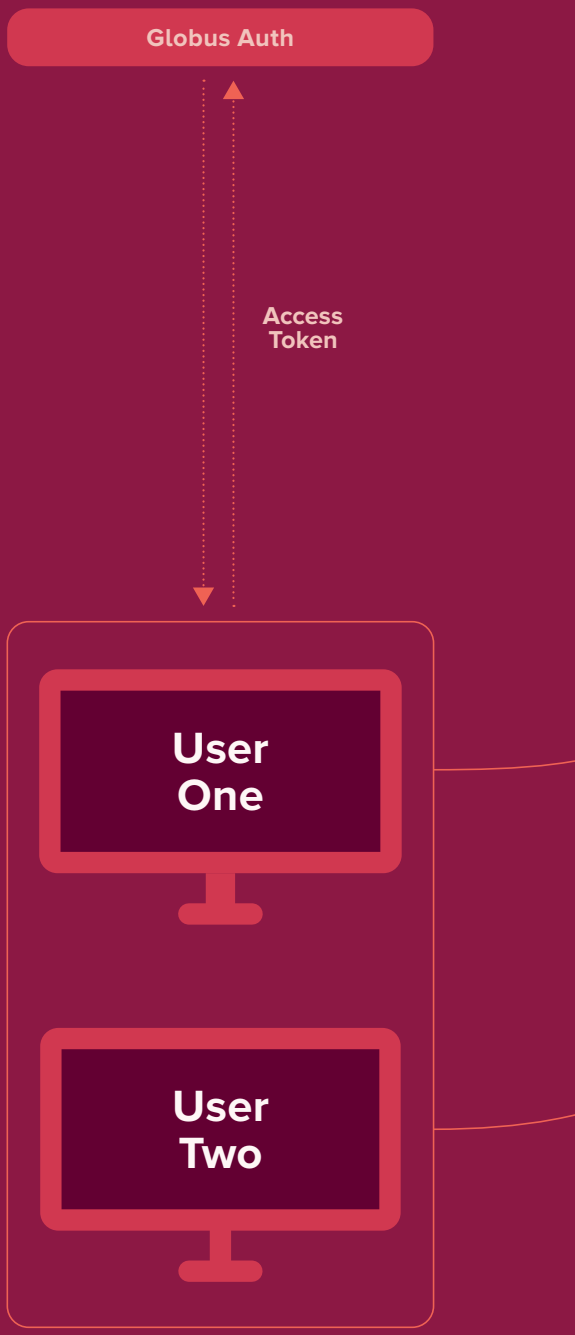
Building Toward the Future of AI for Science

These examples underscore the unique role of the ALCF in advancing AI-driven science. From exascale simulations to domain-specific foundation models, the facility is helping researchers harness AI to accelerate discovery and generate insights that were previously out of reach.

With Aurora entering production and the AI Testbed expanding, the ALCF is poised to support even larger and more ambitious research campaigns in the coming years. By combining advanced computing architectures, innovative AI methods, and deep expertise, the facility is driving progress toward a future where AI and supercomputing work in concert to solve the most complex challenges in science and technology.

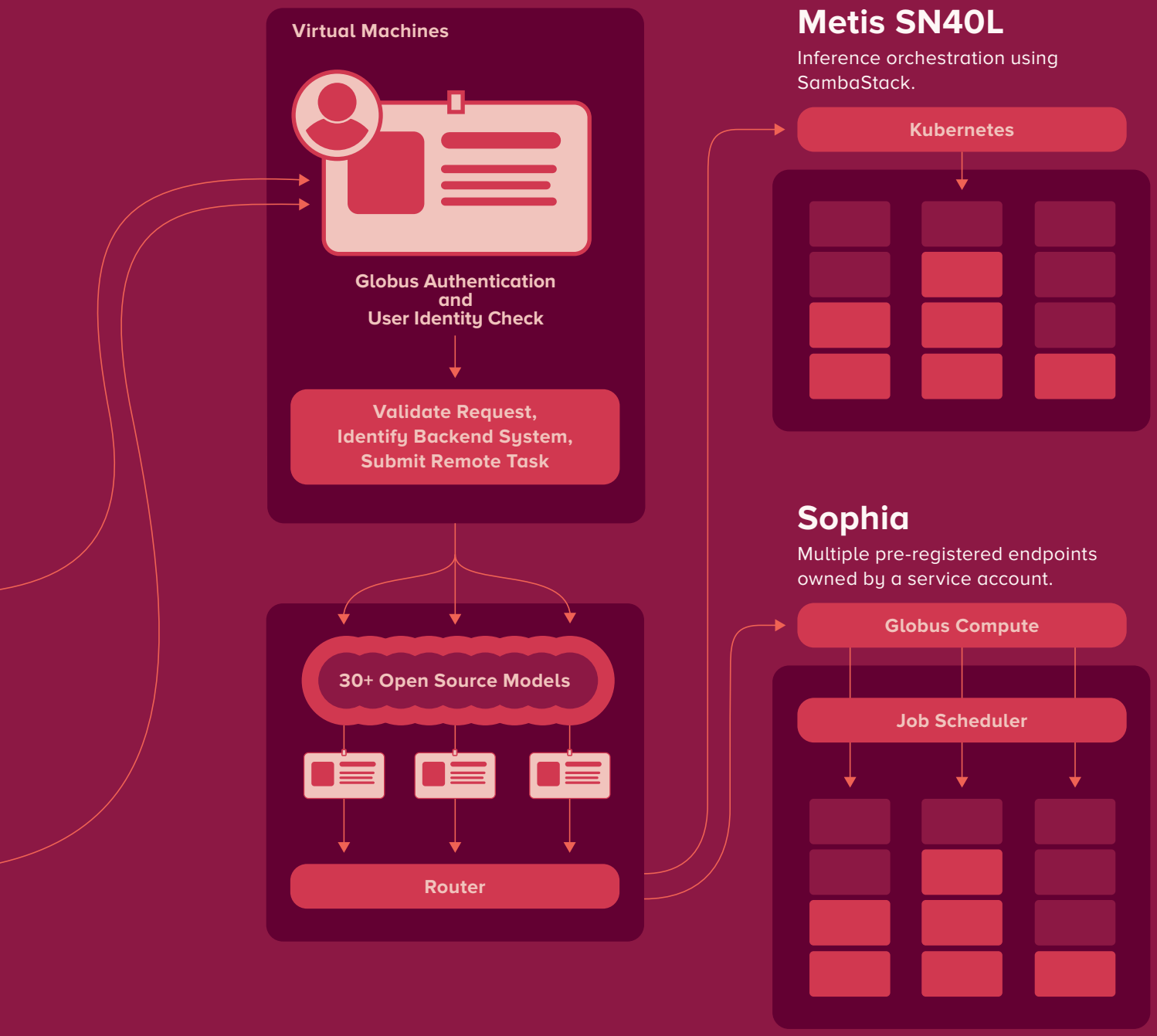
ALCF Launches New Inference-for-Science Service

The ALCF Inference Service enables researchers to run secure, scalable AI inference on Argonne's HPC systems, offering enhanced accessibility, massive scalability, privacy, and performance optimized for scientific workflows to advance data-driven discovery.



Inference Services

Provides the programmatic interfaces to 30+ open-source AI models for science.



Metis SN40L

Inference orchestration using SambaStack.

Sophia

Multiple pre-registered endpoints owned by a service account.

ALCF's Sophia system is one of the machines helping power the ALCF Inference Service.

The ALCF has deployed a first-of-its-kind inference-for-science service, providing its users with cloud-like access to large language models (LLMs), foundation models, and other AI-driven inference workloads to integrate with their existing workflows.

The ALCF Inference Service is intended to accelerate the pace of scientific discovery and breakthroughs by leveraging Argonne HPC resources to offer users access to remote, secure, scalable, and efficient AI inference for a vast and growing array of scientific and research-related applications.

The inference-as-a-service framework advances the use of AI in scientific research. With applications for everything from cosmological surveys to drug discovery to material design to improved design of nuclear reactors, LLMs and foundation models are becoming a key tool for the research community. Researchers are increasingly using these models for tasks such as analyzing vast datasets, interpreting and accelerating complex simulations, while also generating data on which to train new models.

Making Inference Accessible

There is a clear need for efficient and accessible inference services. However, their deployment and management present significant hurdles, computational and otherwise. Such challenges include the complexities of internal authentication policies, job scheduling systems, and resource management.

Tailored for scientific applications running on HPC systems—and agnostic to the underlying HPC clusters—the ALCF Inference Service aims to address these limitations, and to bridge the gap between traditional HPC infrastructure and demand for inference services. Designed in recognition of the key role inference plays in science workflows, the Inference Service combines low-latency model-serving for interactive use and high-throughput batch processing to meet the complex and varied computational needs of contemporary HPC-driven research. It provides for OpenAI API compatibility enabling easy integration with applications.

Working at Scale, Securely

The Inference Service's scaling is dynamic in its acquisition of compute resources, made possible by the system's architecture: an Inference Gateway API manages user requests to avoid overwhelming the system backend.

The system leverages Globus Compute for flexible model configurations and high scalability. Globus Compute, serving as an intermediate communication layer between the Gateway API and HPC resources, enables users to remotely trigger distributed inference tasks across target clusters while integrating with Globus Auth for secure, federated identity management that conforms to institutional policies. Furthermore, the communication layer seamlessly interfaces with job schedulers.

The ALCF AI Testbed, including SambaNova systems, provide powerful inference capabilities.

This layered architecture separates concerns and enables flexible deployment across different types of environments. These components, without user intervention, carry inference requests from users' workstations down to the compute nodes where the LLMs are running.

Myriad Scientific Applications

The ALCF's inference capabilities will help scientists speed progress in fields ranging from energy research to materials design. In fusion energy research, for example, AI inference allows scientists to analyze streams of experimental data in real time and predict plasma disruptions before they happen. By enabling rapid, automated analysis, these capabilities help researchers manage complex fusion reactions more safely and efficiently.

The same inference capabilities are also game-changing for materials science. Researchers can rapidly predict the properties of millions of potential materials, providing insights that could lead to better batteries, catalysts, and electronics. Instead of simulating the behavior of each compound, researchers will be able to use trained AI models to instantly infer how new materials might behave.

Dozens of users have already employed the ALCF Inference Service for a range of scientific applications in its initial deployment on the ALCF's Polaris and Sophia systems to process millions of inference tasks, generating many billions of tokens.

A team is working to develop a chatbot for HPC support—capable of providing user support in complex computing environments—using the ALCF Inference Service's embedding and inference models.

Additional researchers used the Inference Gateway to conduct comprehensive evaluations and comparisons of custom generative pre-trained transformer (GPT) models trained on domain-specific datasets against standard open-source models, including comparisons between AuroraGPT models and general-purpose LLMs to assess their performance in understanding and generating HPC-specific content. They also evaluated domain-specific models trained on private scientific datasets.

Benchmarking performance across different model architectures and sizes, the researchers gleaned insights into the tradeoffs between model size, computational requirements, and performance. The work entailed processing more than 50,000 inference requests across 15 different models; evaluation times reduced by 40 percent compared to manual deployment approaches, with performance remaining consistent.

The ALCF Inference Service was demonstrated to scale effectively throughout these use cases. The developers plan to expand the service.

Seamless Science

From x-ray beamlines to fusion experiments, Argonne is combining supercomputers, workflow tools, and decades of expertise to make integrated, cross-facility science a reality.

DOE's Integrated Research Infrastructure (IRI) initiative aims to accelerate the pace of scientific discovery by seamlessly connecting data-intensive experimental and observational facilities with powerful supercomputers. Argonne National Laboratory is helping to advance this DOE initiative through its Nexus effort, coordinating projects, tools, and collaborations that link large-scale experiments with ALCF computing resources. For more than a decade, Argonne has been leading advances in this space, pioneering the integration of experimental and observational facilities, leadership computing facilities, and workflow tools to create a geographically distributed research ecosystem that streamlines the path to data-driven discoveries.

At the center of the lab's efforts are the ALCF and the Advanced Photon Source (APS), where collaborations have demonstrated core building blocks for IRI, from near-real-time data analysis to new models of computer access. A key enabler of these capabilities is Globus, the data management and compute platform developed at Argonne and the University of Chicago. Globus services—Transfer, Compute, and Flows—provide the automated backbone for securely moving data, running computations, and linking processes across facilities. Combined with ALCF's world-class computing power, staff expertise, and Eagle, the facility's large-scale data-sharing platform, these resources are laying the foundation for the IRI vision.

HPC on Demand

Providing instant access to DOE supercomputers for experiment-time data analysis requires a shift in how computing facilities operate. Traditionally, researchers must create individual user accounts at each supercomputing center, then submit jobs to queues where they wait their turn. This model does not align with the rapid turnaround times needed for IRI.

To overcome these hurdles, Argonne has developed and demonstrated new approaches, including service accounts and on-demand queues, to make its computing resources more accessible. Service accounts provide secure access tied to a specific experiment rather than individual users, allowing teams to run analysis jobs seamlessly regardless of which member is operating the experiment on a given day. On-demand queues dedicate a portion of supercomputer nodes to IRI workloads, ensuring time-sensitive jobs can launch immediately rather than waiting in the standard queue.

Several facilities and projects have run remote analyses on ALCF supercomputers, with some integrating them into production workflows.

ESGF Earth System Grid Federation

JGI Joint Genome Institute

LCLS-II Linac Coherent Light Source II

ALS Advanced Light Source

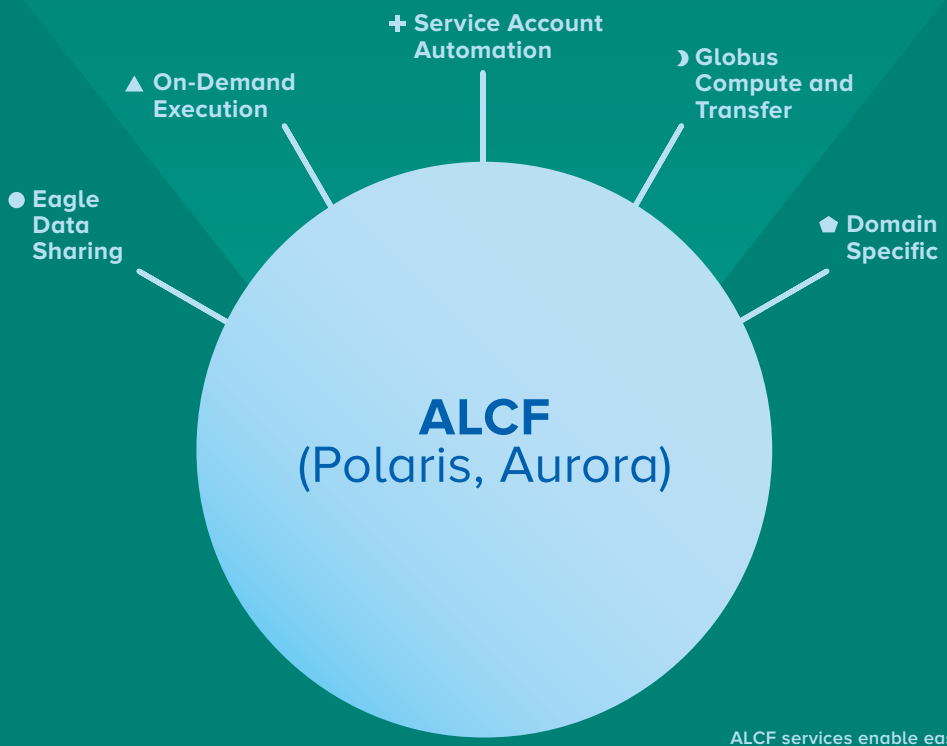
APS Advanced Photon Source

NSLS-II National Synchrotron Light Source II

DIII-D DIII-D National Fusion Facility

SBND Short-Baseline Near Detector

OpenCosmo OpenCosmo Data and Analysis Portal



ALCF services enable easy, secure, and reliable integration with experiments

- ● **Eagle Data Sharing**: Large-scale filesystem enables secure data sharing across collaborations
- ▲ **On-Demand Execution:** Dedicated queues allow time-sensitive analysis jobs to run during experiments
- \+ **Service Account Automation:** Experiment and project-specific accounts enable automated workflow execution
- ◗ **Globus Compute and Transfer:** Globus tools facilitate secure data movement and job execution
- ⬟ **Domain Specific:** ALCF supports tailored integrations for specific experimental or research needs

By integrating ALCF supercomputers with APS beamlines, scientists can analyze data as it is collected, enabling them to steer experiments in real time.

These capabilities have been successfully deployed in APS experiments, enabling fully automated runs that move data from APS beamlines to ALCF supercomputers and back with no humans in the loop. The process relies on Globus to run the computational flows between the two facilities, managing high-speed data transfers, ALCF computations, and the cataloging and distribution of results. This combination of service accounts, on-demand queues, and Globus automation is broadly applicable to IRI-related workflows across DOE facilities.

X-ray Science at APS

The upgraded APS provides up to 500 times brighter x-rays and corresponding increases in the generation of experimental data. To handle the growing deluge of data, Argonne researchers built automated pipelines that connect APS beamlines directly with ALCF's Polaris supercomputer. Using Globus Flows, data is transferred from the beamline, reconstructed on Polaris, and returned to the APS in near-real time, delivering analysis results during active experiments.

This capability was recently demonstrated with Laue microdiffraction experiments, where scientists were able to collect and analyze data 10 times faster than before. Globus services also provide resilience: if ALCF is unavailable, workflows can transfer to another facility such as NERSC, ensuring experiments continue without interruption.

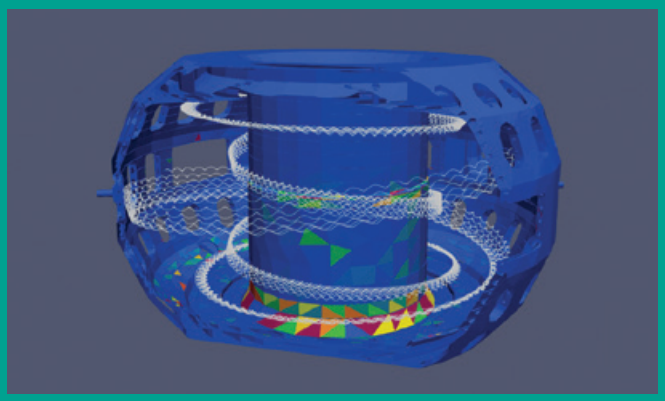
Multi-Facility Imaging with Ptychography

Ptychography, a high-resolution imaging technique, requires sophisticated analysis and large-scale computing resources. Managing data across multiple synchrotron facilities has historically been a major barrier. Argonne and partners are addressing this challenge by pairing the Ptychodus reconstruction tool with Globus Flows to create a standardized, automated workflow for ptychographic imaging.

The workflow is already being used at APS and the Linac Coherent Light Source, and pilots are underway at the National Synchrotron Light Source II and the Advanced Light Source. By using a common set of file readers and reconstruction libraries, Ptychodus ensures data can be ingested from different beamlines, while Globus orchestrates the transfer and execution of analysis tasks at remote HPC sites. This approach allows researchers to dispatch jobs and retrieve results regardless of where the experiment took place, lowering barriers to collaboration and accelerating science across the light source community.

Cosmology with the OpenCosmo Portal

The OpenCosmo Data and Analysis Portal, a collaboration between the Argonne Cosmology Group and ALCF, is designed to simplify access to massive cosmological simulation datasets by providing a browser-based entry point for analysis and visualization. Built on Globus services and linked to ALCF, OLCF, and NERSC computing systems, the portal allows users to run queries, visualize results, and launch HPC-scale analyses of data on the Eagle filesystem, without needing to manage data transfers or system-specific details.



The IonOrb code generates a heat map of deposited high energy particles onto the walls of the DIII-D tokamak (blue). This information is crucial for the protection of sensitive equipment and the prevention of excessive impurity generation. *Image: DIII-D*

Globus Flows manages the movement of data across sites, while Globus Compute executes tasks where the data resides, reducing unnecessary transfers. Security and access control are handled through Globus Groups and facility-specific policies. This approach not only streamlines large-scale cosmology research, but also broadens access by making advanced computational analyses available through simple web interfaces, helping researchers focus on science rather than infrastructure.

Fusion Energy Research at DIII-D

At the DIII-D National Fusion Facility, experiments use plasma pulses, or shots, occurring every 20 minutes to study plasma behavior, generating large volumes of data that must be analyzed quickly to inform adjustments for the next shot. In collaboration with DIII-D and NERSC, ALCF has helped implement two Globus-enabled workflows that are transforming this process.

The Consistent Automatic Kinetic Equilibria (CAKE) workflow runs plasma reconstructions in lockstep with the pulsed operation of the tokamak. By leveraging remote HPC resources, CAKE rapidly produces equilibrium analyses that would otherwise require manual intervention. Similarly, the IonOrb workflow uses GPU-accelerated codes at ALCF or NERSC to track particle trajectories, reducing run times from 4.5 hours to just 7 minutes. These results are returned to researchers in near-real time, enabling them to assess instrument risks and adjust experiments on the fly. Both workflows rely on Globus tools to automate data movement, job submission, and analysis, highlighting how this approach enables more efficient experimental science.

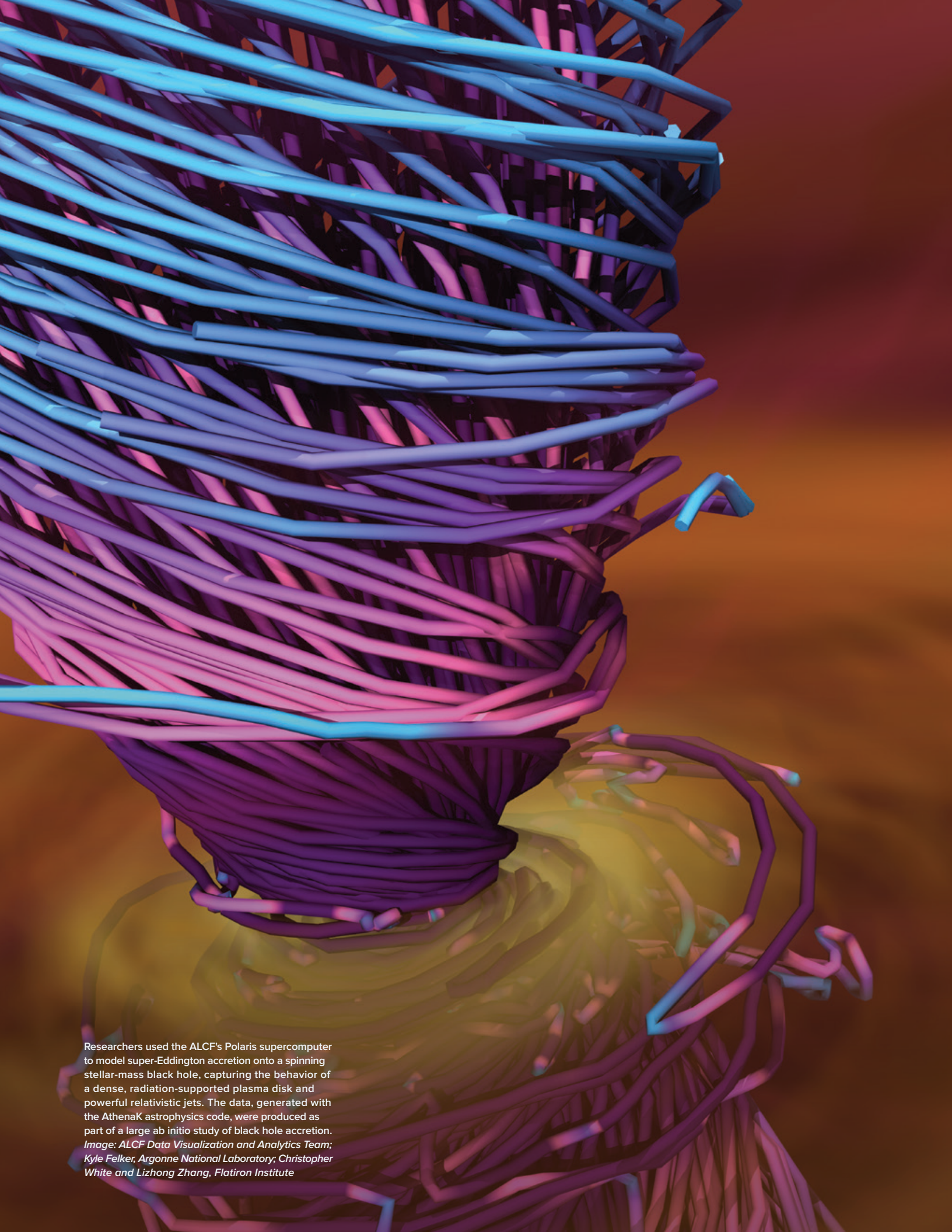
Building Toward an Integrated Future

These projects illustrate how Argonne is not only advancing data-intensive science and experimental research, but also demonstrating the broader capabilities needed for IRI: interoperable workflows, on-demand access to supercomputing resources, multi-site resilience, and user-friendly interfaces for distributed science. Through Argonne's Nexus effort, ALCF has deployed key services for near-real-time computing and data analysis, while Globus provides the secure, scalable framework for integrating experimental and observational facilities with HPC resources.

Beyond technical deployments, ALCF staff also play a leadership role in DOE's IRI program, co-chairing working groups, serving on technical subcommittees, and collaborating closely with DOE user facilities to shape future requirements. Together, these efforts ensure that the DOE research community has the tools, infrastructure, and expertise to make IRI a reality, enabling faster, more interactive, and more modes of discovery between experimental facilities and ALCF.

SCIENCE

The ALCF's AI and supercomputing resources are enabling groundbreaking research across disciplines, from biology and engineering to physics and materials science.



Researchers used the ALCF's Polaris supercomputer to model super-Eddington accretion onto a spinning stellar-mass black hole, capturing the behavior of a dense, radiation-supported plasma disk and powerful relativistic jets. The data, generated with the AthenaK astrophysics code, were produced as part of a large ab initio study of black hole accretion. *Image: ALCF Data Visualization and Analytics Team; Kyle Felker, Argonne National Laboratory; Christopher White and Lizhong Zhang, Flatiron Institute*

Accessing ALCF Resources for Science

As a national user facility dedicated to open science, any researcher in the world with a large-scale computing problem can apply for time on ALCF computing resources.

Researchers gain access to ALCF systems for computational science and engineering projects through competitive, peer-reviewed allocation programs supported by the DOE and Argonne.

Allocation Programs

INCITE

The Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program aims to accelerate scientific discoveries and technological innovations by awarding ALCF computing time and resources to large-scale, computationally intensive projects that address grand challenges in science and engineering.

ALCC

The ASCR Leadership Computing Challenge (ALCC) program allocates ALCF computing resources to projects that advance the DOE mission; help to broaden the community of researchers capable of using leadership computing resources; and serve the national interests for scientific discovery, technological innovation, and economic competitiveness.

Director's Discretionary

Director's Discretionary projects are dedicated to leadership computing preparation, INCITE and ALCC scaling, and efforts to maximize scientific application efficiency and productivity on leadership computing platforms.

ESP

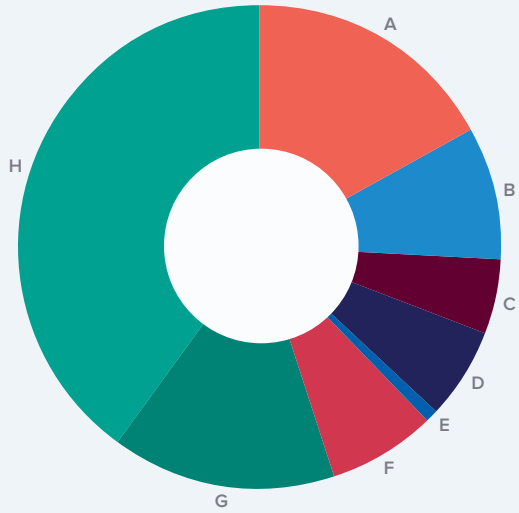
As part of the process of bringing a new supercomputer into production, the ALCF hosts its Early Science Program (ESP) to prepare applications for the architecture and scale of a new system. ESP projects represent a typical system workload at the ALCF and cover key scientific areas and numerical methods.

INCITE/ALCC BY DOMAIN

2025 INCITE

29M NODE HOURS

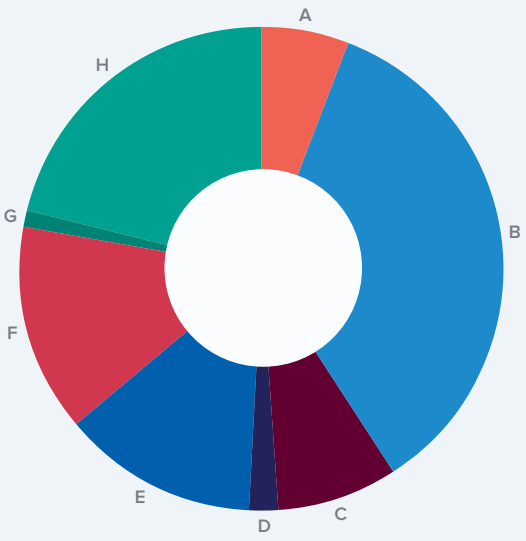
A	Biological Sciences	17 %
B	Chemistry	9
C	Computer Science	5
D	Earth Science	6
E	Energy Technologies	1
F	Engineering	7
G	Materials Science	15
H	Physics	40



2025 ALCC

14M NODE HOURS

A	Biological Sciences	6 %
B	Chemistry	35
C	Computer Science	8
D	Earth Science	2
E	Energy Technologies	13
F	Engineering	14
G	Materials Science	1
H	Physics	21



ALCC data are from calendar year 2025.

2025 Science Highlights

From simulating the universe's evolution to accelerating the discovery of new materials, ALCF users are leveraging the facility's HPC and AI resources to achieve breakthroughs across diverse research fields.

ALCF users employ simulations, AI-driven methods, and large-scale data analyses to advance scientific discovery and drive innovation in tackling some of the most pressing challenges in science and engineering.

From atomic-scale explorations to massive cosmological studies, researchers use ALCF systems to investigate complex physical systems and processes that are too small or large, costly, or dangerous to study in a laboratory.

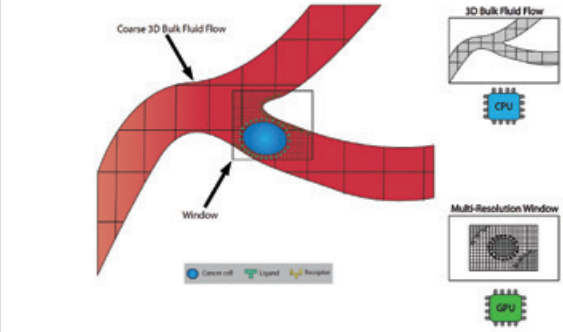
Year after year, ALCF users achieve groundbreaking results, whether by developing novel computational methods, training large-scale AI and foundation models, or publishing high-impact research across a variety of domains.

In the following pages, we highlight a selection of projects that illustrate the range and impact of recent scientific achievements enabled by the ALCF. Researchers carried out expansive cosmological simulations that are helping interpret dark energy observations, created an AI-driven Earth system model that extends forecasting capabilities, and performed high-fidelity simulations to advance our understanding of the aerothermodynamics of hypersonic flight. Other teams developed and employed innovative methods to accelerate protein dynamics simulations, advance x-ray spectroscopy, efficiently parse large datasets, implement digital twin technologies for reactor optimization, and evaluate quantum algorithms for disordered systems.

ALCF resources also enabled breakthroughs in investigating light-matter interactions, quark structures, battery electrolytes and electrodes, and ultrafast topological quantum materials. These highlights demonstrate how the facility's AI and supercomputing capabilities continue to empower researchers to develop new tools, methods, and insights across physics, chemistry, materials science, biology, and engineering.

Establishing Digital Twins for High Throughput Cellular Analysis in Whole Blood

PI Amanda Randles, Duke University
AWARD INCITE, Aurora ESP
SYSTEM Aurora



ADR-AD model overview: A cancer cell (blue) coated with ligands (green) interacts adhesively with wall receptors inside a finely resolved moving window coupled to a coarse bulk fluid domain. Bulk fluid simulations run on CPUs, while cellular calculations run on GPUs. *Image: Amanda Randles, Duke University*

Understanding how cancer spreads through the body requires tracking how circulating tumor cells (CTCs) move through the bloodstream and interact with vessel walls. These adhesive interactions occur at submicrometer scales, making them incredibly difficult and expensive to simulate in anatomically realistic settings. To address this challenge, researchers at Duke University are using ALCF computing resources to develop an adaptive, GPU-accelerated modeling framework that captures high-resolution cancer cell adhesion behavior across large vascular domains.

CHALLENGE Simulating the adhesive dynamics (AD) of cancer cells requires resolving fine-scale molecular interactions between CTC ligands and receptors on blood vessel walls, while also modeling fluid flow over millimeter- to centimeter-scale vascular geometries. Performing such high-fidelity simulations across large domains is infeasible with traditional methods due to the extreme memory and computational demands. Most existing models are limited to small, idealized vessels and cannot scale to anatomically realistic systems. A new approach is needed to bridge the gap between microscale biological accuracy and macroscale anatomical context.

APPROACH The team extended a technique known as adaptive physics refinement (APR), which dynamically couples a finely resolved simulation window to a larger coarse-grained fluid domain. In this work, they advanced the APR framework to allow the window to traverse vessel walls, enabling detailed simulation of ligand-receptor binding events at the cell surface. The model integrates lattice Boltzmann fluid dynamics with stochastic adhesive dynamics and is optimized for GPU execution. Researchers used the ALCF's Aurora supercomputer to run large-scale simulations with significantly reduced memory costs, leveraging heterogeneous CPU-GPU resources and multi-resolution spatial modeling.

RESULTS The APR-AD framework accurately reproduced adhesive interactions between a cancer cell and patterned wall receptors in a complex microfluidic geometry. Simulations showed strong agreement with fully explicit, high-resolution models while requiring far fewer resources—reducing memory usage by more than 25 times. Performance optimizations, including GPU-accelerated random number generation and octree-based receptor searches, significantly sped up the most computationally intensive portions of the workflow. Validation tests confirmed that the framework preserves flow accuracy and bond formation behavior across a range of simulation conditions.

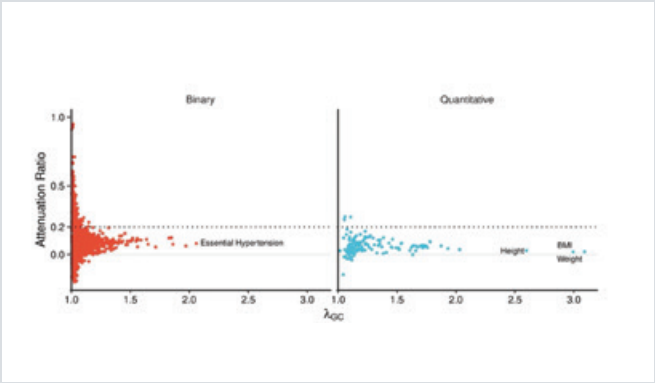
IMPACT This work provides a scalable and efficient method for simulating cancer cell adhesion at anatomical scales, enabling more detailed exploration of how CTCs interact with the vascular system during metastasis. By reducing computational costs without sacrificing biological accuracy, the APR-AD framework opens new possibilities for simulating cancer progression and testing therapeutic strategies. The approach also demonstrates how advanced multiscale modeling techniques and leadership-class computing can be combined to tackle complex problems in biomedical engineering.

PUBLICATIONS
Martin, A., W. Ladd, R. Wu, and A. Randles. "Adaptive Physics Refinement for Anatomic Adhesive Dynamics Simulations," *Computational Science – ICCS 2025* (July 2025), Springer. https://doi.org/10.1007/978-3-031-97626-1_19.

Biological Sciences | Simulation, Data, Learning

Harnessing AI and Supercomputing to Accelerate Biomedical Discovery

PI Ravi Madduri, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



Inflation statistics for within-population meta-analysis. Each dot indicates within-population meta-analysis. The horizontal axes show lambda GC values, and the vertical axis shows attenuation ratio. Dotted lines indicate the level of attenuation ratio of 0.2. *Image: Argonne National Laboratory*

Advancing biomedical science relies on analyzing vast genomic datasets, modeling disease processes at high resolution, and rapidly processing sequencing data. From statistically modeling millions of genetic variants to benchmarking data pipelines that handle terabytes of information in hours, these tasks demand immense computational power. Using ALCF resources, researchers are combining high-performance computing (HPC) and advanced AI methods to enable breakthroughs that could accelerate genetic discovery, refine disease understanding, and improve biomedical data workflows.

CHALLENGE To uncover the genetic and molecular underpinnings of disease, modern biomedical studies must integrate data from diverse sources, such as large-scale biobanks, single-cell experiments, and high-throughput sequencing. Each dataset presents unique challenges: biobank-scale analyses involve billions of variant-trait comparisons, single-cell studies require methods that capture cell-type-specific genetic effects, and sequencing pipelines must process raw data into usable results quickly and reproducibly.

APPROACH Research teams are leveraging ALCF supercomputers to address these challenges. In one large-scale analysis, investigators conducted association tests of more than 1,000 traits and identified instances of pleiotropy where a variant affects multiple traits or diseases. Another team developed scPrediXcan, a transcriptome-wide association framework that uses a deep-learning model to accurately predict single-cell gene expression and discover disease-associated genes from fine-grained biological signals. In a third effort, researchers benchmarked accelerated next-generation sequencing (NGS) analysis pipelines on HPC systems, evaluating their performance and scalability for real-world, high-throughput genomics workloads.

RESULTS The pleiotropy study produced one of the most comprehensive genome-wide assessments to date, revealing extensive shared genetic architecture among traits and over 30,000 independent genetic associations. The scPrediXcan framework improved the detection of disease-associated genes by leveraging cell-type-specific models, producing insights that would be missed in bulk-tissue analyses. The NGS benchmarking effort identified optimized pipelines that dramatically reduce time-to-results for sequencing data, enabling faster turnaround for research and clinical applications. Together, these projects demonstrate how AI-driven methods, novel algorithms, and HPC infrastructure can transform the scale and resolution of biomedical analysis.

IMPACT By applying AI and supercomputing to genetics, transcriptomics, and sequencing workflows, researchers can move from raw data to actionable insights more quickly and with greater precision. These advances lay the groundwork for more targeted disease research, faster diagnostic development, and a future of personalized treatments informed by large-scale genomic and cellular data.

PUBLICATIONS

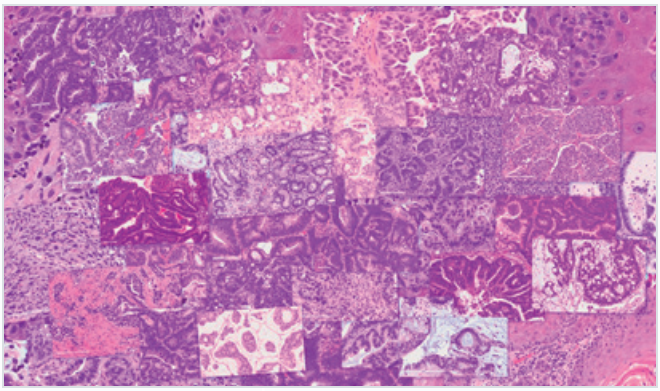
Levin, M. G., S. Koyama, J. Woerner, et al. "Genome-Wide Assessment of Pleiotropy Across >1000 Traits from Global Biobanks," *medRxiv* (preprint), openRxiv. https://doi.org/10.1101/2025.04.18.25326074

Samarakoon, P. S., G. Fournous, L. T. Hansen, A. Wijesiri, S. Zhao, R. A. Alex, T. N. Nandi, R. Madduri, A. D. Rowe, G. Thomassen, E. Hovig, and S. Razick. "Benchmarking Accelerated Next-Generation Sequencing Analysis Pipelines," *Bioinformatics Advances* (May 2025), Oxford University Press. https://doi.org/10.1093/bioadv/vbaf085

Zhou, Y., T. Adeluwa, L. Zhu, et al. "scPrediXcan Integrates Advances in Deep Learning and Single-Cell Data into a Powerful Cell-Type-Specific Transcriptome-Wide Association Study Framework," *Cell Genomics* (May 2025), Elsevier. https://doi.org/10.1016/j.xgen.2025.100875

Innovative Methodologies and New Data for Predictive Oncology Model Evaluation

PI Thomas Brettin, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Aurora, Polaris



Predicting cancer type and drug response using histopathology images from the National Cancer Institute's Patient-Derived Models Repository. *Image: Argonne National Laboratory*

Cancer continues to represent a leading cause of death globally, accounting for some 10 million deaths every year. Supercomputing resources are being used to accelerate the research and development of effective cancer treatments. Limited data and a proliferation of deep learning models without standard benchmarking, however, can restrict the extent to which state-of-the-art AI technologies can be used in such research.

To overcome these challenges, the IMPROVE (Innovative Methodologies and New Data for Predictive Oncology Model Evaluation) project, led by Argonne National Laboratory, aims to collect relevant drug-response prediction models by curating deep learning models and building a framework for their comparison, all while generating high-quality data.

CHALLENGE There are currently dozens of different drug response prediction models circulating throughout the research community. Despite their proliferation, there are no benchmarks and no standard by which to compare the models to each other. The project's aims are twofold. First, focusing on precision medicine, the researchers aim to enable oncologists to use models that recommend treatment based on a genetic, protein, or other molecular profile that can be generated from a biopsy. Second, the researchers aim to improve the predictive performance of the deep learning models.

This means being able to quickly and accurately compare thousands of models and assess which are performing best in as fair and biologically relevant a manner as possible. Moreover, the researchers want the comparison of models and the comparison of the performance impacts of training and validation choice to be as fully automated as can be achieved.

Before neurons can be reconstructed in 3D, the 2D profiles of objects must be aligned between neighboring images in an image stack. Image misalignment can occur when tissue samples are cut into thin sections, or during imaging on the electron microscope. The Feabas application (developed by collaborators at Harvard) uses template-matching and feature-matching techniques to optimize image transformations and align the 2D image content between sections.

APPROACH ALCF supercomputing resources are used to optimize hyperparameters, variables that control the models' learning process; optimization is accomplished by running large ensembles of models, and the performance of each model on specific drugs and cell lines is examined.

Once the researchers have identified the best settings for training the models, they perform cross-study generalization analyses to determine how well models trained on a dataset from the organization where they were experimentally generated perform when applied to data from a different organization.

Data include RNA and DNA sequences of cancer models, in addition to drug screening and response data continuously curated and standardized from the public domain.

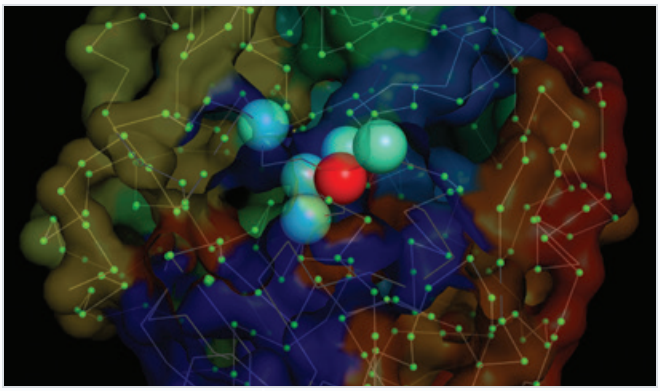
RESULTS Runs involving the training of thousands and tens of thousands of deep learning models have been carried out to assess model uncertainty as a function of specific drugs, as a function of specific cell line cancer types, and as a function of neural network architecture.

IMPACT This work stands to accelerate the research and development of effective cancer treatments.

Biological Sciences | Simulation

Multi-Resolution Genome Folding: Ensemble 3D Structures Across Diverse Tissues

PI Lin Du, Jie Liang, University of Illinois Chicago
AWARD INCITE
SYSTEM Aurora



Single-cell 3D chromatin structure, reconstructed from population Hi-C data using polymer modeling, reveals complex, many-body interactions between a gene (red) and multiple regulatory elements. *Image: Hammad Farooq and Jie Liang, University of Illinois Chicago*

3D genome organization and modifications are fundamental to cellular functions. Genomic DNAs, typically 2 million base pairs long and organized into chromosomes, are compacted within a cell nucleus 10 to 20 micrometers in diameter. Proper folding is crucial for maintaining nuclear organization and facilitating essential cellular processes such as gene expression regulation and cellular specialization.

To explore the relationship between genome 3D structure and function, a research team led by University of Illinois Chicago launched a large-scale computational campaign aimed at constructing detailed 3D models of genome folding across four distinct cell types. *The original PI of this project, Professor Jie Liang of UIC, passed away in December 2024*.

CHALLENGE Techniques such as high-throughput chromosome conformation capture (Hi-C) have provided a wealth of information on nucleus organization and genome important for understanding gene expression regulation. Genome-wide association studies (GWAS) have identified numerous loci associated with complex traits. Expression quantitative trait loci (eQTL) studies have further linked the genetic variants to alteration in expression levels of associated target genes across individuals. However, current joint analyses of Hi-C and eQTLs data lack advanced computational tools, limiting what can be learned from these data.

APPROACH The research team is generating models to study the structural basis of genome folding and its functional implications, thereby providing comprehensive maps of how genes at various loci adopt distinct spatial configurations, influence cellular states, and modulate gene expression.

RESULTS The researchers have developed a computational method for simultaneous analysis of Hi-C and eQTL data, capable of identifying a small set of non-random interactions from all Hi-C interactions. Using these non-random interactions, they reconstructed large ensembles of high-resolution single-cell 3D chromatin conformations with thorough sampling, accurately replicating Hi-C measurements. The results revealed many-body interactions in chromatin conformation at the single-cell level within eQTL loci, providing a detailed view of how 3D chromatin structures form the physical foundation for gene regulation, including how genetic variants of eQTLs affect the expression of associated eGenes. Furthermore, the method used can deconvolve chromatin heterogeneity and investigate the spatial associations of eQTLs and eGenes at subpopulation level, revealing their regulatory impacts on gene expression.

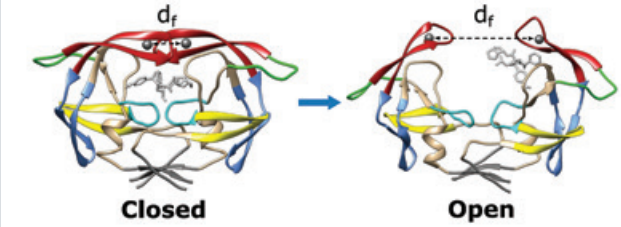
IMPACT The researchers' analysis will delineate tissue-specific master regulatory interactions and conserved interactions across cell types. It will also characterize chromatin structural heterogeneity by identifying major structural clusters in cell subpopulations. Moreover, the work will lead to a high-quality database of enhancer-gene target pairs and enable machine learning predictors to identify them across the genomes of various cell types. The approach taken is crucial for understanding genome topology, gene expression, and discovering potential causal genes for noncoding risk variants identified in GWAS.

PUBLICATIONS
Du, L., H. Farooq, P. Delafrouz, and J. Liang. "Structural Basis of Differential Gene Expression at eQTLs Loci from High-Resolution Ensemble Models of 3D Single-Cell Chromatin Conformations," *Bioinformatics* (January 2025), Oxford University Press.
https://doi.org/10.1093/bioinformatics/btaf050

Biological Sciences | Simulation

Understanding the Rigorous Molecular Mechanism of Drug Resistance to HIV Protease Inhibitors

PI Ao Ma, University of Illinois Chicago
AWARD Director's Discretionary
SYSTEM Polaris



Representative structures of the closed and open states of DRV-bound HIV-PR protease. *Image: Ao Ma, University of Illinois Chicago*

Understanding how proteins undergo conformational changes, or alterations in their three-dimensional shape, is essential for uncovering their roles in health and disease, including how they interact with drugs. However, simulating this complex process in atomic detail has long been a major challenge with traditional methods. Using ALCF supercomputers, researchers from the University of Illinois Chicago developed a new physics-based method that dramatically accelerates simulations and enhances their predictive power by identifying a protein's true reaction coordinates (tRCs), which are key drivers of conformational changes.

CHALLENGE Traditional molecular dynamics simulations struggle to capture slow conformational changes in proteins because they operate on microsecond timescales, while the relevant biological processes occur over milliseconds to hours. Enhanced sampling techniques have been developed to extend simulation capabilities, but they depend on user-defined collective variables (CVs) that are often based on intuition rather than the true driving forces of conformational change. When CVs are inaccurate, these methods either fail to accelerate transitions or produce non-physical behavior.

APPROACH With support from ALCF's Polaris supercomputer, the team developed a novel approach to computing tRCs from energy relaxation simulations, which are short and inexpensive. Building on their earlier work at ALCF, in which they identified reaction coordinates from traditional molecular dynamics trajectories, the team applied the generalized work functional (GWF) to energy relaxation simulations, enabling a faster and more efficient extraction of singular coordinates that reveal the tRCs controlling both energy relaxation and conformational activation. This method enables predictive sampling of protein dynamics from a single structure. The team used this approach on Polaris to simulate flap opening and drug dissociation in HIV-1 protease and to study allosteric regulation in the PDZ2 protein domain.

RESULTS By biasing simulations along the newly identified tRCs, the team achieved up to a 10^{15}-fold acceleration in sampling key conformational transitions, enabling processes that normally take milliseconds to hours to complete in just picoseconds. The accelerated trajectories followed realistic, physically valid transition pathways and were validated on multiple proteins, confirming the general applicability of the method. This approach also enabled the first predictive simulations of large-scale conformational changes in PDZ2, revealing a plausible mechanism for allosteric changes that have eluded intensive studies for two decades.

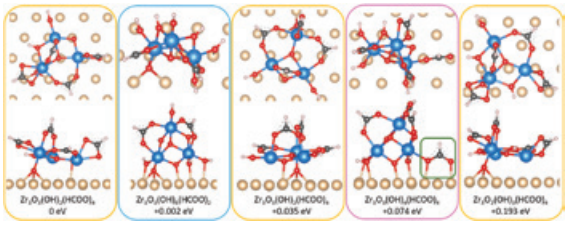
IMPACT The team's novel computational method provides a transformational tool for molecular dynamics in structural biology, overcoming a long-standing bottleneck and enabling exploration of protein functions that were previously inaccessible. By providing a fast, rigorous way to identify a protein's true reaction coordinates, the method has the potential to drive advances in biomedical research, including drug and enzyme design.

PUBLICATIONS
Li, H., and A. Ma. "Enhanced Sampling of Protein Conformational Changes via True Reaction Coordinates from Energy Relaxation," *Nature Communications* (January 2025), Springer Nature. https://doi.org/10.1038/s41467-025-55983-y

Chemistry | Simulation

Heterogeneous Catalysis as a Collective Phenomenon Within a Dynamic Ensemble of Sites

PI Anastassia Alexandrova, University of California, Los Angeles
AWARD INCITE
SYSTEM Polaris



Catalyst configurations with lowest free energy from each defined family and metastable structures that dominate the catalytic activity. *Image: Yang et al. J. Am. Chem. Soc. 2025, 147, 18, 15294–15306*

Led by researchers from the University of California, Los Angeles, this project aims to identify effective methods for converting CO_2 emissions into value-added chemicals and fuels, such as methanol, to close the carbon cycle.

CHALLENGE CO_2 activation is kinetically challenging, but recent studies have demonstrated that modifying the industrial Zn/Cu catalyst to include a ZrO layer is effective. Such catalysts are dynamic in reaction conditions, and a thorough sampling of catalyst configurations is required to grasp the full ensemble of dynamically accessed structures and the quantity and diversity of adsorbates. The configurational makeup of the catalyst strongly depends on the conditions and may vary even within the same reactor. Within a large ensemble of configurations, all species can contribute to the reaction, making evaluation of the catalytic activity difficult.

APPROACH Catalyst configurations were generated using a three-atom Zr cluster and various quantities of different adsorbates placed on top of a Cu slab. The bottom two Cu layers were held fixed during geometry optimization with the VASP code on the ALCF's Polaris supercomputer. The relative free energies of the catalyst structures were calculated from total pressure, CO_2 to H_2 ratio, temperature, and partial reactant and product pressures.

RESULTS A large variation in reaction energies was observed among the initial 13 catalyst configurations optimized. The lowest unoccupied and highest occupied molecular orbital energy levels most strongly correlate with the reaction energy. Optimal electronic properties for the Zr/Cu inverse catalyst feature a partially reduced Zr cluster. Formate groups configurations present highly distinct reactivity for their hydrogenation, with varying reaction energies and reaction barriers. Rare metastable catalyst configurations showed a markedly enhanced rate constant.

IMPACT The researchers developed an efficient computational method for examining the reactivity of many coexisting catalyst structures, offering a practical and effective methodology for studying the reactivity of inverse catalysts with structural diversity under operating conditions, and providing insights into the rational design of improved catalytic systems.

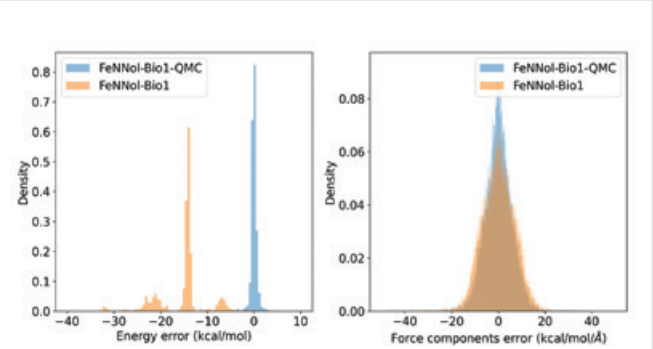
Ultimately, this work aims to a develop fundamental theory of heterogeneous thermal and electrocatalysis, and a realistic statistical and dynamical description of the catalytic interface in reaction conditions. This will enable the understanding of catalytic mechanisms, and the design of new efficient catalysts.

PUBLICATIONS
Yang, Z., S. Kumari, A. N. Alexandrova, and P. Sautet. "Catalytic Activity of an Ensemble of Sites for CO_2 Hydrogenation to Methanol on a ZrO_2-on-Cu Inverse Catalyst," *Journal of the American Chemical Society* (April 2025), ACS Publications.
https://doi.org/10.1021/jacs.5c00848

Chemistry | Simulation, Data, Learning

High-Accuracy Quantum Simulations in Cancer Therapy Using Exascale Computing

PI Anouar Benali, Qubit Pharmaceuticals
AWARD INCITE, Aurora ESP
SYSTEM Aurora, Polaris



Distribution of the energy and force components errors on the QMC validation set, before and after transfer learning. *Image: Anouar Benali, Qubit Pharmaceuticals*

Predicting the behavior of molecular systems is a key challenge in chemistry, biology, and materials science, with applications ranging from cancer drug discovery to the design of novel catalysts. Machine learning models can accelerate such predictions, but their accuracy often depends on computationally demanding quantum chemistry calculations. Using ALCF supercomputers, researchers have developed a scalable framework for training machine learning foundation models that learn from large datasets of lower-cost simulations while retaining the accuracy of high-level quantum results.

CHALLENGE High-fidelity quantum chemistry methods, such as quantum Monte Carlo (QMC), provide highly accurate molecular predictions but require significant computational resources, limiting their accessibility. Lower-cost methods, such as density functional theory (DFT), allow simulations at greater scale but with reduced precision. This tradeoff between cost and fidelity presents an obstacle for creating machine learning models that can be both accurate and broadly applicable. Bridging this gap requires new approaches that can integrate information from multiple simulation fidelities while minimizing the computational burden.

APPROACH To address these challenges, the team developed a multi-fidelity pretraining framework for molecular machine learning foundation models. They began by using ALCF's Polaris and Cineca's Leonardo supercomputers to carry out large-scale pretraining on millions of molecular structures using DFT data to capture general molecular patterns. A high-fidelity dataset of thousands of molecular configurations was then computed on Aurora using QMC forces combined with multideterminant selected-Configuration Interaction wavefunctions, providing unprecedented accuracy for fine-tuning the model. The researchers employed graph neural networks to represent molecular structures and distributed workloads across multiple GPUs to efficiently train models on large, diverse datasets.

RESULTS Aurora's exascale capabilities made it possible, for the first time, to perform full QMC force calculations and combine multideterminant QMC energies and forces at the complete basis-set limit. The team fully implemented and optimized these methods in QMCPACK, computing forces for 2,000 molecular configurations with 1.4 million determinants per molecule, achieving 52 petaflops. Models trained with this multi-fidelity approach achieved accuracy comparable to high-level quantum methods while maintaining DFT-scale efficiency. They outperformed single-fidelity models on benchmark molecular property tests and generalized effectively to new molecular systems and target properties.

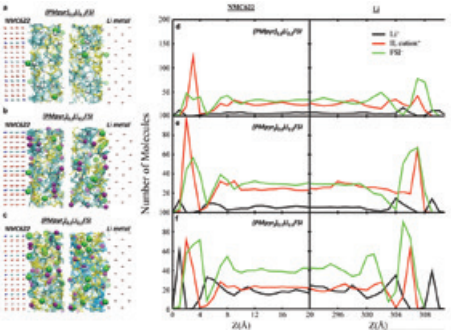
IMPACT This work demonstrates a practical path toward creating large-scale, accurate molecular foundation models without prohibitive computational costs. By integrating data from simulations of varying fidelity, the approach enables researchers to scale up machine learning training while retaining the precision needed for demanding scientific applications. The methodology could accelerate discovery pipelines in pharmaceuticals, energy materials, and catalysis, offering a powerful tool for scientists seeking to explore chemical space more efficiently.

PUBLICATIONS
Benali, A., T. Plé, O. Adjoua, V. Agarawal, T. Applencourt, M. Blazhynska, R. Clay III, K. Gasperich, K. Hossain, J. Kim, C. Knight, J. T. Krogel, Y. Maday, M. Maria, M. Montes, Y. Luo, E. Posenitskiy, C. Villot, V. Vishwanath, L. Lagardère, and J.-P. Piquemal. "Pushing the Accuracy Limit of Foundation Neural Network Models with Quantum Monte Carlo Forces and Path Integrals," *arXiv* (preprint), https://doi.org/10.48550/arXiv.2504.07948

Chemistry | Simulation

Microscopic Insight into Transport Properties of Li-Battery Electrolytes

PI Wei Jiang, Argonne National Laboratory
AWARD ALCC
SYSTEM Polaris, Theta



Molecular dynamics simulation showing how different electrolytes interact with NMC622 cathodes and lithium anodes. Colors indicate different ions and molecular backbones (cyan = IL cation, white = H on PMpyr+, purple = F on PMpyrf+, yellow = FSI−, green = Li+). Density profiles along the surface are also shown. *Image: Liu et al., Nat Commun 14, 3678 (2023)*

To advance high-energy-density lithium battery chemistries, researchers are working to develop protective interphases that enhance stability and performance. With this project, a team from Argonne National Laboratory is using ALCF supercomputers to perform large-scale molecular dynamics (MD) simulations to uncover the mechanisms behind electrolyte fluorination, a key process for creating such interphases. By identifying how fluorination affects interphase formation and lithium-ion transport, the team's research provides key insights for designing next-generation electrolytes.

CHALLENGE Creating stable, high-performance lithium batteries requires a robust interphase formed by electrolyte decomposition, but identifying the optimal electrolyte formulations is costly and time-consuming. Electrolyte fluorination, the process of introducing fluorine atoms to the electrolyte molecules, has shown to be a promising solution for creating durable protective layers and improving battery stability. However, developing an effective fluorination strategy without relying on extensive lab work remains a significant obstacle.

APPROACH Using the NAMD code on ALCF supercomputers, the team conducted extensive MD simulations to study how electrolyte fluorination affects interphase formation and charge conduction under realistic conditions. The researchers employed novel all-atom models for high-capacity Li-metal anodes and Mn-Ni-O cathodes to investigate atomic-scale interface structural changes induced by fluorination. Enhanced sampling techniques were applied to efficiently explore the dynamics of viscous electrolytes, requiring large timescales beyond brute-force simulations. Multiple replicas of each simulated system were run in parallel, allowing for comprehensive configurational sampling to accurately capture the effects of fluorination on electrolyte-electrode interactions.

RESULTS The team's simulations provide new insights into the fundamental mechanism of fluorination and its effects on battery stability, lifetime, and capacity. Their research identified how fluorination facilitates the formation of a robust protective interphase, which enhances electrolyte stability and improves charge conduction. The work also demonstrated that the electrolytes exhibit distinct behaviors based on their molecular structure—fluorinated diluents with cyclic structures improved lithium-ion mobility, while those with linear structures restricted charge transport. Statistical analysis revealed that fluorination must be applied to specific positions on electrolyte molecules or particular molecular structures of diluents to achieve optimal performance, paving the way for the rational design of future fluorinated electrolytes.

IMPACT This research accelerates electrolyte discovery by reducing reliance on extensive wet-lab experiments, improving safety, and enhancing the rational design of fluorinated electrolytes. The insights gained from the team's simulations open new pathways for optimizing lithium battery performance, supporting the development of more efficient and longer-lasting energy storage solutions.

PUBLICATIONS

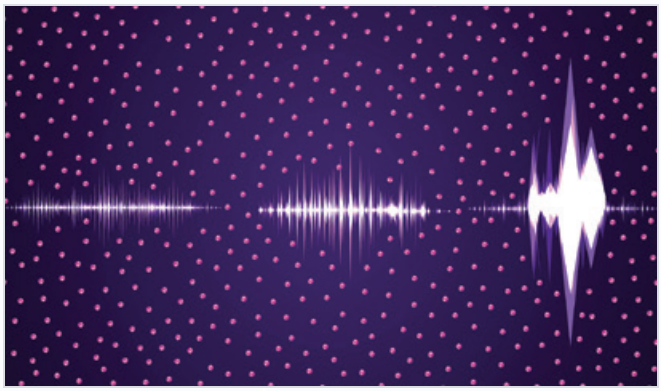
Liu, Q., J. Xu, W. Jiang, J. Gim, A.P. Tornheim, R. Pathak, Q. Zhu, P. Zuo, Z. Yang, K.Z. Pupek, E. Lee, C. Wang, C. Liu, J.R. Croy, K. Xu, and Z. Zhang. "High-Energy $LiNiO_2$ Li Metal Batteries Enabled by Hybrid Electrolyte Consisting of Ionic Liquid and Weakly Solvating Fluorinated Ether," *Advanced Science* (October 2024), John Wiley and Sons. https://doi.org/10.1002/advs.202409662.

Liu, Q., W. Jiang, J. Xu, Y. Xu, Z. Yang, D.-J. Yoo, K.Z. Pupek, C. Wang, C. Liu, K. Xu, and Z. Zhang. "A Fluorinated Cation Introduces New Interphasial Chemistries to Enable High-Voltage Lithium Metal Batteries," *Nature Communications* (June 2023), Springer Nature. https://doi.org/10.1038/s41467-023-38229-7.

Chemistry | Simulation

Super-Resolution Stimulated X-ray Raman Spectroscopy

PI Phay Ho, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Theta



An incoming x-ray light wave (left) made up of a chaotic distribution of very fast spikes interacts with atoms (purple dots) in a gas to amplify specific spikes (right) in the light wave. *Image: Stacy Huang, Argonne National Laboratory*

Understanding how electrons move and interact on ultrafast timescales is key to advancing chemistry and materials science, but existing x-ray techniques lack the resolution to capture these dynamics on excited electronic state surfaces in sufficient detail. Using ALCF supercomputers, an international research team led by Argonne National Laboratory is developing a new x-ray spectroscopy technique that resolves the electronic states of materials with unprecedented clarity.

CHALLENGE Stimulated x-ray Raman scattering (SXRS) has long been viewed as a promising method for probing electron motion in atoms and molecules, but achieving high-resolution measurements poses a significant challenge. Conventional approaches are restricted by the broad bandwidth of x-ray free-electron laser pulses and the resolution limits of spectrometers, preventing clear separation of closely spaced energy levels. For elements like neon, these subtle electronic structure feature carry critical information about many-body electron interactions. Overcoming these constraints requires both novel experimental methods and advanced modeling to interpret complex light-matter interactions.

APPROACH The team developed a stochastic SXRS method that exploits natural fluctuations in x-ray laser pulses to extract high-resolution spectra. By applying covariance analysis, which correlates each incoming pulse's fine spectral structure with its Raman-scattered signal, in conjunction with super-resolution peak finding methods awarded the Chemistry Nobel Prize in 2014, they bypassed conventional resolution limits. Working closely with ALCF staff, they optimized their TDSE/MWE code for the Theta supercomputer and implemented software tools to efficiently run large ensembles of calculations. These simulations modeled how intense, spectrally structured x-ray pulses propagate through neon gas and how stimulated Raman scattering interacts with competing processes like x-ray lasing, providing valuable insights for interpreting experimental results and refining the technique.

RESULTS The team demonstrated SXRS with spectroscopic precision, detecting valence-excited states in neon with a near Fourier-limited joint energy-time resolution of 0.1 eV–40 femtoseconds. Using stochastic covariance analysis over 18,000 single shots, they surpassed both the ~8 eV bandwidth of the incident beam and the ~0.2 eV instrumental resolution, creating super-resolution conditions analogous to super-resolved fluorescence microscopy. The experiments, supported by ab initio propagation simulations, revealed the competition between lasing in the ion and stimulated Raman scattering in the neutral atom. The method achieved orders-of-magnitude higher efficiency than spontaneous Raman scattering, with enhanced signal collection and a broad excitation window enabling rapid, high-fidelity mapping of neon's electronic structure.

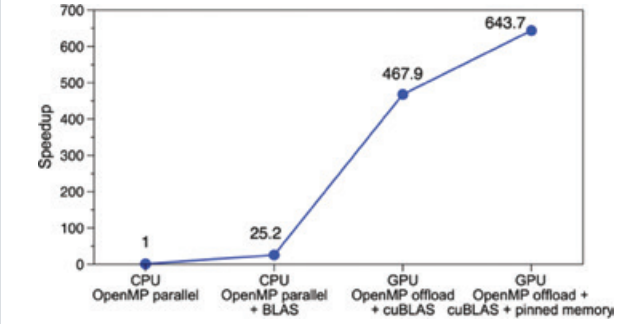
IMPACT By integrating high-performance computing with advanced x-ray facilities, the team's approach transforms what was once considered experimental noise into a tool for ultrafast, high-precision spectroscopy. The method can be extended to study complex molecules and materials, advancing research in areas such as catalysis, quantum materials, and energy science.

PUBLICATIONS
Li, K., C. Ott, M. Agåker, P. J. Ho, G. Doumy, A. Magunia, M. Rebholz, M. Simon, T. Mazza, A. De Fanis, T. M. Baumann, J. Montano, N. Rennhack, S. Usenko, Y. Ovcharenko, K. Chordiya, L. Cheng, J.-E. Rubensson, M. Meyer, T. Pfeifer, M. B. Gaarde, and L. Young. "Super-resolution Stimulated X-ray Raman Spectroscopy," *Nature* (July 2025), Springer Nature. https://doi.org/10.1038/s41586-025-09214-5

Computer Science | Simulation

Accelerating Light-Matter Dynamics Simulations with DCMESH

PI Aiichiro Nakano, University of Southern California
AWARD INCITE, Aurora ESP
SYSTEM Aurora, Polaris



Speedup over the baseline DCMESH code on a single Polaris node resulting from a series of code optimizations. Measurement was made using a single OpenMP thread for simplicity. *Image: Aiichiro Nakano, Taufeq Mohammed Razakh, University of Southern California*

Simulating quantum systems and complex materials is among the most computationally demanding challenges in science. Such simulations are critical for advancing technologies in energy, electronics, and quantum information, but the calculations often scale beyond the limits of conventional methods. To prepare for the exascale era, researchers from the University of Southern California are leveraging ALCF resources to develop GPU-optimized algorithms for density functional theory (DFT) simulations and precision-aware strategies for electronic structure calculations. Central to both efforts is the team's Divide-and-Conquer Maxwell-Ehrenfest Surface Hopping (DCMESH) framework, which enables scalable simulations of light-matter interactions.

CHALLENGE For light-matter interactions, modeling electron dynamics in real time requires propagating wavefunctions with high fidelity across thousands of time steps, a process that can be prohibitively slow. DFT-based electronic structure codes bottleneck as they approach exascale, with matrix operations and floating-point precision choices heavily influencing performance and accuracy. Overcoming these challenges requires not just more powerful supercomputers but also innovations in algorithms tailored to modern architectures.

APPROACH For light-matter dynamics, the team optimized DCMESH kernels for GPUs, minimizing CPU-GPU data transfer using shadow dynamics and hierarchical offloading while accelerating the most compute-intensive linear algebra routines. These simulations were run on ALCF's Polaris and on the Intel Data Center GPU Max Series hardware that powers Aurora. The team also systematically studied how different BLAS (Basic Linear Algebra Subprograms) precision modes affect performance and accuracy in DCMESH, identifying the optimal balance for large-scale DFT simulations.

RESULTS The team's DCMESH implementation achieved up to 644x speedup on Nvidia A100 GPUs compared with CPU-based methods, enabling detailed simulations of ultrafast light-induced phenomena. On a single Intel Data Center GPU Max Series 1550, their use of mixed-precision BLAS routines delivered a 1.35x speedup while maintaining accuracy in key outputs such as excited electron counts, current density, and kinetic energy, reaching up to 3.9x for large-scale operations. Weak-scaling tests on 256 Polaris nodes (1,024 GPUs) maintained parallel efficiency above 96 percent, demonstrating the framework's scalability. These advances establish the foundation for even larger, extreme-scale simulations on Aurora and other exascale systems.

IMPACT By advancing both GPU acceleration for quantum dynamics and precision-aware strategies for electronic structure calculations, this work lays critical groundwork for the efficient use of Aurora and other exascale systems. These innovations will help researchers probe fundamental processes in light-matter interactions, accelerate the discovery of new materials, and extend the scientific reach of quantum simulations. The team's precision-aware approach is broadly applicable to other HPC workloads dominated by linear algebra.

PUBLICATIONS

Razakh, T. M., T. Linker, Y. Luo, R. K. Kalia, K.-I. Nomura, and P. Vashishta. "Accelerating Quantum Light-Matter Dynamics on Graphics Processing Units," *2024 IEEE International Parallel and Distributed Processing Symposium Workshops* (May 2024), IEEE. https://doi.org/10.1109/IPDPSW63119.2024.00176

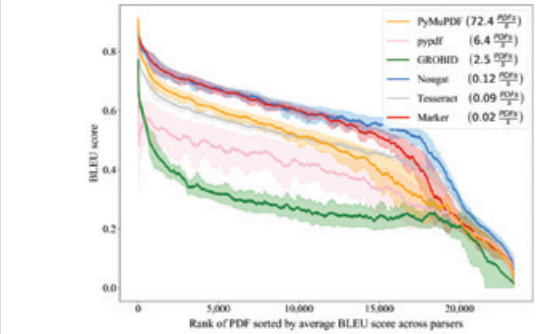
Piroozan, N., S. J. Pennycook, T. M. Razakh, P. Caday, N. Kumar, and A. Nakano. "Impact of Varying BLAS Precision on DCMESH," *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis* (November 2024), IEEE. https://doi.org/10.1109/SCW63240.2024.00187

AdaParse: Smart PDF Processing for Scientific AI Training

PI Rick Stevens, Argonne National Laboratory and University of Chicago

AWARD INCITE

SYSTEM Polaris



The team evaluated the performance and scalability of seven different parsers on the ALCF's Polaris supercomputer. *Image: Carlo Siebenschuh, Argonne National Laboratory and University of Chicago*

Scientific progress increasingly depends on large-scale analysis of research papers, but most academic writing is organized in PDF format, which is difficult to process reliably. Even small parsing errors, such as changing "hyperthyroidism" to "hypothyroidism" or "pH" to "Ph," can completely alter scientific meaning and compromise downstream research. To address this problem, a team from Argonne National Laboratory and the University of Chicago is leveraging ALCF supercomputers to develop the Adaptive Parallel PDF Parsing and Resource Scaling Engine (AdaParse) and to apply it to parse large numbers of scientific papers.

CHALLENGE Parsing PDFs presents a fundamental obstacle for building multimodal science models and AI tools. PDFs are optimized for visual appearance rather than readability, meaning that text may be scrambled, chemical formulas corrupted, or figures and tables misinterpreted. Traditional parsing tools face a trade-off between speed and accuracy. Fast parsers often introduce serious errors, while high-quality approaches are slow and computationally expensive, limiting their use in large-scale processing efforts. Overcoming these constraints requires new methods that are simultaneously accurate, efficient, and scalable to millions of documents across diverse domains.

APPROACH The team developed AdaParse with an adaptive design that dynamically adjusts parsing strategies based on the input data. AdaParse uses a three-stage adaptive process. First, it quickly extracts text to assess document characteristics. Then, a machine learning model predicts which parsing method will yield the highest-quality results. Finally, the selected parser is applied across large collections of documents while distributing workloads across multiple nodes to optimize computational efficiency. The researchers incorporated human-in-the-loop principles and human expert feedback through direct preference optimization, aligning the system's choices with what scientists prefer. The team evaluated the framework using ALCF's Polaris supercomputer, enabling tests with large-scale, production-like log datasets and stress-testing performance on distributed-memory architectures representative of exascale systems.

RESULTS In benchmarking tests with 25,000 scientific documents spanning eight research domains and six major publishers, parses produced by AdaParse were preferred by scientists more often than any single parser, while running up to 17 times faster than state-of-the-art high-quality approaches. The adaptive selection process reduced errors like corrupted formulas or scrambled characters while maintaining high throughput. By scaling efficiently across distributed-memory architectures, AdaParse demonstrated the ability to process millions of documents at rates suitable for training advanced AI systems on scientific text.

IMPACT AdaParse speeds up the creation of AI systems trained on scientific literature, leading to better AI research assistants, improved scientific discovery tools, and more accessible scientific knowledge. It has the potential to expand access to large-scale scientific datasets needed for training advanced AI models.

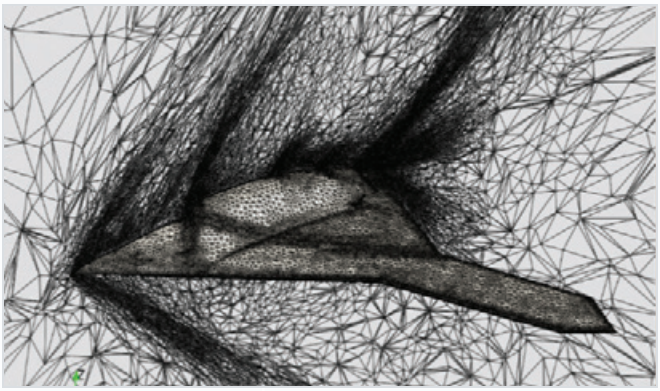
PUBLICATIONS

Siebenschuh, C., K. Hippe, O. Gokdemir, A. Brace, A. M. Khan, K. Hossain, Y. Babuji, N. Chia, V. Vishwanath, A. Ramanathan, R. L. Stevens, I. Foster, and R. Underwood. "AdaParse: An Adaptive Parallel PDF Parsing and Resource Scaling Engine," *Proceedings of the 8th MLSys Conference* (May 2025) https://doi.org/10.48550/arXiv.2505.01435

Computer Science | Simulation, Learning

Autonomy for DOE Simulations

PI Allan Grosvenor, MSBAI
AWARD ALCC
SYSTEM Aurora, Polaris



Solution-adaptive mesh refinement example of GURU autonomously setting up CFD simulations for high-speed aerodynamics. *Image: Allan Grosvenor, MSBAI*

Setting up high-fidelity simulations for complex engineering tasks, such as modeling aircraft aerodynamics or predicting satellite behavior, can take experts hours or even days to complete. These time-intensive processes remain a major barrier to broader use of high-performance computing (HPC) and simulation tools. To overcome this challenge, researchers at MSBAI are using ALCF computing resources to advance GURU, an AI-powered hybrid intelligence system that autonomously configures engineering simulations, enabling faster, more accessible use of HPC tools across industries.

CHALLENGE Mission-critical simulation workflows demand decision-making that is not only fast and accurate, but also explainable, scalable, and adaptable to novel scenarios. Traditional AI models often struggle to meet all of these requirements: symbolic systems are interpretable but brittle, while deep learning approaches are powerful but opaque and inflexible. In simulation-driven engineering, this tradeoff has limited automation efforts—particularly in tasks like computational fluid dynamics (CFD) meshing, where incorrect configurations can lead to failure or unusable results. Bridging this gap requires a new AI architecture that combines the strengths of both symbolic reasoning and data-driven learning.

APPROACH The MSBAI team developed a hybrid intelligence framework inspired by cognitive theories like Global Workspace Theory. Their system integrates multiple layers of AI agents—including symbolic rule-based systems, graph neural networks, reinforcement learning agents, and transformers—to automate end-to-end workflows. Using ALCF's Aurora system, the team trained and scaled these agents across thousands of GPUs, applying them to complex CFD and space-domain awareness tasks. For CFD setup, they implemented an AutoML-guided optimizer capable of refining mesh parameters, while planning agents orchestrated workflow steps based on real-time performance feedback.

RESULTS The GURU platform successfully automated the CFD mesh-generation process across hundreds of aircraft geometries. The system raised boundary-layer capture propagation from an initial 8 percent to 98 percent while cutting mesh failure rates from 88 percent to just 2 percent. Reinforcement learning agents, combined with symbolic constraints and context-aware embeddings, played a critical role in this success. The hybrid architecture also scaled efficiently on leadership-class systems, sustaining over 88 percent parallel efficiency on more than 1,000 compute nodes and showing strong performance in a second domain: detecting satellite maneuvers in near-real time.

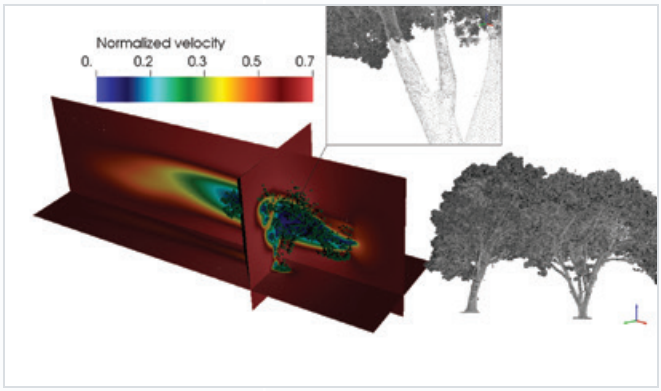
IMPACT This work demonstrates a powerful new model for AI-driven automation of engineering simulations. By blending symbolic logic with large-scale machine learning, the GURU platform offers a scalable, interpretable, and generalizable solution for accelerating HPC adoption across industries. Its success paves the way for broader integration of intelligent agents into simulation workflows for aerospace, energy, manufacturing, and beyond, advancing the DOE's exascale computing goals and helping industry unlock massive design and innovation potential.

PUBLICATIONS
Grosvenor, A., A. Zemlyansky, A. Wahab, K. Bohachov, A. Dogan, and D. Deighan. "Hybrid Intelligence Systems for Reliable Automation: Advancing Knowledge Work and Autonomous Operations with Scalable AI Architectures," *Frontiers in Robotics and AI* (July 2025), Frontiers Media. https://doi.org/10.3389/frobt.2025.1566623

Computer Science | Simulation

Bridging Gaps in Simulation Analysis through a General Purpose, Bidirectional Steering Interface with Ascent

PI Victor Mateevitsi, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



An example flow field for a real tree case from Lidar data by USGS.
Image: Zhuoting Wu, United States Geological Survey

Interactive steering of large-scale scientific simulations promises to enhance scientific productivity by enabling real-time user intervention. However, existing in situ workflows typically rely on automated triggers and lack flexible human-in-the-loop control mechanisms. To address this gap, researchers developed a general-purpose bidirectional steering interface integrated with the Ascent framework, allowing scientists to pause simulations and interactively adjust parameters and workflows during runtime.

CHALLENGE Traditional in situ analysis and visualization workflows for large-scale scientific simulations rely on automated triggers to execute tasks at predefined simulation states. While effective for many applications, these automated mechanisms limit the ability of scientists to interactively steer simulations in real time, especially when domain expertise is essential to determine appropriate interventions. Existing in situ frameworks either lack flexible bidirectional steering interfaces or require complex, simulation-specific customizations, creating a barrier to human-in-the-loop simulation control.

APPROACH Researchers developed a novel, general-purpose bidirectional steering interface built on the Ascent in situ framework, enabling interactive human-in-the-loop control of existing simulations without requiring bespoke infrastructure. This interface supports registering, listing, and invoking simulation callbacks and system commands via two user steering interfaces: a lightweight terminal-based interface and a Jupyter-notebook interface connected through an SSH tunnel. Leveraging Ascent's trigger infrastructure, simulations can be paused at key moments, allowing users to interactively adjust parameters, invoke callbacks, and resume execution without costly restarts.

RESULTS Two use cases demonstrated the interface's flexibility and productivity benefits. The first used a lattice Boltzmann CFD simulation to rescue unstable runs by interactively reverting to stable checkpoints and adjusting timestep sizes, successfully preventing crashes without restarting. The second optimized a vegetation canopy CFD simulation by dynamically swapping Lidar dataset resolutions and adjusting forcing parameters during runtime, thereby streamlining the process of finding optimal simulation parameters. In both cases, interactive steering reduced overhead, accelerated troubleshooting, and enabled nuanced scientific judgment in steering decisions.

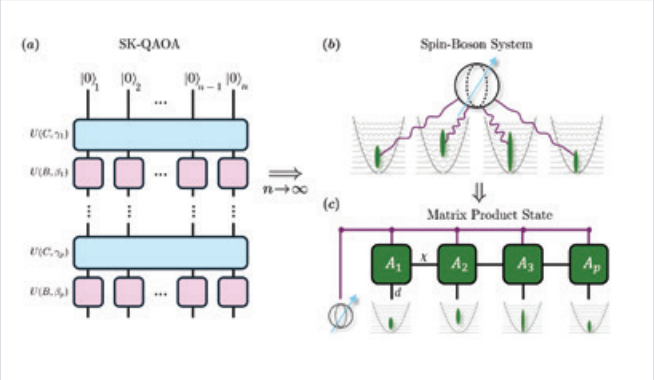
IMPACT This general-purpose bidirectional steering mechanism empowers scientists to perform real-time, human-in-the-loop interventions in large-scale simulations, improving efficiency and enabling more flexible exploratory workflows. Integrated into the widely used Ascent framework, it provides a practical path to advance interactive simulation steering across diverse scientific domains, addressing a critical gap in existing in situ methods. This capability holds promise to accelerate scientific discovery by allowing researchers to steer simulations responsively, reducing downtime and improving insight.

PUBLICATIONS
Sewell, A., D. K. Fytanidis, V. A. Mateevitsi, C. Harrison, N. Marsaglia, T. Marrinan, S. Rizzi, J. A. Insley, M. E. Papka, and S. Petruzza. "Bridging Gaps in Simulation Analysis through a General Purpose, Bidirectional Steering Interface with Ascent," *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2024), IEEE. https://doi.org/10.1109/SCW63240.2024.00119

Computer Science | Simulation

Evidence that the QAOA Optimizes the Sherrington-Kirkpatrick Model Efficiently in the Average Case

PI Ruslan Shaydulin, JPMorganChase
AWARD Director's Discretionary
SYSTEM Polaris



Overview of the main technical result: the equivalence between the task of computing the infinite-size limit QAOA energy in the average case and the task of simulating a spin-boson system. *Image: Abid Khan, JPMorganChase*

Quantum computers offer a path to solving optimization problems that are intractable for even the most powerful classical machines. A foundational method in quantum computing research, the Quantum Approximate Optimization Algorithm (QAOA) is a leading hybrid classical-quantum algorithm designed to tackle complex combinatorial optimization problems in domains such as finance, physics, and machine learning. Using the ALCF's Polaris supercomputer, researchers from JPMorganChase and Argonne National Laboratory have demonstrated that QAOA can efficiently approximate solutions to a challenging model for disordered systems, providing new insights into the performance and scalability of quantum algorithms.

CHALLENGE The Sherrington-Kirkpatrick (SK) model has become a key tool for understanding the behavior of complex energy landscapes. Finding its exact ground state is an NP-hard optimization problem in the worst case, and even approximate solutions typically require super-polynomial time on classical machines. While quantum algorithms like QAOA have been proposed to outperform classical methods, past studies of QAOA on the SK model were limited by exponential computational costs that prevented exploration at large circuit depths.

APPROACH To overcome this bottleneck, the researchers leveraged the ALCF's Polaris supercomputer to develop a method to compute average QAOA energy for the SK model in the infinite-size limit by mapping it to a spin-1/2 particle coupled to multiple bosonic modes. The team used tensor network simulations, specifically matrix product state (MPS) techniques, to model this spin-boson system efficiently while constraining entanglement and truncating the Fock-space dimension. By distributing workloads across up to 160 NVIDIA GPUs, Polaris enabled large-scale parallel simulations, with QAOA parameters optimized up to depth p = 80 and energies computed up to p = 160.

RESULTS The team demonstrated that QAOA achieves progressively better approximations to the SK model's optimal energy as circuit depth increases, reaching a 2.2 percent deviation at depth 160. Their analysis suggests that a $(1 - \varepsilon)$ approximation can be achieved in time scaling as $O(n/\varepsilon^{1.13})$, providing numerical evidence for a potential polynomial quantum speedup over classical approaches. When applied to finite-size SK instances (up to 30 qubits) using parameters from the infinite-size simulations, QAOA achieved success probabilities consistent with theoretical predictions, demonstrating that it can efficiently approximate the SK model in the average case.

IMPACT The team's work provides the strongest numerical evidence to date that QAOA can efficiently approximate solutions to the SK model. The techniques developed here not only extend the frontier of QAOA simulation but also provide a practical benchmark for evaluating quantum advantage. As quantum hardware continues to evolve, these findings support the promise of QAOA as a scalable algorithm for combinatorial optimization problems in finance, physics, and beyond.

PUBLICATIONS
Boulebnane, S., A. Khan, M. Liu, J. Larson, D. Herman, R. Shaydulin, and M. Pistoia. "Evidence that the Quantum Approximate Optimization Algorithm Optimizes the Sherrington-Kirkpatrick Model Efficiently in the Average Case," *arXiv* (preprint). https://doi.org/10.48550/arXiv.2505.07929

Opportunities for Retrieval and Tool-Augmented Large Language Models in Scientific Facilities

PI Michael Prince, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris

Overview of CALMS: CALMS uses a large language model in conjunction with conversational memory, document stores, and experimental tools to answer user queries or take action to drive an instrument. *Image: Prince et al. npj Comput Mater 10, 251 (2024).*

Upgrades to advanced scientific user facilities and instruments such as next-generation x-ray light sources, nanoscience centers, and neutron facilities are revolutionizing our understanding of materials across the physical sciences, from biology to microelectronics. However, these upgrades significantly increase complexity over previous generations of facilities and instruments. Driven by more exacting scientific needs, tools and experiments become more intricate each year, making it ever more challenging for domain scientists to design experiments that effectively leverage the capabilities of and operate on these advanced instruments. Large language models (LLMs) can perform complex information retrieval, assist in knowledge-intensive tasks across applications, and provide guidance on tool usage. Researchers led by Argonne National Laboratory explored the potential for Context-Aware Language Model for Science (CALMS) to assist scientists with instrument operations and complex experimentation.

CHALLENGE With the ability to retrieve relevant information from documentation, CALMS can answer simple questions on scientific capabilities and other operational procedures. Designed to interface with software tools and experimental hardware, CALMS can conversationally operate scientific instruments. By making information more accessible and acting on user needs, LLMs could expand and diversify scientific facilities' users and accelerate scientific output.

APPROACH The core of CALMS consists of four components: an LLM, conversational history that allows follow-on queries, semantic search over document stores to retrieve the most relevant context given a question from the user, and instrument tools that the LLM can use when instructed to by the user. The team compared responses from two state-of-the-art LLMs, OpenAI's GPT-3.5 Turbo and an open-source model Vicuna, over questions related to experimental planning assistance, operation, and ability to drive an instrument successfully, noting that the CALMS framework is independent of the chosen LLM.

RESULTS In a paper published in *npj Computational Materials*, the researchers demonstrated that when provided with appropriate context and tools, CALMS can either answer complex technical questions of which the LLM has no prior knowledge, or it can execute an experiment or perform a computation.

IMPACT The researchers anticipate LLMs capable of leveraging decades of information recorded in e-logs, as well as fully autonomously extracting code or commands and executing experimental workflows. Context- and tool-aware language models such as CALMS have the potential to streamline the transfer of knowledge between instrument experts and domain users, with the time saved enabling researchers to accomplish more and explore new science opportunities. This could enable a paradigm shift in the way users interact with the beamline instruments, paving the path to full automation and handling of exceptions based on past experiences recorded in the form of e-logs.

PUBLICATIONS
Prince, M. H., H. Chan, A. Vriza, T. Zhou, V. K. Sastry, Y. Luo, M. T. Dearing, R. J. Harder, R. K. Vasudevan, and M. J. Cherukara. "Opportunities for Retrieval and Tool Augmented Large Language Models in Scientific Facilities," *npj Computational Materials* (November 2024), Springer Nature. https://doi.org/10.1038/s41524-024-01423-2

Computer Science | Simulation

Ponte Vecchio Across the Atlantic: Single-Node Benchmarking of Two Intel GPU Systems

PI Thomas Applencourt, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Aurora

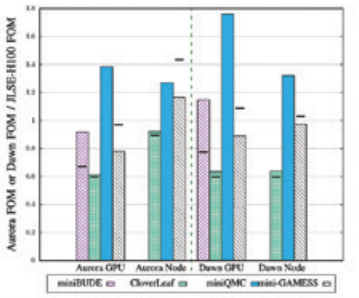


Fig. 3: Figure of Merits on Aurora and Dawn Relative to JLSE-H100. The black bars are the expected relative performance based on the microbenchmarking results and theoretical peaks.

Figures of merit for Aurora and Dawn relative to JLSE-H100. The black bars show the expected relative performance based on microbenchmarking results and theoretical peaks. *Image: PMBS*

This project provides a comprehensive performance evaluation of Intel's Ponte Vecchio GPUs with the supercomputers Aurora and Dawn, using microbenchmarks, mini-apps, and real scientific applications. The results show that the Intel GPU delivers competitive or superior performance compared to other tested GPUs, offering valuable insights for developers targeting these new architectures.

CHALLENGE High-performance computing increasingly relies on GPUs to accelerate complex scientific simulations, yet Intel's new Data Center GPU Max 1550 (Ponte Vecchio or PVC) lacks extensive benchmarking data to guide developers. Unlike the widely studied NVIDIA and AMD GPUs, PVC's novel architecture and deployment in the Aurora and Dawn supercomputers present uncertainty about its achievable performance and optimization strategies for application developers.

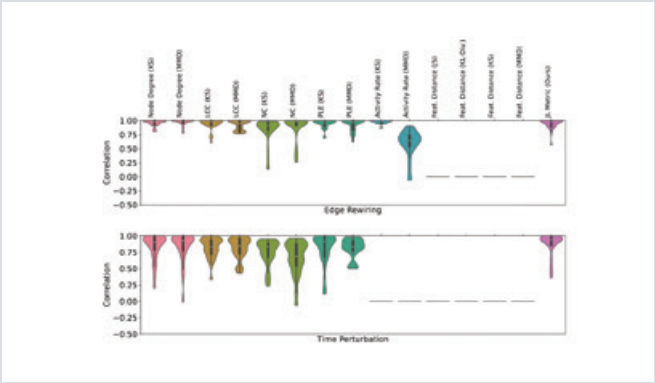
APPROACH The researchers conducted detailed microbenchmarking on single GPUs and nodes of Dawn and Aurora systems using high-level programming models (OpenMP, SYCL, MPI) to measure key performance metrics such as floating-point throughput, memory bandwidth, communication latency, and matrix multiplication performance. They evaluated four mini-applications and two full applications on PVC, comparing results with NVIDIA H100 and AMD MI250 GPUs to contextualize performance differences and identify architectural bottlenecks.

RESULTS Microbenchmarking revealed PVC's compute and memory performance is competitive with or exceeds that of NVIDIA H100 and AMD MI250 GPUs, with mini-apps on a single PVC achieving 0.6–1.8x H100 and 0.8–7.5x MI250 performance. Real-world applications like OpenMC and CRK-HACC ran efficiently on PVC-based nodes, with Aurora's 6x PVC node outperforming the NVIDIA H100 node by 1.7x in OpenMC. The study also highlighted node-level design impacts on performance and showed the robustness of PVC's software stack across multiple programming models.

IMPACT This comprehensive performance characterization provides critical reference data and insights for developers targeting PVC-based supercomputers, accelerating application optimization and adoption of Intel GPUs in high-performance computing. By demonstrating competitive performance and identifying system-level bottlenecks, this work lays the groundwork for future benchmarking and application development on emerging Intel GPU architectures, supporting scientific advancements across computational domains.

Quality Measures for Dynamic Graph Generative Models

PI Filippo Simini, Venkatram Vishwanath, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



Distributions of Spearman rank correlations across all datasets and random seeds. Correlation is measured between metric response and perturbation probability. *Image: Ryien Hosseini, University of Chicago*

Deep generative models have recently achieved significant success in representing graph data, including dynamic graphs, where both topology and features change over time. These models have applications in diverse areas, including social network analysis, infrastructure analysis of grids and networks, the study of biological functions, materials design, and financial fraud detection. However, unlike in vision or language domains, evaluating generative models for dynamic graphs is challenging because visual inspection is impractical, and existing metrics often fail to capture temporal dependencies or the interplay between node and edge features. To address this challenge, a research collaboration between Argonne National Laboratory and University of Chicago used ALCF resources to develop a new metric that provides a quantitative, interpretable measure of dynamic graph similarity, enabling more reliable evaluation of generative models.

CHALLENGE Generative models for dynamic graphs promise to accelerate discovery by producing realistic system representations at scale, but their accuracy is difficult to assess. Traditional metrics often discretize continuous-time graph data into static snapshots and compare graph statistics, assuming snapshots are independent. This approach overlooks temporal dependencies, feature evolution, and interactions between topology and features. Without metrics that capture these complexities, it is challenging to determine whether a generative model has captured the essential dynamics of a system or how to improve it.

APPROACH The researchers introduced a metric based on the Johnson-Lindenstrauss lemma, applying random projections directly to dynamic graph data to produce fixed-dimensional embeddings of variable-length node interactions. These embeddings preserve both topological and feature information while accounting for temporal evolution. The method produces a unified scalar measure that can be applied across graphs of varying sizes and time spans. Using ALCF supercomputers, the researchers scaled their framework to analyze large datasets, enabling evaluations that compare generated dynamic graphs with ground truth sequences.

RESULTS The team's method revealed subtle differences between generative models, identifying discrepancies in feature values, edge dynamics, and temporal patterns that traditional metrics often miss. By providing a scalar, expressive measure of similarity, their approach allowed generative models to be ranked and assessed systematically. Sensitivity analyses demonstrated that the metric could detect issues such as mode collapse or missing dynamic behaviors, offering a robust tool for model evaluation.

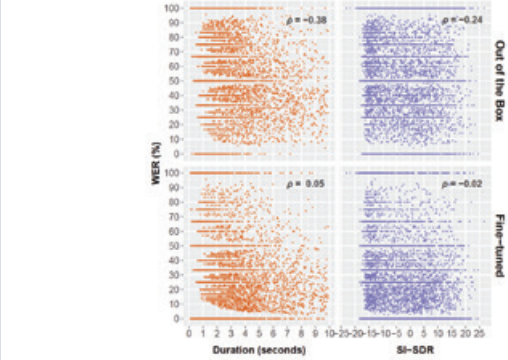
IMPACT By providing an efficient, interpretable, and quantitative evaluation method for dynamic graphs, the team's approach can help researchers assess and improve generative models more reliably. Its general applicability and scalability make it a valuable tool for modeling dynamic systems across diverse fields, from network science to biology, supporting the development of more accurate and trustworthy AI models for discovery and design.

PUBLICATIONS
Hosseini, R., F. Simini, V. Vishwanath, R. Willett, and H. Hoffmann. "Quality Measures for Dynamic Graph Generative Models," *The Thirteenth International Conference on Learning Representations (ICLR 2025)* (April 2025), Singapore, OpenReview
https://doi.org/10.48550/arXiv.2503.01720

Computer Science | Data

Speech Recognition for Analysis of Police Radio Communication

PI Christopher Graziul, University of Chicago
AWARD Director's Discretionary
SYSTEM Polaris



Word error rate (WER) vs. audio quality (SI-SDR) and utterance duration for NeMo FastConformer CTC (616M) on the dev set. *Image: Christopher Graziul, University of Chicago*

Police radio communications are a unique source of information about everyday police activity and emergency response, offering valuable insights for research into law enforcement operations and public safety. However, these audio streams are often noisy and difficult to transcribe accurately, especially at scale. A team of researchers is using ALCF computing resources to evaluate and improve automatic speech recognition (ASR) systems for police radio, enabling more effective analysis of this valuable data source.

CHALLENGE Broadcast police communication (BPC) audio is challenging for ASR systems due to multiple factors, including narrow-band audio channels, a multitude of speakers, background noise, and specialized vocabulary. Publicly available ASR systems, which are often trained on clean audio, struggle with high word error rates in this domain. Creating accurate transcripts is essential for downstream tasks such as event detection, call classification, and trend analysis, but retraining ASR models for BPC data requires substantial computational power and domain-specific expertise.

APPROACH The team assembled a new corpus of roughly 62,000 manually transcribed police radio transmissions (about 46 hours of audio) from the Chicago Police Department to benchmark model performance. Using ALCF computing resources, they evaluated state-of-the-art ASR systems, including Whisper Large (v3), and customized ESPNet models with domain-specific token vocabularies and language models. Their work involved using multiple audio feature extractors, including log Mel-filterbank features, HuBERT Large, WavLM Large, and a feature fusion model, and conducting large-scale fine-tuning on the ALCF's Polaris supercomputer. In addition, the team performed detailed error analyses to identify recurring transcription failures and guide future improvements.

RESULTS Baseline evaluations showed that off-the-shelf ASR models performed poorly on BPC audio, with high rates of substitution, deletion, and insertion errors. Fine-tuned models trained on Polaris achieved substantial accuracy gains over their baseline counterparts. The researchers also quantified inter-annotator agreement to better understand the inherent ambiguity in transcribing BPC, helping to set realistic performance expectations for ASR systems. Their work establishes new benchmarks for ASR performance in police radio transcription and provides a framework for systematically improving accuracy in this domain.

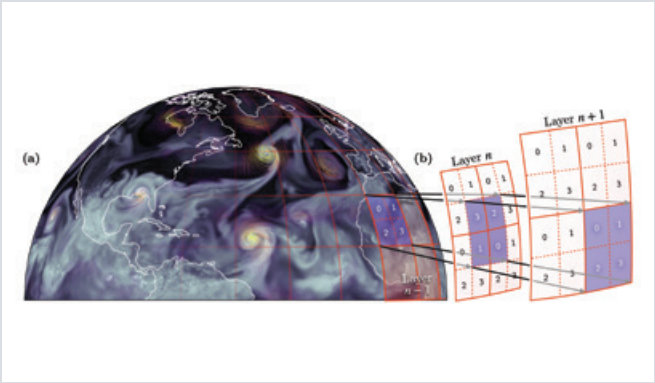
IMPACT By advancing speech recognition for noisy, domain-specific audio like police radio communications, this work supports the development of tools for analyzing public safety and emergency response data. The team made its large corpus of police radio transmissions, along with the associated data annotation pipeline, available to the research community to enable further study on the large-scale transcription and analysis of police communications.

PUBLICATIONS
Srivastava, T., J.-C. Chou, P. Shroff, K. Livescu, and C. Graziul. "Speech Recognition for Analysis of Police Radio Communication," *2024 IEEE Spoken Language Technology Workshop* (December 2024), IEEE. https://doi.org/10.1109/SLT61566.2024.10832157

Earth Science | Simulation, Data, Learning

AERIS: Argonne Earth Systems Model for Reliable and Skillful Predictions

PI Rao Kotamarthi, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Aurora



Forecast over the Atlantic basin highlighting Hurricane Teddy, Tropical Storm Sally, and Post-Tropical Cyclone Paulette (a); sequence-window parallelism in AERIS showing how example windows are partitioned across network layers (b). *Image: Argonne National Laboratory*

Modeling complex non-linear systems, such as the Earth system, to make skillful predictions beyond a few days is challenging due to computational costs and decreasing predictability over time caused by chaotic behavior inherent in the systems. To address this, researchers are using ALCF's Aurora exascale supercomputer to develop AERIS, a multi-billion-parameter probabilistic foundation model for weather and climate forecasting, trained to generate efficient, accurate forecasts from storm-scale events to seasonal, planetary circulation.

CHALLENGE Earth system processes span vast spatial and temporal ranges, from localized convection to global circulation patterns. Current approaches often rely on specialized models at each scale, where errors accumulate across them, and computational complexity restricts the number of ensemble members that can be generated. Data-driven models show promise but face difficulties scaling to realistic resolutions, embedding physical constraints, and managing the extreme computational demands of AI training. Advancing research in this area requires a next-generation model that is capable of unifying local and global scales, maintaining physical fidelity, producing probabilistic forecasts and training efficiently on extreme-scale supercomputers.

APPROACH The team built AERIS using a generative, pixel-level Swin Transformer adapted for data-driven weather modeling. Central to its scalability is the team's novel method called SWiPe (Sequence-Window Pipeline parallelism), which integrates window, sequence, and pipeline parallelism to efficiently shard the model without added communication overhead. Using ALCF's Aurora system, the researchers trained AERIS on 16 terabytes of ECMWF Reanalysis v5 (ERA5) data encompassing several years of global atmospheric conditions. SWiPe reduced activation memory requirements and minimized communication costs, enabling stable training of models ranging from 1.3 to 80 billion parameters. This marks the first pixel-level diffusion model in the billion parameter regime guided by physical priors in AI for Science.

RESULTS On Aurora, AERIS achieved a sustained, mixed-precision performance of 10.21 exaflops and a peak of 11.21 exaflops on 10,080 nodes (120,960 GPU tiles)—over 95 percent of the full machine. The training strategy demonstrates nearly linear scaling across thousands of GPUs while maintaining numerical stability with a generative, physics-informed loss function. In forecasting tests, AERIS delivered skillful medium-range predictions competitive with state-of-the-art numerical and machine learning models, and remained stable in 90-day rollouts, highlighting its potential for subseasonal-to-seasonal forecasting.

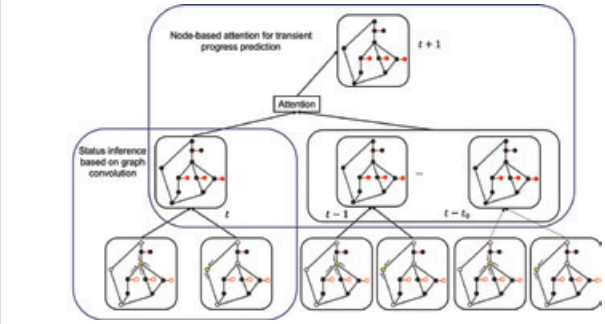
IMPACT By demonstrating how generative models can perform across spatial scales, AERIS provides a blueprint for next-generation probabilistic forecasting tools for seasonal-to-subseasonal time scales. Trained on Aurora using a novel parallelization strategy, the model advances the scientific community's ability to generate stable, high-resolution forecasts, improve subseasonal outlooks, and inform energy production and impacts decisions sensitive to extreme weather and environmental risk.

PUBLICATIONS
Hatanpää, V., E. Ku, J. Stock, M. Emani, S. Foreman, C. Jung, S. Madireddy, T. Nguyen, V. Sastry, R. A. O. Sinurat, S. Wheeler, H. Zheng, T. Arcomano, V. Vishwanath, and R. Kotamarthi. "AERIS: Argonne Earth Systems Model for Reliable and Skillful Predictions," *arXiv* (preprint). https://doi.org/10.48550/arXiv.2509.13523

Energy Technologies | Simulation, Learning

Development of Whole System Digital Twins for Advanced Reactors

PI Rui Hu, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Sophia



Heterogenous graph neural networks-based digital twin: combination of graph convolution and node attention for transient progress prediction. *Image: Liu et al., Nuclear Technology 211(9), 2206–23 (2024).*

Modeling and predicting the behavior of complex nuclear reactor systems at scale presents a critical challenge for advancing reactor technology. Conventional simulation methods often struggle to deliver rapid, accurate insights necessary for real-time reactor operation and decision-making. To overcome this, Argonne researchers have developed an innovative digital twin technology that leverages artificial intelligence to improve reactor efficiency, reliability, and safety.

CHALLENGE Traditional reactor simulation tools, while accurate, are computationally intensive and slow, limiting their utility for real-time monitoring and control. Furthermore, capturing the intricate interactions among diverse reactor components requires a modeling approach that can handle complex relational data efficiently. Addressing these challenges is essential for next-generation reactors, including small modular and microreactors, which demand enhanced operational agility.

APPROACH The team introduced a novel methodology that represents entire nuclear reactor systems as heterogeneous graphs, where nodes correspond to various physical components and edges represent their interconnections. Using graph neural networks (GNNs), which excel at capturing relationships in such data, the researchers developed a digital twin that models the reactor's dynamic behavior. Training utilized simulation data generated by Argonne's System Analysis Module (SAM), and computing resources from the ALCF enabled scalable training and uncertainty quantification.

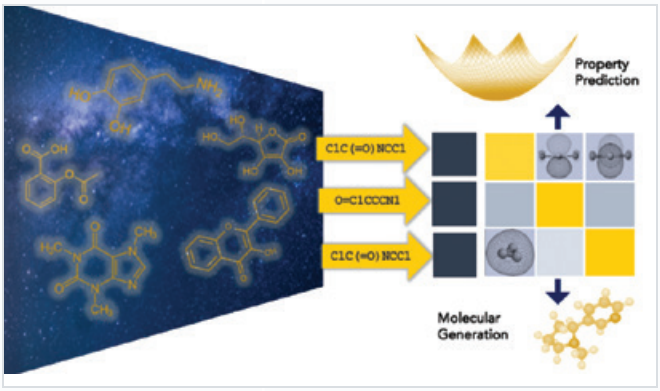
RESULTS The GNN-based digital twin rapidly and accurately predicts reactor responses to operational transients, including power fluctuations and cooling system changes. Case studies on the Experimental Breeder Reactor II (EBR-II) and a generic fluoride-salt-cooled high-temperature reactor (gFHR) demonstrated the model's precision and speed, surpassing traditional simulations. The digital twin can infer whole-system status from sparse sensor data and detect anomalies early, supporting proactive maintenance and reducing operational costs.

IMPACT This AI-driven digital twin approach offers a transformative tool for managing advanced nuclear reactors, enhancing safety, autonomy, and efficiency. By enabling real-time system monitoring and predictive control, the technology supports longer component lifespans and lowers operational expenses. The work represents a significant advancement toward intelligent reactor operations, with broad implications for the future deployment of safe, cost-effective nuclear energy solutions.

PUBLICATIONS
Liu, Y., F. Alsafadi, T. Mui, D. O'Grady, and R. Hui. "Development of Whole System Digital Twins for Advanced Reactors: Leveraging Graph Neural Networks and SAM Simulations," *Nuclear Technology* (October 2024), Taylor and Francis Group. https://doi.org/10.1080/00295450.2024.2385214

Foundation Models for Molecular Design for Energy Storage and Conversion

PI Venkat Viswanathan, University of Michigan
AWARD INCITE
SYSTEM Aurora, Polaris



The team is using ALCF supercomputers to develop foundation models that aim to accelerate the discovery of new battery materials. *Image: Anoushka Bhutani, University of Michigan*

Developing safer, longer-lasting, and higher-performing batteries is critical for applications ranging from electric vehicles to grid storage and aviation. Using ALCF supercomputers, a University of Michigan-led team is building massive foundation models to accelerate battery materials research, providing tools that help researchers predict key properties and discover next-generation battery electrolytes and electrodes.

CHALLENGE The chemical design space for potential battery materials is immense with an estimated 10^{60} possible molecular compounds, making exhaustive experimental or computational testing infeasible. Foundation models can help researchers explore this space more efficiently, but training such models on data from billions of known molecules requires computing power beyond the capabilities of most research labs. Before gaining access to ALCF resources, the team's largest models topped out at 100 million molecules, falling short of the performance of other state-of-the-art AI models. The researchers also found that widely used molecular tokenization methods, which break chemical structures into machine-readable elements, often fail to capture critical details, limiting model accuracy.

APPROACH With computing time on the ALCF's Polaris and Aurora systems awarded through DOE's INCITE program, the team is scaling training to billions of molecules, starting with small-molecule electrolytes and expanding to crystal structures for electrodes. Molecular structures are represented using SMILES strings. To improve model processing, the researchers developed Smirk and Smirk-GPE —two open-vocabulary tokenizers that fully cover the OpenSMILES specification. These methods avoid the gaps and ambiguities of conventional tokenizers while maintaining computational efficiency. The team has also worked closely with ALCF staff at INCITE hackathons to optimize workflows for large-scale AI training.

RESULTS In their work to develop a foundation model for electrolytes on Polaris, the researchers successfully trained one of the largest chemical foundation models to date, unifying capabilities that previously required multiple separate AI tools and outperforming their earlier property prediction models. Building on this work, they are now using Aurora to develop a second foundation model aimed at battery electrodes. The team's new tokenizers have demonstrated improved chemical coverage and flexibility compared to existing chemistry-specific tokenizers. Smirk's design allows for integration of nuclear, electronic, and geometric information, broadening its potential applications beyond batteries to areas such as biology, agriculture, and pharmacology.

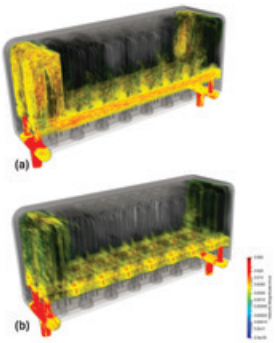
IMPACT By enabling the screening of billions of potential materials in a fraction of the time required by traditional methods, the team's work could greatly accelerate the discovery of safe, high-performance battery components. Beyond energy storage, the team's advances in large-scale AI training and tokenization provide tools and approaches that could be applied to other scientific domains where vast chemical or material design spaces remain largely unexplored.

PUBLICATIONS
Wadell, A., A. Bhutani, and V. Viswanathan. "Tokenization for Molecular Foundation Models," *arXiv* (preprint). https://doi.org/10.48550/arXiv.2409.15370

Engineering | Simulation

Exascale MHD Simulations for Liquid Metal Fusion Blankets

PI Misun Min, Argonne National Laboratory
AWARD ALCC
SYSTEM Aurora, Polaris



Volume rendering of NekRS simulations demonstrating the significance of MHD effects, comparing velocity throughout the Water-Cooled Lithium Lead Test Blanket Module between MHD off (top) and MHD on (bottom). *Image: Misun Min, Argonne National Laboratory*

Researchers led by Argonne National Laboratory leveraged exascale supercomputing resources and the application NekRS to perform pioneering simulations of lead-lithium and water flows in the ITER fusion blanket system. This work introduces magnetohydrodynamics (MHD) capabilities to NekRS, an advancement enabling realistic modeling of fusion blankets under prototypical conditions.

CHALLENGE Fusion energy holds promise as a reliable and nearly limitless energy source capable of supporting long-term energy security and prosperity. The recent demonstration of ignition at the National Ignition Facility represents a watershed moment for the field. However, the practical deployment of fusion concepts remains a significant hurdle due to unresolved technological challenges, including material degradation under extreme radiation and thermal conditions, the complexities of tritium breeding and fuel cycle closure, and the efficient extraction and conversion of thermal energy into electricity. Modeling and simulation are poised to play a pivotal role in the design, optimization, and qualification of modern fusion reactor concepts, but this task remains formidable because of the inherent multiphysics, spanning wide spatial and temporal scales.

Fusion breeding blankets are key components in future fusion devices, serving to harness and manage the energy produced during fusion reactions. They facilitate heat transfer and thermal insulation, convert heat into usable energy, protect structural components from neutron irradiation, and breed tritium. The principal computational challenge in their design is time advancement of the governing equations for incompressible MHD, which this work aims to overcome.

Before neurons can be reconstructed in 3D, the 2D profiles of objects must be aligned between neighboring images in an image stack. Image misalignment can occur when tissue samples are cut into thin sections or during imaging on the electron microscope. The Feabas application (developed by collaborators at Harvard) uses template- and feature-matching techniques to optimize image transformations and align the 2D image content between sections.

APPROACH The Argonne-developed NekRS is run on supercomputing resources, including Aurora, to generate simulations. NekRS is a high-performance, open-source solver for incompressible and low-Mach number flows, heat transfer, and combustion, optimized for turbulent flows in complex geometries.

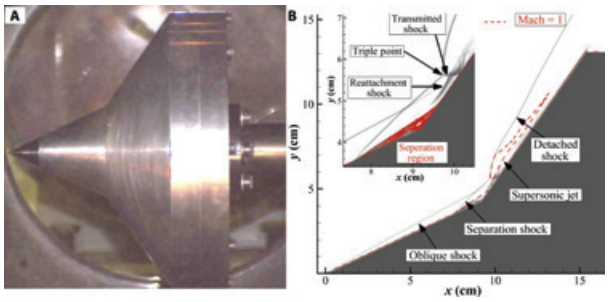
RESULTS With Aurora, the researchers performed the largest high-order simulation of a Water-Cooled Lithium Lead Test Blanket Module ever generated, marking the first exascale demonstration of scalable high-order MHD. The simulations incorporate MHD with detailed heat and fluid flow modeling in both the lead-lithium and water-cooling systems.

IMPACT These simulations significantly advance the analysis of fusion systems and are already being used to benchmark and refine predictions. Their significance will continue to grow as fusion transitions from a physics-centric challenge to an engineering-driven endeavor.

PUBLICATIONS
Min, M., Y.-H. Lan, P. Fischer, E. Merzari, T. Nguyen, Y. Guo, H. Yuan, P. Shriwise, S. Kerkemeier, A. Davis, A. Dubas, R. Eardley-Brunt, N. Bhatia, R. Akers, V. Mateevitsi, K. Row, and R. Balakrishnan. "Exascale MHD Simulations for Liquid Metal Fusion Blankets," *The International Journal of High Performance Computing and Applications* (preprint), SAGE Publications.

First-Principles Simulation of Hypersonic Flight

PI Maninder Grover, University of Dayton Research Institute and Air Force Research Laboratory
AWARD INCITE
SYSTEM Theta



(A) Test article in the CUBRC test facility before the experiment. (B) Numerical shadowgraph obtained from the DMS calculation, highlighting key flow features and the sonic line (Mach = 1) on the flow field. The inset in (B) provides a magnified view of the biconic junction. *Image: Sci. Adv. 11, eads2147 (2025)*

Hypersonic flight—the ability to fly at five times the speed of sound—has become a research priority for its potential to transform national security, aviation, and space exploration. At the core of this research is the need for accurate prediction of the extreme environments surrounding hypersonic vehicles, where strong shock waves, intense heating, and chemical reactions can impact performance and integrity. To advance this goal, researchers from the University of Dayton Research Institute and the Air Force Research Laboratory are using ALCF computing resources to perform large-scale molecular simulations that capture the complex interactions driving hypersonic flow.

CHALLENGE Hypersonic aerothermodynamics involve tightly coupled phenomena, including shock waves, molecular vibrations, and reactive chemistry. Experiments remain crucial but are limited in resolution and accessibility. While CFD numerical simulations provide a powerful way to study these flows, accurately modeling detailed mechanisms for high-Mach conditions requires molecular-level calculations and massive computational resources. Benchmark geometries, such as the canonical double cone test case, are particularly difficult, producing flow patterns highly sensitive to surface geometry and freestream conditions.

APPROACH To address these challenges, the team employed the massively parallel SPARTA code, a particle-based method that captures molecular-level dynamics and chemistry. For the double cone test case, which simulates reactive oxygen flow over sharp-angled geometry, they used the direct molecular simulation (DMS) method, an ab initio variant of the direct simulation Monte Carlo (DSMC) approach, to incorporate quantum-accurate collision dynamics. In a complementary study, the team applied DMS-tuned DSMC to shock tunnel experiments in pure oxygen at Texas A&M University, enabling direct comparisons between simulation results and spectroscopy data. Working with ALCF staff, the researchers optimized throughput on the facility's systems, generated detailed scientific visualizations, and began preparing their codes for GPU-powered exascale supercomputers.

RESULTS In the double cone study, the team carried out the first quantum-mechanically guided simulation of a hypersonic ground test. Their work revealed molecular-level flow details and identified regions where the gas deviated from Boltzmann energy distributions, providing unprecedented insight into nonequilibrium effects. In the shock study, the researchers used DSMC to reproduce coherent anti-Stokes Raman spectroscopy (CARS) measurements at Texas A&M, achieving one of the first direct, apples-to-apples comparisons between numerical simulations and high-fidelity spectroscopy. These advances reduced uncertainty in oxygen reaction rates that are central to predicting hypersonic flows in air.

IMPACT By combining quantum chemistry with particle-based flow models, the team is enhancing predictive capabilities for hypersonic environments. Their work is helping advance the development of next-generation hypersonic vehicles by providing benchmarks for validating fluid dynamics codes, informing thermal protection design, and refining reaction rate models.

PUBLICATIONS

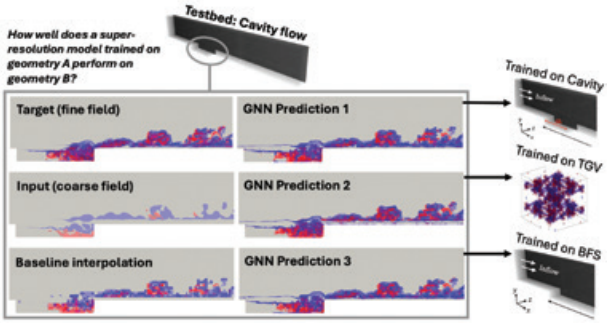
Grover, M. S., P. Valentini, and N. Bisek. "Ab Initio Investigation of a Hypersonic Double Cone Experiment," *Science Advances* (February 2025), AAAS. https://doi.org/10.1126/sciadv.ads2147

Grover, M. S., et al. "Direct Simulation Monte Carlo Investigation of a Hypervelocity Mach Stem Experiment," *Physics of Fluids* (preprint), AIP Publishing.

Engineering | Simulation, Learning

Mesh-Based Super-Resolution of Fluid Flows with Multiscale Graph Neural Networks

PI Shivam Barwey, Pinaki Pal, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



Visualizations of vorticity contours for the cavity flow at Reynolds number 1600 show examples of GNN fine-scale reconstruction and geometry extrapolation capability. *Image: Shivam Barwey, Argonne National Laboratory*

Capturing the complex behavior of turbulent flows is essential for advancing applications in energy, propulsion, power generation, and manufacturing, but simulating these flows at high resolution can be extremely costly with traditional methods. To address this challenge, researchers at Argonne National Laboratory are using ALCF computing resources to develop advanced AI models based on graph neural networks (GNN) that enhance the accuracy and efficiency of fluid flow simulations.

CHALLENGE Turbulent flows span a wide range of spatial and temporal scales, making them challenging to simulate accurately and efficiently. High-resolution simulations that resolve this multiscale behavior require significant computing power, particularly for problems with complex geometries. Traditional coarse-grid models reduce computational cost but can miss critical flow features, while conventional super-resolution methods are incompatible with unstructured meshes in three spatial dimensions used in realistic scientific applications. Bridging this gap requires a framework that is both data-driven and mesh-aware, providing the capability to resolve localized flow structures across varied geometries.

APPROACH Using the ALCF's Polaris supercomputer, the researchers developed a multiscale GNN architecture designed to achieve super-resolution on unstructured meshes. Instead of processing entire flow fields, the model focuses on small, localized patches of simulation data, learning to reconstruct more detailed velocity fields. The network uses a combination of message-passing layers and an unpooling layer to map information between coarse and fine scales. To train and validate the model, the researchers generated high-fidelity simulation data using Argonne's GPU-accelerated NekRS flow solver on Polaris. GNN training was distributed across multiple GPUs to efficiently process large datasets from canonical and complex flows, including Taylor–Green vortex, backward-facing step, and cavity configurations.

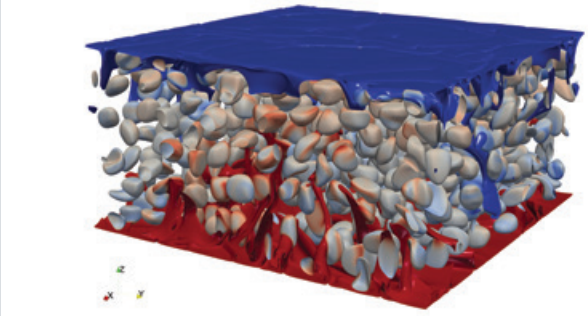
RESULTS The team's GNN model accurately reconstructed high-resolution flow fields from coarse inputs across all test cases, outperforming traditional interpolation methods. Key flow features and energy spectra were preserved, with model performance improving when more information from neighboring elements was included. At higher Reynolds numbers, reconstruction errors increased as expected due to greater turbulence complexity, but the model maintained consistent accuracy trends across configurations. Importantly, models trained on one geometry (e.g., backward-facing step) generalized well to others (e.g., cavity) without retraining, demonstrating strong extrapolation capabilities.

IMPACT This work introduces a flexible and scalable AI-driven framework for enhancing fluid flow simulations. By capturing detailed flow behavior without the full cost of direct numerical simulation, the team's method can accelerate modeling and design in areas such as aerospace engineering and combustion engines. The team's research also highlights how AI and high-performance computing can be integrated to advance multiscale modeling in computational fluid dynamics.

PUBLICATIONS
Barwey, S., P. Pal, S. Patel, R. Balin, B. Lusch, V. Vishwanath, R. Maulik, and R. Balakrishnan. "Mesh-Based Super-Resolution of Fluid Flows with Multiscale Graph Neural Networks," *Computer Methods in Applied Mechanics and Engineering* (August 2025), Elsevier. https://doi.org/10.1016/j.cma.2025.118072.

Rayleigh-Bénard Turbulent Convection in Emulsions

PI Parisa Mirbod, University of Illinois Chicago
AWARD INCITE
SYSTEM Polaris



Schematic of the 3D turbulent Rayleigh–Bénard convection cell with a Cartesian coordinate system. Heating is applied at the bottom wall (red), while the top wall is cooled (blue). The image shows instantaneous iso-surfaces of temperature and dispersed droplets at a short time ($\Delta t = 0.4$) after introducing droplets with a volume fraction of $\Phi = 0.2$ into the initially single-phase flow. *Image: Bilondi et al., J. Fluid Mech. 2024*

High-fidelity simulations reveal how heat affects emulsion dynamics, offering insights for energy-efficient industrial processes. By capturing the complex interactions between droplets in turbulent, heated flows, this work advances our understanding of multiphase fluid behavior in practical engineering environments.

CHALLENGE Emulsions, a mixture of immiscible liquids like oil and water, are central to applications in pharmaceuticals, food, cosmetics, and oil extraction. However, under heat, emulsions exhibit complex and poorly understood behaviors due to droplet deformation, movement, and interaction in turbulent flows. Traditional models fall short in capturing the interfacial physics and heat transfer dynamics essential to optimizing these systems.

APPROACH To address this, researchers used FluTAS (Fluid Transport Accelerated Solver), a high-resolution, interface-resolved direct numerical simulation (DNS) tool, to model incompressible multiphase flows with heat transfer. Leveraging CPU parallelization (MPI/OpenMP) and GPU acceleration (OpenACC), the team simulated thirteen scenarios at a grid resolution of 1024x1024x512. Simulations varied volume fraction, viscosity ratio, and thermal diffusivity ratio to study their effects on heat transfer and droplet behavior. Single-phase baseline runs were also conducted for direct comparison.

RESULTS The simulations revealed that deformable droplets enhance local mixing by promoting energy transfer to smaller turbulent scales. Heat transfer was consistently higher in multiphase simulations than in the single-phase case. Droplets tended to migrate away from wall boundaries, resulting in a dispersed phase with higher thermal diffusivity. The study also confirmed two distinct droplet scaling regimes, dominated by either coalescence or breakup, depending on size. These findings provide a deeper understanding of how heat and flow interact within emulsions.

IMPACT This research fills a longstanding knowledge gap in multiphase thermal fluid dynamics by showing how emulsion microstructure affects heat transfer. The insights gained could improve pharmaceutical drug delivery through controlled release, reduce energy consumption in oil-water separation, and enhance food and cosmetic product stability. With the support of the ALCF Catalyst team, the project optimized FluTAS for next-generation high-performance computing systems like Aurora, laying the groundwork for even more advanced simulations in the future.

PUBLICATIONS

Bilondi, A. M., H. Brynteson, P. Mirbod, M. Papka, L. Brandt, and N. Scapin. “Effects of Employing Liquid–Liquid Emulsions on Heat Transfer Within a Turbulent Rayleigh–Bénard Convection,” *76th Annual Meeting of the APS Division of Fluid Dynamics* (November 2023), Washington, DC, APS.

Bilondi, A. M., N. Scapin, L. Brandt, and P. Mirbod. “Heat Transfer Enhancement in Turbulent Rayleigh-Bénard Convection with Liquid–Liquid Emulsions,” *76th Annual Meeting of the Division of Fluid Dynamics* (November 2023), Washington, DC, APS.

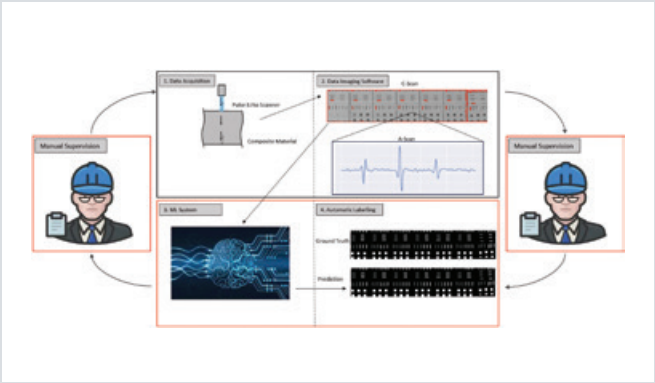
Bilondi, A. M., N. Scapin, L. Brandt, and P. Mirbod. “Multi-Phase Turbulent Rayleigh–Bénard Convection with Bubbles,” *75th Annual Meeting of the Division of Fluid Dynamics* (November 2022), Indianapolis, IN, APS.

Bilondi, A. M., N. Scapin, L. Brandt, and P. Mirbod. “Turbulent Convection in Emulsions: The Rayleigh–Bénard Configuration,” *Journal of Fluid Mechanics* (November 2024), Cambridge University Press. https://doi.org/10.1017/jfm.2024.765

Engineering | Learning

Tool for AI-Enabled Automated Non-Destructive Inspection of Composite Aerostructures Manufacturing

PI Rajkumar Kettimuthu, Ian Foster, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris, Sophia



Automatic inspection of composite aerospace components using AI. *Image: Austin Yunker, Argonne National Laboratory*

In aerospace manufacturing, composite materials such as carbon fiber reinforced polymers are critical for lightweight, high-performance structures. To ensure structural integrity, every part must undergo rigorous ultrasonic inspection. This process can significantly slow down manufacturing flow and increase factory energy use. A cross-sector team from Spirit AeroSystems, Argonne National Laboratory, Northern Illinois University, and TRI Austin collaborated to develop an AI-assisted tool that aims to accelerate the inspection process while maintaining high safety and quality standards.

CHALLENGE Traditional AI approaches to defect detection rely on large labeled datasets. In aerospace inspection, however, defects are rare and labeling is inconsistent. Moreover, defect patterns are unpredictable and can appear anywhere in the part, making supervised learning approaches difficult to apply at scale. The team needed to build a robust system capable of performing reliably on complex, real-world production data.

APPROACH Using ALCF's Polaris and Sophia systems, the team developed an anomaly detection system trained on non-defective ultrasonic scan data from Spirit's archive. The model learns the characteristics of normal signals and flags outliers as potential defects. To enhance performance and interpretability, the team embedded domain-specific knowledge such as signal amplitude and material thickness. The system outputs a configurable anomaly score that helps inspectors focus on areas of interest.

RESULTS The AI model was integrated into a standalone software package and evaluated alongside current workflows at Spirit. In pilot testing, it reduced ultrasonic inspection evaluation time by up to 24 percent, resulting in an estimated 8.8 percent reduction in inspection-related manufacturing flow time and about 3 percent energy savings per aircraft. The system was also tested on additional composite material systems and configurations, demonstrating its adaptability.

IMPACT By augmenting manual inspection with AI-guided analysis, the tool improves inspection efficiency and supports energy-conscious, high-throughput composite manufacturing. Although not yet deployed for full production use, the project establishes a scalable foundation for applying AI to non-destructive inspection across the aerospace industry.

PUBLICATIONS

Yunker, A, R. Kettimuthu, and Z. Kral. "DOC-DICAM: Domain Aware One Class Defect Identification in Composite Aerostructure Material," *2024 International Conference on Machine Learning and Applications (ICMLA)* (December 2024), Miami, FL, IEEE. https://doi.org/10.1109/ICMLA61862.2024.00157

Yunker, A., R. Kettimuthu, and Z. Kral. "OC-DICAM: One Class Defect Identification in Composite Aerostructure Material," *ASME 2024 Aerospace Structures, Structural Dynamics, and Materials Conference* (April 2024), Renton, WA, ASME. https://doi.org/10.1115/SSDM2024-121466

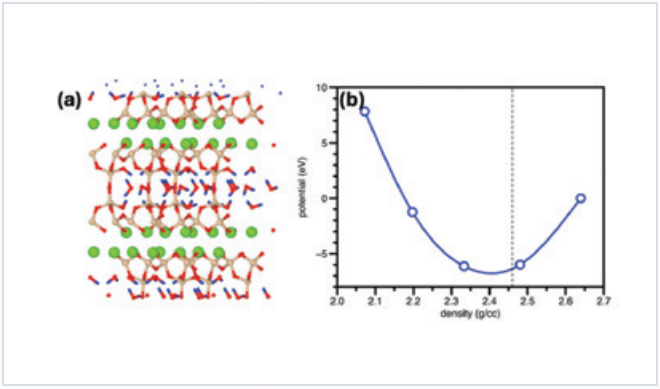
Yunker, A., Z. Kral, and R. Kettimuthu. "Whose Plane to Take? A Novel Approach to Identify the Optimal Model for Defect Detection in Composite Aerostructure Material," *ASME 2025 Aerospace Structures, Structural Dynamics, and Materials Conference* (May 2025), Houston, TX, ASME. https://doi.org/10.1115/SSDM2025-152304

Yunker, A., R. Lake, R. Kettimuthu, and Z. Kral. "Comparative Study on Deep Learning Methods for Defect Identification and Classification in Composite Aerostructure Material," *2023 50th Annual Review of Progress in Quantitative Nondestructive Evaluation* (July 2023), Austin, TX, ASME. https://doi.org/10.1115/QNDE2023-108602

Yunker, A., R. Lake, R. Kettimuthu, and Z. Kral. "Evaluating Model Robustness for Defect Identification and Classification in a Composite Aerostructure Material," *Journal of Nondestructive Evaluation, Diagnostics and Prognostics of Engineering Systems* (July 2024), ASME. https://doi.org/10.1115/1.4065474

Allegro-FM: Toward an Equivariant Foundation Model for Exascale Molecular Dynamics Simulations

PI Aiichiro Nakano, University of Southern California
AWARD INCITE
SYSTEM Aurora, Sophia



Snapshot of the tobermorite 11 Å (T11) structure (a). Equation of state of T11 using Allegro-FM, with the dotted-line showing an experimental density of T11 at 2.46 g/cm (b). *Image: Nomura et al., J. Phys. Chem. Lett. 2025*

Foundation models (FM) represent a paradigm shift in AI, where a single universal model acquires sufficient robustness and generalizability to enable diverse, out-of-distribution downstream tasks. Leveraging ALCF supercomputing resources, a research team led by the University of Southern California has introduced Allegro-FM to perform exascale molecular dynamics simulations to accurately describe a wide variety of material properties and processes with a single pretrained model. Allegro-FM covers 89 elements in the periodic table and can be applied to a wide range of tasks, including structural correlations, reaction kinetics, mechanical strengths, fracture, and solid/liquid dissolution, exhibiting emergent capabilities for which the model was not trained.

CHALLENGE Fracture mechanics of silicate materials are studied for technological importance as well as to answer fundamental scientific questions. To examine the applicability of Allegro-FM to fracture behavior, the researchers performed a tensile test on a tobermorite 11Å (T11) crystal, a calcium silicate mineral and the key ingredient for fire-resilient concrete. The researchers also simulated the carbonation process—important for carbon sequestration—for a nanoparticle of T11 placed in a mixture of H_2O and CO_2 molecules.

APPROACH The researchers have developed an exascalable universal machine learning interatomic potential by leveraging an E(3) equivariant network architecture and a set of large-scale organic and inorganic materials data sets merged by the Total Energy Alignment framework.

RESULTS In a paper featured on the cover of the *The Journal of Physical Chemistry Letters*, the researchers demonstrated Allegro-FM's excellent agreement with high-level quantum chemistry theories for describing structural, mechanical, and thermodynamic properties, as well as its emergent capabilities for structural correlations, reaction kinetics, mechanical strengths, fracture, and solid/liquid dissolution, for which the model has not been trained. The researchers also demonstrated the robust predictability and generalizability of Allegro-FM for chemical reactions using Transition1x and confirmed the scalability of Allegro-FM on the ALCF's exascale Aurora supercomputer, with a 97.5 percent perfect speedup for 4.08 billion atoms across 4,096 graphics processing units. The obtained molecular dynamics trajectory for the carbonation process simulation was robust, with no system failures or spurious events observed.

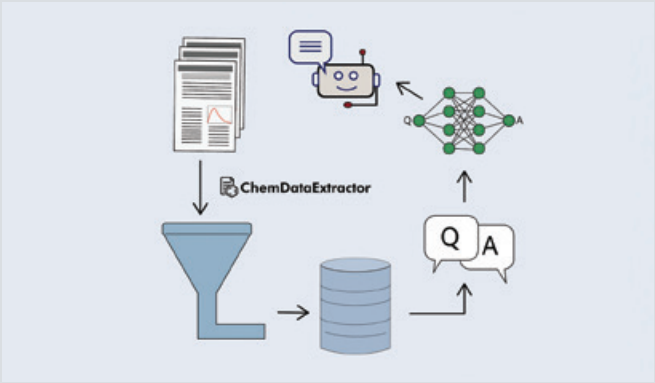
IMPACT This work demonstrates the potential of FM use to accelerate and advance novel materials design and discovery via large-scale atomistic simulations.

PUBLICATIONS
K. Nomura, S. Hattori, S. Ohmura, I. Kanemasu, K. Shimamura, N. Dasgupta, A. Nakano, R. K. Kalia and P. Vashishta. "Allegro-FM: Towards an Equivariant Foundation Model for Exascale Molecular Dynamics Simulations," *The Journal of Chemical Physics Letters* (June 2025), ACS Publications. https://doi.org/10.1021/acs.jpclett.5c00605

Materials Science | Simulation, Data, Learning

AI Tools for Data-Driven Molecular Engineering

PI Jacqueline Cole, University of Cambridge
AWARD ALCC, Director's Discretionary
SYSTEM Polaris



A schematic showing how scientific literature is mined using ChemDataExtractor to build materials databases. These databases are then employed to generate QA pairs, which are used to fine-tune efficient, materials-domain-specific language models. *Image: Mayank Shreshtha, University of Cambridge*

As the volume of scientific literature grows, materials scientists face the challenge of extracting the most relevant information quickly enough to guide experiments and discover new materials for a wide range of applications. With support from ALCF computing resources, researchers from the University of Cambridge are developing AI tools that can mine millions of research papers, build structured materials databases, and train specialized language models to accelerate discovery.

CHALLENGE Large language models (LLMs) show promise for answering scientific questions, but training them from scratch on domain-specific data requires enormous computing resources. Even fine-tuning general-purpose models can be less effective if the training data lack the precision needed for specialized scientific tasks. For many areas of materials science, there are also gaps in the availability of structured, machine-readable property data to enable such training.

APPROACH To overcome these barriers, researchers are leveraging ALCF resources to combine automated text mining with innovative AI training approaches at scale. Using their ChemDataExtractor toolkit, the team auto-generates high-quality datasets from scientific articles and builds structured materials databases, which are then fed into small, efficient language models, bypassing the need for costly pretraining of LLMs. ALCF supercomputers have been central to developing these workflows, enabling large-scale data extraction, algorithm development, and model training.

RESULTS In one study, the team transformed a photovoltaic materials database into hundreds of thousands of structured question-and-answer (QA) pairs, achieving up to 20 percent higher accuracy on solar cell–related tasks than models trained on general-purpose data. They applied the same strategy to other domains, including building a stress–strain property database with more than 720,000 records to train MechBERT, a model that outperforms standard tools in predicting material behavior under stress. Optoelectronics-related work showed that computational costs for domain-adaptive pretraining can be reduced by more than 80 percent compared to traditional methods, while maintaining or improving accuracy. Most recently, the group generated nearly 100,000 QA pairs from a thermoelectric materials database, demonstrating that a small BERT model fine-tuned on this dataset can outperform models trained on either source alone.

IMPACT By turning unstructured literature into targeted, machine-readable knowledge, the team is making AI tools more accessible to the research community. Their efforts illustrate a scalable approach to building smaller, faster language models for specialized materials science domains and help scientists tailor models to their own areas, design better experiments, interpret results, and accelerate the path from discovery to application.

PUBLICATIONS

Kumar, P., S. Kabra, and J. M. Cole. "A Database of Stress-Strain Properties Auto-generated from the Scientific Literature Using ChemDataExtractor," *Scientific Data* (November 2024), Springer Nature. https://doi.org/10.1038/s41597-024-03979-6

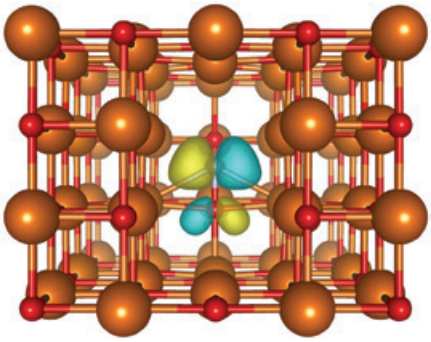
Kumar, P., S. Kabra, and J. M. Cole. "MechBERT: Language Models for Extracting Chemical and Property Relationships about Mechanical Stress and Strain," *Journal of Chemical Information and Modeling* (January 2025), American Chemical Society. https://doi.org/10.1021/acs.jcim.4c00857

Li, Z., and J. M. Cole. "Auto-generating Question-Answering Datasets with Domain-Specific Knowledge for Language Models in Scientific Tasks," *Digital Discovery* (February 2025), Royal Society of Chemistry. https://doi.org/10.1039/D4DD00307A

Sierepeklis, O., and J. M. Cole. "Autogenerating a Domain-Specific Question-Answering Data Set from a Thermoelectric Materials Database to Enable High-Performing BERT Models," *Journal of Chemical Information and Modeling* (August 2025), American Chemical Society. https://doi.org/10.1021/acs.jcim.5c00840

Large-Scale Simulations of Materials for Energy and Quantum Information Science

PI Giulia Galli, Argonne National Laboratory and University of Chicago
AWARD ALCC
SYSTEM Polaris



Magnesium atoms (orange) and oxygen atoms (red) surround the nitrogen-vacancy center in magnesium oxide, shown by a transparent representation of a nitrogen atom under the missing magnesium atom. The yellow and blue spots show how electrons localize around the vacancy. *Image: Giulia Galli, University of Chicago/ Argonne National Laboratory*

Quantum technologies have the potential to revolutionize computing, communication, and sensing, but their advancement requires the discovery of stable, scalable materials for quantum bits, or qubits. Researchers from Argonne National Laboratory, the University of Chicago, and Linköping University have identified a promising defect in magnesium oxide that could serve as a new spin qubit. Using DOE supercomputers, the team carried out large-scale calculations to predict the defect's electronic and optical properties, providing a foundation for experimental synthesis and characterization.

CHALLENGE Qubits can take many forms, including spin defects, which are atomic-scale irregularities in a material that can host quantum information. Identifying new host materials with long coherence times and favorable electronic, spin, and optical properties is essential, but the space of possible candidates is vast. Countless defects can exist in a single material, and predicting their properties with the accuracy needed for quantum applications requires significant computational power.

APPROACH Building on predictions in prior work that oxides such as magnesium oxide could host spin defects with long coherence times, the team first carried out high-throughput computational screening of nearly 3,000 candidate defects. Automated filters narrowed the list to a manageable set, and the researchers then applied advanced electronic structure methods to the most promising candidate, the nitrogen-vacancy center. To get a more precise characterization of this promising defect, they conducted detailed simulations with open-source codes from the Midwest Integrated Center for Computational Materials, using ALCF's Polaris and NERSC's Perlmutter systems. This analysis predicted key properties relevant for quantum information use, including electronic structure, optical properties, zero-field splitting, and hyperfine interactions, as well as the interaction between the defect and the surrounding magnesium oxide host.

RESULTS The simulations confirmed that the nitrogen-vacancy center in magnesium oxide possesses stable triplet ground and excited states, and could support spin-dependent optical readout, indicating potential for hybrid quantum applications such as sensing and single-qubit control. The study also explored design strategies to mitigate vibronic coupling, a factor that can reduce performance, pointing to practical routes for improving defect stability. The team's computational framework offers a scalable approach to discovering and characterizing new spin defects in a wide range of host materials.

IMPACT By combining high-throughput screening with advanced electronic structure calculations, the team has demonstrated a powerful protocol for accelerating the search for viable quantum defects. Their results highlight magnesium oxide as a new candidate host and establish a methodology for evaluating defects across other oxides and materials. This integrated computational strategy will help guide experimental efforts and advance the development of materials for emerging quantum technologies.

PUBLICATIONS
Somjit, V., J. Davidsson, Y. Jin, and G. Galli. "An NV− Center in Magnesium Oxide as a Spin Qubit for Hybrid Quantum Technologies," *npj Computational Materials* (March 2025), Springer Nature. https://doi.org/10.1038/s41524-025-01558-w

Materials Science | Simulation, Learning

Multiscale Light-Matter Dynamics in Quantum Materials: From Electrons to Topological Superlattices

PI Aiichiro Nakano, University of Southern California
AWARD INCITE, Aurora ESP
SYSTEM Aurora

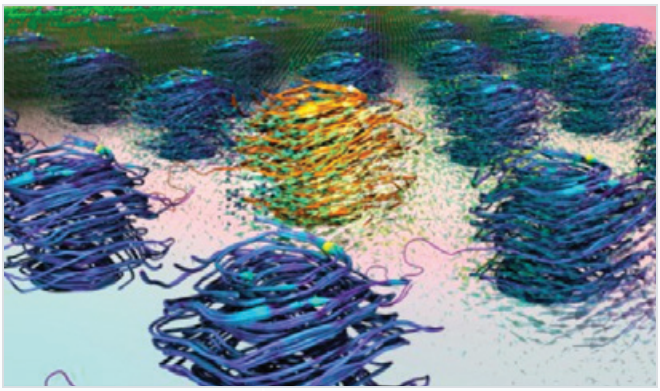


Photo-switching of a ferroelectric skyrmion superlattice in $PbTiO_3$. *Image: ALCF Visualization and Data Analytics Team*

Light-matter dynamics (LMD) in topological quantum materials enables ultralow-power, ultrafast devices with great promise to meet the demands of ubiquitous AI. A key challenge is to simulate multiple field and particle equations for light, electrons, and atoms over vast spatiotemporal scales on exascale computers with increased heterogeneity and low-precision focus. A team of researchers led by the University of Southern California achieved the first end-to-end exa-deployed multiscale LMD simulation and machine learning for light, electrons, and atoms, achieving 152x and 3,780x improvements in time-to-solution, establishing a new computational paradigm.

CHALLENGE LMD presents researchers with problems that encompass multiphysics of vastly different computational characteristics at vastly different spatiotemporal scales. This work aims to solve the multiscale/multiphysics/heterogeneity/low-precision challenge.

APPROACH The researchers introduce a paradigm shift that harnesses hardware heterogeneity and low-precision arithmetic. Divide-conquer-recombine (DVR) algorithms partition a given problem into spatial and physical subproblems that are mapped onto optimally selected hardware units; metamodel-space algebra minimizes communications and precision requirements.

RESULTS Using 60,000 Aurora GPUs, multiscale LMD software respectively performed 152 and 3,780 times faster than the state-of-the-art for 15.4 million-electron and 1.23 trillion-atom $PbTiO_3$ materials, the former achieving 1.87 exaflops per second. This enabled the first study of light-induced switching of topological superlattices for future ferroelectric topotronics.

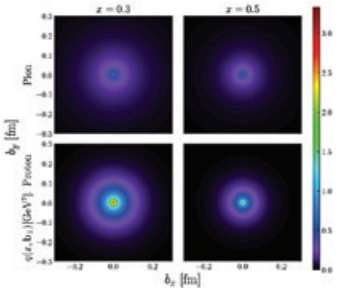
IMPACT This work unleashed the power of AI-enhanced multiscale/multiphysics simulations. The new DVR/metamodel-space-algebra paradigm offers viable algorithm-hardware codesign pathways for a wide range of problems in the coming post-exascale era and is metascalable at the emerging nexus of post-exascale high-performance computing, AI, and quantum computing. Aurora software developed for this work is informing x-ray free-electron laser experiments; integrated computational-experimental studies of this kind will be critical future topotronics and other advanced electronics.

PUBLICATIONS
Kazakh, T. M., T. Linker, Y. Luo, N. Piroozan, J. Pennycook, N. Kumar, A. Musaelian, A. Johansson, B. Kozinsky, R. K. Kalia, P. Vashishta, F. Shimojo, S. Hattori, K. Nomura, and A. Nakano. "Multiscale Light-Matter Dynamics in Quantum Materials: From Electrons to Topological Superlattices," *arXiv* (preprint). https://doi.org/10.48550/arXiv.2509.00966

Physics | Simulation

3D Imaging of Strong Interaction Nambu-Goldstone Bosons

PI Yong Zhao, Argonne National Laboratory
AWARD INCITE
SYSTEM Aurora, Polaris



Distribution of up-type quarks in the transverse plane (perpendicular to the direction of motion) of a fast-moving pion (top) or proton (bottom) for different longitudinal momentum fractions '*x*' (fraction of pion momentum carried by the quark in the direction of pion motion) (*left, right*), illustrating how the particle "shrinks" as one quark carries more of the motion. *Image: Yong Zhao, Argonne National Laboratory*

A fundamental question in nuclear physics is understanding how visible matter forms from elementary particles like quarks and gluons, which can be profoundly addressed by quantifying the multi-dimensional structure of nucleons and mesons. Using ALCF resources and advanced theoretical frameworks, a research team has generated a detailed 3D quark image of the pion, the lightest meson responsible for the spontaneous breaking of chiral symmetry in the strong force.

CHALLENGE Scientists have long been interested in finding and understanding the distribution of quarks within composite particles held together by the strong nuclear force. For the lightest of these particles, the pion, there are no experimental results to rely on, so scientists are using large-scale computation to reveal their internal structure. The research team utilized the Polaris supercomputer to simulate the fundamental theory of the strong force and unveiled a detailed 3D image of the pion, the lightest particle composed of a quark and an antiquark.

APPROACH Using ALCF resources, the researchers calculated the quark generalized parton distribution (GPD) of the pion to create a detailed 3D image of it. This calculation is based on the lattice formulation of quantum chromodynamics, the fundamental theory of the strong force, and employs advanced techniques from the Large Momentum Effective Theory framework. The pion GPD is determined with controlled systematic uncertainties across different quark longitudinal momentum fractions x (fraction of quark momentum along the direction of the pion motion) and in the transverse plane (perpendicular to the pion motion).

RESULTS The results reveal that the transverse size of the pion decreases as x increases (a pattern also seen in the proton) and that the effective size of the pion is smaller than that of the proton at moderate x values.

IMPACT Because there currently are no experimental measurements of the pion GPD, the obtained theoretical findings will provide valuable guidance and support for experiments such as those at Jefferson Lab and the future Electron-Ion Collider at Brookhaven National Laboratory.

PUBLICATIONS
Ding, H.-T., X. Gao, S. Mukherjee, P. Petreczky, Q. Shi, S. Syritsyn, and Y. Zhao. "Three-Dimensional Imaging of Pion Using Lattice QCD: Generalized Parton Distributions," *Journal of High Energy Physics* (February 2025), Springer Nature. https://doi.org/10.1007/JHEP02(2025)056

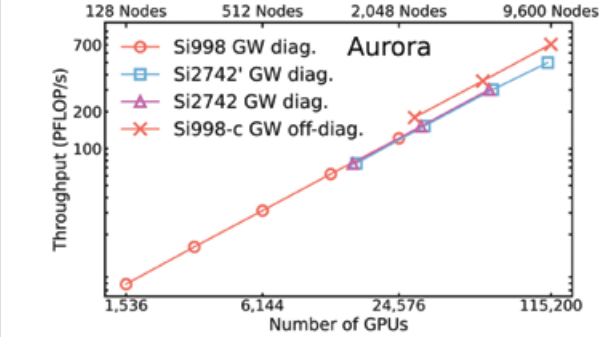
Physics | Simulation, Data

Advancing Quantum Many-Body GW Calculations on Exascale Supercomputing Platforms

PI Zhenglu Li, University of Southern California
Mauro Del Ben, Lawrence Berkeley National Laboratory

AWARD INCITE, Aurora ESP

SYSTEM Aurora



Throughput performance of the BerkeleyGW's sigma module on Aurora, showing scalability to nearly the whole machine. *Image: Benran Zhang, University of Southern California*

Advanced ab initio materials simulations face growing challenges as increasing systems and phenomena complexity require higher accuracy, thereby increasing computational demands. Quantum many-body GW methods represent the state-of-the-art for treating electronic excited states and couplings but are often hindered by complexity. This project aims to present innovative implementations of advanced GW methods within the BerkeleyGW package, enabling large-scale simulations on exascale supercomputers, including the ALCF's Aurora system. The team's work has been recognized as a finalist for the 2025 Gordon Bell Prize.

CHALLENGE Quantum materials research has entered a new era where a broad range of emerging and increasingly heterogeneous materials systems and various many-body phenomena are becoming central topics of study. Predictive first-principles methods are necessary for their investigation, as well as to design next-generation electronic, optical, and quantum devices.

The BerkeleyGW software package offers a range of widely adopted first-principles methodologies for electronic and optical excitations based on the GW approximation. To apply these accurate quantum many-body methods to the investigation of complex quantum materials systems such as solid-state defects and moiré superlattices, their computational bottlenecks must be addressed.

APPROACH The researchers have implemented significant methodological and algorithmic innovations in BerkeleyGW, including portability across architectures and kernel optimizations with hardware-optimized programming languages. This enabled scaling to full-system runs on Aurora, achieving excellent time-to-solution, a high fraction of peak performance, and over 1 exaflops of sustained performance.

RESULTS The researchers' approach has demonstrated exceptional versatility for complex heterogeneous systems with up to 17,574 atoms and has achieved true performance portability across GPU architectures. Results indicate excellent strong and weak scaling to thousands of nodes. This work represents a breakthrough in the use of exascale computing for quantum materials simulations, delivering unprecedented predictive capabilities for the rational design of future quantum technologies.

IMPACT The advancements implemented in this work enable at-scale study of not only quasiparticle excited-state properties but also important electron-phonon coupling phenomena, with the latter being critical to the transport, optical absorption, and decoherence and lifetimes of quantum states. The portable, efficient utilization of resources, and the ability to describe heterogeneous systems with the inclusion of quantum many-body interactions, establish a new vision for studying increasingly complex quantum materials, phenomena, and devices in the exascale era.

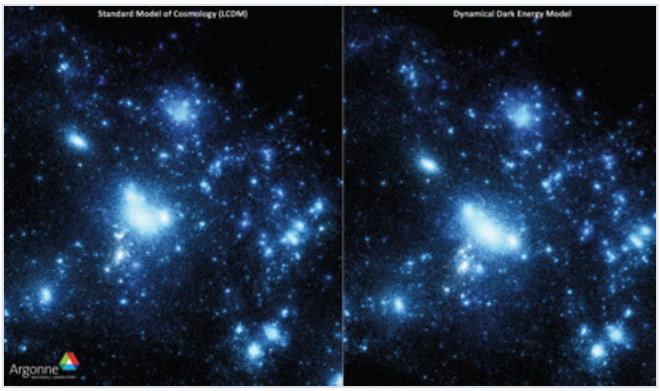
PUBLICATIONS

Zhang, B., D. Weinberg, C.-E. Hsu, A. R. Altman, Y. Shi, J. B. White III, D. Vigil-Fowler, S. G. Louie, J. R. Deslippe, F. H. da Jornada, Zhenglu Li, and M. Del Ben. "Advancing Quantum Many-Body GW Calculations on Exascale Supercomputing Platforms," *SC25: The International Conference for High Performance Computing, Networking, Storage, and Analysis* (November 2024), St. Louis, MO, IEEE. https://doi.org/10.48550/arXiv.2509.23018

Physics | Simulation, Data, Learning

Illuminating the Physics of Dark Energy with the Discovery Simulations

PI Katrin Heitmann, Argonne National Laboratory
AWARD Aurora ESP
SYSTEM Aurora



Visual comparison of a small region in the simulations (left: standard model of cosmology; right: dynamical dark energy model). The differences are subtle but still clearly visible at the substructure level. *Image: ALCF Visualization and Data Analytics Team and the HACC Collaboration*

Dark energy, the mysterious force thought to drive the accelerated expansion of the universe, has long been considered constant in the standard model of cosmology. Recent observations from the Dark Energy Spectroscopic Instrument (DESI) suggest that it may evolve over time, which could reshape our understanding of the cosmos. To help the scientific community investigate this possibility, Argonne researchers turned to ALCF's Aurora exascale system to simulate the growth of cosmic structure under competing dark energy models.

CHALLENGE DESI has created the most precise 3D map of the universe to date, spanning nearly 15 million galaxies and quasars. While its observations largely support the standard model of cosmology, growing tensions appear when the data are combined with other measurements from studies of the cosmic microwave background, supernovae, and weak gravitational lensing. The inconsistencies point to the possibility that dark energy may evolve with time rather than remain constant. To test this idea, researchers need extremely large, high-resolution simulations that can separate true signatures of evolving dark energy from effects introduced by data collection and analysis.

APPROACH Building on years of large-scale cosmology simulations carried out on DOE's leadership supercomputers, the Argonne team used the HACC (Hardware/Hybrid Accelerated Cosmology) code on Aurora to perform a pair of high-resolution N-body simulations with identical starting conditions: one assuming a constant cosmological constant and another allowing dark energy to vary over time. Aurora's immense processing power enabled the pair to achieve the fine resolution necessary to detect subtle structural differences, completing each simulation in days rather than the weeks required on earlier systems. The team also applied on-the-fly analysis methods to process data as it was generated, eliminating storage and postprocessing bottlenecks and providing rapid insights to refine simulations in real time.

RESULTS The team's simulations revealed clear structural contrasts between constant and evolving dark energy models, establishing a controlled benchmark for interpreting DESI data. By making these datasets publicly available, the Argonne team has provided the cosmology community with a valuable resource for testing analysis strategies and refining models at a pivotal moment as DESI observations continue to strengthen the case for evolving dark energy.

IMPACT Serving as a foundation for interpreting DESI data, the team's simulations are enabling researchers to explore the potential evolution of dark energy, a possibility that would represent a major shift in our understanding of the universe. Their work highlights the important role that exascale systems like Aurora can play in testing theoretical models against observations at unprecedented scale and speed, allowing scientists to respond rapidly to new results from DESI and other state-of-the-art cosmological surveys.

PUBLICATIONS

Beltz-Mohrmann, G. D., A. Pope, A. Alarcon, M. Buehlmann, N. Frontiere, A. P. Hearin, K. Heitmann, S. Ortega-Martinez, A. Pearl, E. Rangel, S. Rizzi, T. Uram, and E. Xhakaj. "Illuminating the Physics of Dark Energy with the Discovery Simulations," *The Open Journal of Astrophysics* (June 2025), Maynooth Academic Publishing. https://doi.org/10.33232/001c.140866

Physics | Simulation, Data

Pathfinding Integrated and Automatic Experimental Analyses for DIII-D Research

PI Mark Kostuk, General Atomics
AWARD ALCC, Director's Discretionary
SYSTEM Polaris

The experimental DIII-D National Fusion Facility is connected with the advanced computing resources at ALCF and NERSC via ESnet. *Image: ESnet*

At the DIII-D National Fusion Facility, researchers study plasmas under fusion-relevant conditions to advance the physics and technology needed for future fusion reactors. This work requires analyzing experimental data quickly and accurately, so scientists can make informed decisions during live operations. To accelerate this process, researchers from DIII-D and ALCF are leveraging integrated workflows and computing capabilities to link experiments directly to DOE supercomputers.

CHALLENGE Plasma behavior inside a tokamak is complex and constantly evolving, requiring accurate reconstructions and particle tracking to guide experiments and protect sensitive equipment. Traditional analysis can take hours, delaying feedback and limiting opportunities to adjust experimental parameters in real time. Faster turnaround requires infrastructure that moves data between facilities, runs compute-intensive simulations at scale, and delivers actionable results within minutes.

APPROACH The team pioneered on-demand remote analysis between DIII-D and ALCF, and is now advancing their work through a DOE Integrated Research Infrastructure (IRI) Pathfinder Project and an ALCC award of computing time at ALCF and NERSC. Central to their approach is a combination of advanced computing capabilities—on-demand queues, direct networking access to compute nodes, and Globus services—that enable fast, secure, and automated analysis. The Consistent Automatic Kinetic Equilibria (CAKE) workflow uses this infrastructure to launch high-fidelity plasma reconstructions in sync with DIII-D's plasma shots. The IonOrb workflow runs a GPU-accelerated particle-tracking code to simulate high-energy particle trajectories and identify where they deposit energy on the walls of the tokamak vessel. To return results quickly, IonOrb runs on 20 nodes of ALCF's Polaris (80 GPUs), far beyond the resources available at DIII-D. Both workflows automate data transfers between DIII-D and ALCF to enable remote processing, while on-demand queues provide immediate access to computing resources.

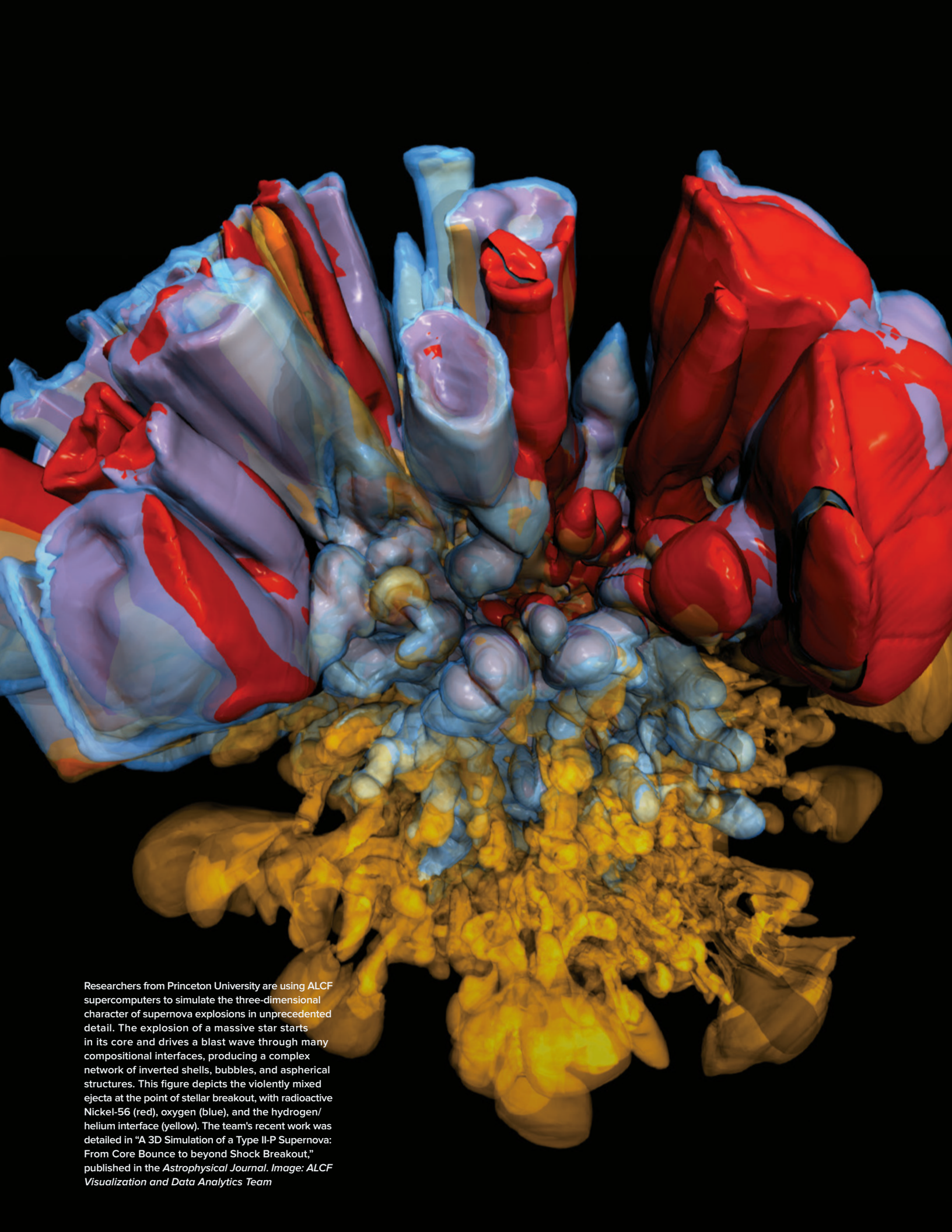
RESULTS The CAKE workflow now delivers high-fidelity kinetic equilibrium reconstructions closely aligned with DIII-D's time-intensive experiment cycle, allowing scientists to evaluate plasma performance during ongoing operations. IonOrb is providing power deposition information in minutes, enabling near-real-time analysis, and the identification of potential hot spots caused by the 20MW of neutral beam particle injectors. The team shared their advances at the SC24 conference, where researchers presented a paper at the XLOOP workshop and led a live demonstration of a fusion experiment workflow linking plasma shots to rapid analysis on DOE supercomputers.

IMPACT By integrating DIII-D experiments with DOE supercomputers, the team is advancing both the speed and fidelity of plasma analysis. This work serves as a model for cross-facility science under DOE's IRI program and lays the foundation for more compute-intensive analysis at DIII-D, with the potential to bring additional applications into the experiment-time analysis cycle.

PUBLICATIONS

Kostuk, M., T. Uram, T. Evans, D. Orlov, M. Papka, and D. Schissel. "Automatic Between-Pulse Analysis of DIII-D Experimental Data Performed Remotely on a Supercomputer at Argonne Leadership Computing Facility," *Fusion Science and Technology* (February 2018), Taylor & Francis. https://doi.org/10.1080/15361055.2017.1390388

Smith, S. P., Z. A. Xing, T. B. Amara, S. S. Denk, E. W. DeShazer, O. Meneghini, T. Neiser, et al. "Expediting Higher Fidelity Plasma State Reconstructions for the DIII-D National Fusion Facility Using Leadership Class Computing Resources," *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2024), IEEE. https://doi.org/10.1109/SCW63240.2024.00265



Researchers from Princeton University are using ALCF supercomputers to simulate the three-dimensional character of supernova explosions in unprecedented detail. The explosion of a massive star starts in its core and drives a blast wave through many compositional interfaces, producing a complex network of inverted shells, bubbles, and aspherical structures. This figure depicts the violently mixed ejecta at the point of stellar breakout, with radioactive Nickel-56 (red), oxygen (blue), and the hydrogen/helium interface (yellow). The team's recent work was detailed in "A 3D Simulation of a Type II-P Supernova: From Core Bounce to beyond Shock Breakout," published in the *Astrophysical Journal*. *Image: ALCF Visualization and Data Analytics Team*

ALCF Projects

INCITE 2025

BIOLOGICAL SCIENCES

COMbining Deep-Learning with Physics-Based AffinIty EstimatiOn 3 (COMPBIO3)

PI Peter Coveney, University College London
HOURS ALCF: 795,000 Node-Hours
OLCF: 1,000,000 Node-Hours

Establishing Digital Twins for High-Throughput Cellular Analysis in Whole Blood

PI Amanda Randles, Duke University
HOURS ALCF: 200,000 Node-Hours
OLCF: 450,000 Node-Hours

ExaCortex: Exascale Reconstruction of Human Cerebral Cortex

PI Nicola Ferrier, Argonne National Laboratory
HOURS ALCF: 250,000 Node-Hours

Foundation Models for Predictive Molecular Epidemiology

PI Arvind Ramanathan, Argonne National Laboratory
HOURS ALCF: 1,750,000 Node-Hours

Large-Scale Simulations of Inner-Ear Mechanotransduction Complexes

PI Marcos Sotomayor, University of Chicago
HOURS ALCF: 200,000 Node-Hours

Multi-Resolution Genome Folding: Ensemble 3D Structures Across Diverse Tissues

PI Jie Liang, University of Illinois Chicago
HOURS ALCF: 1,000,000 Node-Hours

OpenFold-Powered Machine Learning of Protein-Protein Interactions and Complexes

PI Mohammed AlQuraishi, Columbia University
HOURS ALCF: 100,000 Node-Hours

CHEMISTRY

Exascale Catalytic Chemistry

PI David Bross, Argonne National Laboratory
HOURS ALCF: 425,000 Node-Hours

Heteropolymer Design Harnessing New and Emerging Computing Technologies

PI Vikram Mulligan, Flatiron Institute
HOURS ALCF: 1,500,000 Node-Hours

High-Accuracy Quantum Simulations in Cancer Therapy Using Exascale Computing

PI Anouar Benali, Qubit Pharmaceuticals
HOURS ALCF: 120,000 Node-Hours

High-Throughput Lead Optimization with Amber

PI Darrin York, Rutgers University
HOURS ALCF: 200,000 Node-Hours

COMPUTER SCIENCE

AuroraGPT: A Large-Scale Foundation Model for Advancing Science

PI Rick Stevens, Argonne National Laboratory and The University of Chicago
HOURS ALCF: 920,000 Node-Hours

Experimental Realization of Certified Randomness from Quantum Supremacy

PI Ruslan Shaydulin, JPMorgan Chase
HOURS ALCF: 300,000 Node-Hours

EARTH SCIENCE

A Climate-Informed, Large-Scale, and High-Resolution Inundation Modeling Framework

PI Sudershan Gangrade, Oak Ridge National Laboratory
HOURS ALCF: 100,000 Node-Hours
OLCF: 1,300,000 Node-Hours

Energy Exascale Earth System Model

PI Peter Caldwell, Lawrence Livermore National Laboratory
HOURS ALCF: 1,000,000 Node-Hours
OLCF: 1,000,000 Node-Hours

High-Resolution GPU-Enabled SCREAM RRM Simulations for Extreme Weather and Climate Events

PI Brandi Gamelin, Argonne National Laboratory
HOURS ALCF: 150,000 Node-Hours
OLCF: 300,000 Node-Hours

ENERGY TECHNOLOGIES

Foundation Models for Crystals

PI Venkatasubramanian Viswanathan, University of Michigan
HOURS ALCF: 200,000 Node-Hours
OLCF: 300,000 Node-Hours

ENGINEERING

Advancing Fusion and Fission Energy through Exascale

PI Elia Merzari, Pennsylvania State University
HOURS ALCF: 200,000 Node-Hours
OLCF: 250,000 Node-Hours

Exascale Simulations of Rayleigh-Benard Convection

PI Janet Scheel, Occidental College
HOURS ALCF: 800,000 Node-Hours

High-Fidelity Simulations of Open Fan/Wing Integration for Sustainable Aviation

PI Stephan Priebe, GE Aerospace Research
HOURS ALCF: 40,000 Node-Hours
OLCF: 800,000 Node-Hours

Hypersonic Transition Control via Porous and Wavy Walls at High Reynolds Numbers

PI Carlo Scalo, Purdue University
HOURS ALCF: 150,000 Node-Hours

Online Machine Learning for Large Scale Turbulent Simulations

PI Kenneth Jansen, University of Colorado Boulder
HOURS ALCF: 375,000 Node-Hours

Scalar Transport in Turbulent Bubbly Flows: Insights from Interface-Resolved Simulations

PI Parisa Mirbod, University of Illinois Chicago
HOURS ALCF: 60,000 Node-Hours

Using Exascale Computing for New Insights into High-Lift Ground Testing

PI Eric Nielsen, NASA Langley Research Center
HOURS ALCF: 500,000 Node-Hours

MATERIALS SCIENCE

Advancing Fusion Reactor Materials Through Exascale Simulations

PI Tim Frolov, Lawrence Livermore National Laboratory
HOURS ALCF: 20,000 Node-Hours
OLCF: 980,000 Node-Hours

AI-Guided Exascale Simulations of Quantum Materials Manufacturing and Control

PI Aiichiro Nakano, University of Southern California
HOURS ALCF: 825,000 Node-Hours

Carbon at Extremes: Discovery Science with Exascale Computers

PI Ivan Oleynik, University of South Florida
HOURS ALCF: 300,000 Node-Hours
OLCF: 800,000 Node-Hours

Exascale Simulation of Correlated Election-Phonon Coupling in Quantum Materials

PI Zhenglu Li, University of Southern California
HOURS ALCF: 100,000 Node-Hours
OLCF: 450,000 Node-Hours

Exascale Simulations of Quantum Materials

PI Paul Kent, Oak Ridge National Laboratory
HOURS ALCF: 750,000 Node-Hours
OLCF: 600,000 Node-Hours

Heterogeneous Reaction Dynamics for Energy Storage and Hydrogen Production

PI Boris Kozinsky, Harvard University
HOURS ALCF: 50,000 Node-Hours
OLCF: 150,000 Node-Hours

High-Throughput Calculation of Materials Properties at Finite Temperature

PI Chris Wolverton, Northwestern University
HOURS ALCF: 150,000 Node-Hours

MOFA: Generative AI-Driven MOF Discovery for Carbon Capture at Exascale

PI Eliu Huerta, Argonne National Laboratory
HOURS ALCF: 400,000 Node-Hours
OLCF: 370,000 Node-Hours

QMC-HAMM: High-Accuracy Multiscale Models Using Quantum Monte Carlo

PI Lucas Wagner, University of Illinois
HOURS ALCF: 760,000 Node-Hours

PHYSICS

3D Imaging of Strong Interaction Nambu-Goldstone Bosons

PI Yong Zhao, Argonne National Laboratory
HOURS ALCF: 600,000 Node-Hours

Ab Initio Nuclear Structure and Nuclear Reactions

PI Gaute Hagen, Oak Ridge National Laboratory
HOURS ALCF: 600,000 Node-Hours
OLCF: 1,000,000 Node-Hours

Advances in Quark and Lepton Flavor Physics with Lattice QCD

PI Andreas Kronfeld, Fermi National Accelerator Laboratory
HOURS ALCF: 150,000 Node-Hours
OLCF: 1,000,000 Node-Hours

Exascale Gyrokinetic Study of ITER Challenge on Power-Exhaust and ELM-Free Edge

PI CS Chang, Princeton Plasma Physics Laboratory
HOURS ALCF: 1,200,000 Node-Hours
OLCF: 1,300,000 Node-Hours

Exascale Models of Astrophysical Thermonuclear Explosions

PI Michael Zingale, Stony Brook University
HOURS ALCF: 50,000 Node-Hours
OLCF: 450,000 Node-Hours

Exascale Simulation and Deep Learning Model for Energetic Particles in Burning Plasmas

PI Zhihong Lin, University of California, Irvine
HOURS ALCF: 100,000 Node-Hours
OLCF: 100,000 Node-Hours

Exascale Simulations of Compact Binary Mergers

PI David Radice, Pennsylvania State University
HOURS ALCF: 500,000 Node-Hours

High-Fidelity Turbulence-Based Predictions of Stellarator Reactor Performance

PI Noah Mandell, Type One Energy
HOURS ALCF: 200,000 Node-Hours

Probing the Primordial Universe with Exascale Simulations

PI Nicholas Frontiere, Argonne National Laboratory
HOURS ALCF: 1,050,000 Node-Hours

Radiation-Dominated Black Hole Accretion

PI James Stone, Institute for Advanced Studies
HOURS ALCF: 1,100,000 Node-Hours
OLCF:1,000,000 Node-Hours

State-of-the-Art High-Resolution 3D Simulations of Core-Collapse Supernovae

PI Adam Burrows, Princeton University
HOURS ALCF: 250,000 Node-Hours
OLCF: 3,000,000 Node-Hours

QCD Under Extreme Conditions

PI Zoltan Fodor, Pennsylvania State University
HOURS ALCF: 100,000 Node-Hours
OLCF: 100,000 Node-Hours

ALCC 2025–2026

BIOLOGICAL SCIENCES

Advanced Modeling for Two-Dimensional Materials and Transition Metal Oxides

PI Daniel Mejia Rodriguez, Pacific Northwest National Laboratory
HOURS ALCF: 275,000 Node-Hours
OLCF: 275,000 Node-Hours

EMERGE: ExaEpi Calibration Runs and Surrogate Models

PI Peter Nugent, Lawerence Berkeley National Laboratory
HOURS ALCF: 425,000 Node-Hours
OLCF: 425,000 Node-Hours
NERSC: 100,000 Node-Hours

Kinetic Plasma Model Investigation of Z Pinch Physics at Fusion Conditions

PI Noah Reddell, Zap Energy Inc.
HOURS ALCF: 200,000 Node-Hours
OLCF: 800,000 Node-Hours

COMPUTER SCIENCE

Causal Online Alignment for Reliable Foundation Models

PI Emmanouil Koukoumidis, Oumi PBC
HOURS ALCF: 100,000 Node-Hours
NERSC: 147,500 Node-Hours

Privacy-Preserving Federated Learning for Foundation Models

PI Kibaek Kim, Argonne National Laboratory
HOURS ALCF: 409,600 Node-Hours
OLCF: 614,400 Node-Hours
NERSC: 102,400 Node-Hours

Space Traffic Autonomy: Leadership Computing Advances Hierarchical Planning R&D

PI Allan Grosvenor, Microsurgeonbot Inc.
HOURS ALCF: 685,000 Node-Hours
OLCF: 685,000 Node-Hours

Tensor-Compressed Sustainable Pre-Training of Extreme-Scale Foundation Models

PI Zheng Zhang,
University of California, Santa Barbara
HOURS ALCF: 700,000 Node-Hours
NERSC: 80,000 Node-Hours

Training Multimodal Models for HPC Code and Data

PI Harshitha Menon,
Lawrence Livermore National Laboratory
HOURS ALCF: 200,000 Node-Hours
OLCF: 300,000 Node-Hours
NERSC: 200,000 Node-Hours

CHEMISTRY

High-Precision Heterogeneous Catalysis by Quantum Monte Carlo

PI Michal Bajdich,
SLAC National Accelerator Laboratory
HOURS ALCF: 1,500,000 Node-Hours

Microscopic Insight into Transport Properties of Li-Battery Electrolytes

PI Wei Jiang, Argonne National Laboratory
HOURS ALCF: 600,000 Node-Hours

ENERGY TECHNOLOGIES

Development of AI-Powered Application for Advanced Nuclear Reactor Design

PI Jun Fang, Argonne National Laboratory
HOURS ALCF: 100,000 Node-Hours

High Energy Density Physics of Novel Inertial Fusion Energy Ablator Materials

PI Ivan Oleynik, University of South Florida
HOURS ALCF: 800,000 Node-Hours
OLCF: 1,000,000 Node-Hours
NERSC: 100,000 Node-Hours

High-Fidelity Simulations of Helium-Air Mixing in HTGR Cavities 2.0

PI Taehun Lee, The City College of New York
HOURS ALCF: 70,000 Node-Hours

LES and DNS Simulation on Flow and Heat Transfer Behavior in Involute Plate Research Reactor

PI Yiqi Yu, Argonne National Laboratory
HOURS ALCF: 150,000 Node-Hours
NERSC: 550,000 Node-Hours

Pathfinding Integrated and Automatic Experimental Analyses for DIII-D Research

PI Mark Kostuk, General Atomics
HOURS ALCF: 15,000 Node-Hours
NERSC: 35,000 Node-Hours

Turbulence Database for Fusion Energy Science

PI Jeff Candy, General Atomics
HOURS ALCF: 300,000 Node-Hours
OLCF: 300,000 Node-Hours

Validation of Stability Simulations and Predictive Scaling Toward Next Step FRC Fusion Device

PI Roelof Groenewald, TAE Technologies Inc.
HOURS ALCF: 250,000 Node-Hours

ENGINEERING

High-Fidelity Simulations of Bubble-Laden Turbulent Flows with Surfactants

PI Suhas Jain, Georgia Institute of Technology
HOURS ALCF: 1,000,000 Node-Hours
OLCF: 1,000,000 Node-Hours

High-Fidelity Simulation of SAF End Use: Emissions and Operability Studies

PI Bruce Perry,
National Renewable Energy Laboratory
HOURS ALCF: 300,000 Node-Hours
OLCF: 600,000 Node-Hours

Integrated and Detailed Simulation of Combustor and Turbine Interaction in a Jet Engine

PI Dheeraj Kapilavai, GE Aerospace Research
HOURS ALCF: 425,000 Node-Hours

Predicting Flow Distortions in Serpentine Engine Inlets

PI Parviz Moin, Stanford University
HOURS ALCF: 1,700,000 Node-Hours
OLCF: 1,600,000 Node-Hours

PHYSICS

Enabling Exascale Discovery in Neutrino Science

PI Thomas Wester, University of Chicago
HOURS ALCF: 200,000 Node-Hours

Enhancing APS-Enabled Research Through Integrated Research Infrastructure

PI Nicholas Schwarz,
Argonne National Laboratory
HOURS ALCF: 150,000 Node-Hours
OLCF: 10,000 Node-Hours
NERSC: 10,000 Node-Hours

Hadronic Contributions to the Muon g-2 from Lattice QCD

PI Thomas Blum, University of Connecticut
HOURS ALCF: 2,000,000 Node-Hours
OLCF: 2,000,000 Node-Hours

Simulating Large-scale Long-lived Neutron Star Remnants from Binary Neutron Star Mergers

PI Ore Gottlieb, Flatiron Institute
HOURS ALCF: 667,000 Node-Hours

ALCC 2024–2025

BIOLOGICAL SCIENCES

Building Digital Twin of a Model Host-Pathogen System for Enhancing Biopreparedness

PI Margaret S. Cheung,
Pacific Northwest National Laboratory
HOURS ALCF: 200,000 Node-Hours
OLCF, 50,000 Node-Hours
NERSC: 300,000 Node-Hours

Foundation Neuroscience AI Model-NeuroX

PI Shinjae Yoo,
Brookhaven National Laboratory
HOURS ALCF: 290,000 Node-Hours
OLCF: 152,000 Node-Hours
NERSC: 120,000 Node-Hours

CHEMISTRY

Exploring Exascale Quantum Chemical Methods for Transition Metal Chemistry

PI Daniel Mejia Rodriguez,
Pacific Northwest National Laboratory
HOURS ALCF: 313,442 Node-Hours
OLCF: 145,628 Node-Hours

Highly Scalable Ab Initio Simulations of N-Doped Porous Materials for Carbon Capture

PI Mark Gordon, Ames National Laboratory
HOURS ALCF: 2,000,000 Node-Hours

COMPUTER SCIENCE

Scalable and Resilient Modeling for Federated Learning Systems and Applications

PI Xiaoyi Lu, University of California
HOURS ALCF: 44,800 Node-Hours
OLCF: 15,360 Node-Hours
NERSC: 207,520 Node-Hours

ENERGY TECHNOLOGIES

DNS of Buoyancy-Driven Flows for Developing NN-Informed High-Fidelity Turbulence Closures

PI Som Dutta, Utah State University
HOURS ALCF: 100,000 Node-Hours
OLCF: 300,000 Node-Hours
NERSC: 200,000 Node-Hours

Exascale Computing for Energy Applications

PI Misun Min, Argonne National Laboratory
HOURS ALCF: 250,000 Node-Hours
OLCF: 200,000 Node-Hours
NERSC: 50,000 Node-Hours

High-Fidelity CFD Enabling Advanced Nuclear Power

PI Dillon Shaver, Argonne National Laboratory
HOURS ALCF: 150,000 Node-Hours
OLCF: 200,000 Node-Hours

High-Fidelity Numerical Analysis on Flow and Heat Transfer Behavior in Involute Plate Research Reactor to Support the Conversion Program

PI Yiqi Yu, Argonne National Laboratory
HOURS ALCF: 200,000 Node-Hours
OLCF: 500,000 Node-Hours

High-Fidelity Simulations of Helium-Air Mixing in High-Temperature Gas Reactor Cavities

PI Saumil Patel, Argonne National Laboratory
HOURS ALCF: 115,000 Node-Hours

ENGINEERING

Autonomy for DOE Simulations

PI Allan Grosvenor, MSBAI
HOURS ALCF: 30,000 Node-Hours
OLCF: 100,000 Node-Hours

MATERIALS SCIENCE

High Energy Density Physics of Novel Inertial Fusion Energy Ablator Materials

PI Ivan Oleynik, University of South Florida
HOURS ALCF: 600,000 Node-Hours
OLCF: 900,000 Node-Hours

Machine Learning-Enabled Atomistic Simulation of Iron at Extreme Pressure

PI Robert Rudd,
Lawrence Livermore National Laboratory
HOURS ALCF: 150,000 Node-Hours

Predicting Heterogeneous Photocatalysts Using Large-scale Ab Initio Calculations

PI Felipe Jornada, Stanford University
HOURS ALCF: 100,000 Node-Hours
OLCF: 140,000 Node-Hours

PHYSICS

Continuum Limit Latice Calculation of Direct CP-Violation in Kaon Decays

PI Christopher Kelly,
Brookhaven National Laboratory
HOURS ALCF: 135,000 Node-Hours

Energy Partition and Particle Acceleration in Laser-Driven Laboratory Magnetized Shocks

PI Frederico Fiuza,
SLAC National Accelerator Laboratory
HOURS ALCF: 300,000 Node-Hours
NERSC: 150,000 Node-Hours

AURORA EARLY SCIENCE PROGRAM

Accelerated Deep Learning Discovery in Fusion Energy Science

PI William Tang,
Princeton Plasma Physics Laboratory

Dark Sky Mining

PI Salman Habib,
Argonne National Laboratory

Data Analytics and Machine Learning for Exascale Computational Fluid Dynamics

PI Kenneth Jansen,
University of Colorado Boulder

Enabling Connectomics at Exascale to Facilitate Discoveries in Neuroscience

PI Nicola Ferrier,
Argonne National Laboratory

Exascale Computational Catalysis

PI David Bross, Argonne National Laboratory

Extending Moore's Law Computing with Quantum Monte Carlo

PI Anouar Benali,
Argonne National Laboratory

Extreme-Scale Cosmological Hydrodynamics

PI Katrin Heitmann,
Argonne National Laboratory

Extreme-Scale In-Situ Visualization and Analysis of Fluid-Structure-Interaction Simulations

PI Amanda Randles, Duke University and
Oak Ridge National Laboratory

Extreme-Scale Unstructured Adaptive CFD

PI Kenneth Jansen,
University of Colorado Boulder

High-Fidelity Simulation of Fusion Reactor Boundary Plasmas

PI C.S. Chang,
Princeton Plasma Physics Laboratory

Machine Learning for Lattice Quantum Chromodynamics

PI William Detmold,
Massachusetts Institute of Technology

Many-Body Perturbation Theory Meets Machine Learning to Discover Singlet Fission Materials

PI Noa Marom, Carnegie Mellon University

NWChemEx: Tackling Chemical, Materials, and Biochemical Challenges in the Exascale Era

PI Thom Dunning,
Pacific Northwest National Laboratory

Simulating and Learning in the ATLAS Detector at the Exascale

PI Walter Hopkins,
Argonne National Laboratory

Virtual Drug Response Prediction

PI Rick Stevens,
Argonne National Laboratory

DIRECTOR'S DISCRETIONARY

The following list provides a sampling of the many Director's Discretionary projects at the ALCF.

BIOLOGICAL SCIENCES

Free Energy Landscapes of Membrane Transport Proteins

PI Benoît Roux, University of Chicago

Scaling Genomic Variant Callers to Leadership-Class Systems: A Collaboration Between VA-MVP and DOE

PI Ravi Madduri,
Argonne National Laboratory

Structural Modeling of High-Confident Direct Protein Interactions Using AlphaFold

PI Kevin Drew, University of Illinois Chicago

Understand Cryptic Site Binding and Allosteric Mechanisms of NNRTIs

PI Ao Ma, University of Illinois Chicago

Understanding the Rigorous Molecular Mechanism of Drug Resistance to HIV Protease Inhibitors

PI Ao Ma, University of Illinois Chicago

CHEMISTRY

Exascale Simulations of Non-Equilibrium Shock Dynamics of Rocket Fuels

PI Subramanian Sankaranarayanan,
Argonne National Laboratory

High-Fidelity Simulations of Flame Flashback in Hydrogen-Fueled Gas Turbine Engines

PI Pinaki Pal, Argonne National Laboratory

A Huckel Method-Inspired Similarity Algorithm to Bridge the Materials Gap in Heterogeneous Metal Catalysis

PI Siddharth Deshpande,
University of Rochester

Interactive Quantum Chemistry

PI Murat Keçeli, Argonne National Laboratory

Speeding Up the Computation of Long-Range Electrostatic Interactions in NWChemEx

PI Ryan M. Richard, Iowa State University

COMPUTER SCIENCE

APS Beamline Data Processing and Analysis

PI Rafael Vescovi, Argonne National Laboratory

Deep Learning IO Performance Evaluation on HPC Storage and File System and AI Testbeds

PI Huihuo Zheng, Argonne National Laboratory

Diaspora: Resilience-Enabling Services for Science from HPC to Edge

PI Ryan Chard, Argonne National Laboratory

FAIR Surrogate Benchmarks Supporting AI and Simulation Research

PI Pete Beckman, Northwestern University

Graph Analytics Codesign on GPUs

PI Sayan Ghosh, Pacific Northwest National Laboratory

MPICH - A High Performance and Widely Portable MPI Implementation

PI Kenneth Raffenetti, Argonne National Laboratory

QAOA Performance with Finite-Size Correction

PI Minzhao Liu, JPMorganChase

Scalable Graph Foundation Models for Atomistic Materials Modeling

PI Massimiliano Lupo Pasini, Oak Ridge National Laboratory

Workflow Scaling on Aurora

PI Thomas Uram, Argonne National Laboratory

EARTH SCIENCE

Development of a Stochastic AI Seasonal to Subseasonal Weather Forecasting Modeling System (SAFS)

PI Rao Kotamarthi, Argonne National Laboratory

Physics-Based Modeling for Fuel Management

PI Marcos Vanella, National Institute of Standards and Technology

ENERGY TECHNOLOGIES

Robust Film Cooling Under Manufacturing Uncertainty For Improved Jet Engine Life Cycle Energy Efficiency

PI Pinaki Pal, Argonne National Laboratory

Towards Exascale Simulations of Transportation and Power Generation Applications

PI Mushin Ameen, Argonne National Laboratory

ENGINEERING

Compressible Flow Code Evaluation

PI Francesco Salvadore, Cineca

Scalable In-Situ SciML+CFD for Long-Time Simulations of Turbulent Reacting Flows

PI Shivam Barwey, Argonne National Laboratory

SPH Simulations of Cell Separation in Spiral Channels by Inertial Microfluidics

PI Zhangli Peng, University of Illinois Chicago

MATERIALS SCIENCE

Broad-Scope Reasoning Artificial Intelligence for Nano-Micro Material and Devices Identification, Assessment, and Categorization

PI Rui Ding, University of Chicago

Computational Design of Novel Semiconductors for Power and Energy Applications

PI Feliciano Giustino, University of Texas at Austin

Data-Driven Molecular Engineering of Solar-Powered Windows

PI Jacqueline Cole, University of Cambridge

DFT Calculations

PI Vishwas Hebbur Venkata Subba Rao, Argonne National Laboratory

Metascalable Layered Materials Genome

PI Aiichiro Nakano, University of Southern California

Metastable Phase Diagram of Material

PI Subramanian Sankaranarayanan, Argonne National Laboratory

PHYSICS

3D Imaging of the Pion from Lattice QCD

PI Yong Zhao, Argonne National Laboratory

Enabling Precise Measurement of Neutrino Interactions Using Advanced Computing in DUNE

PI Aleena Rafique, Argonne National Laboratory

Lattice Quantum Chromodynamics Calculations for Particle and Nuclear Physics

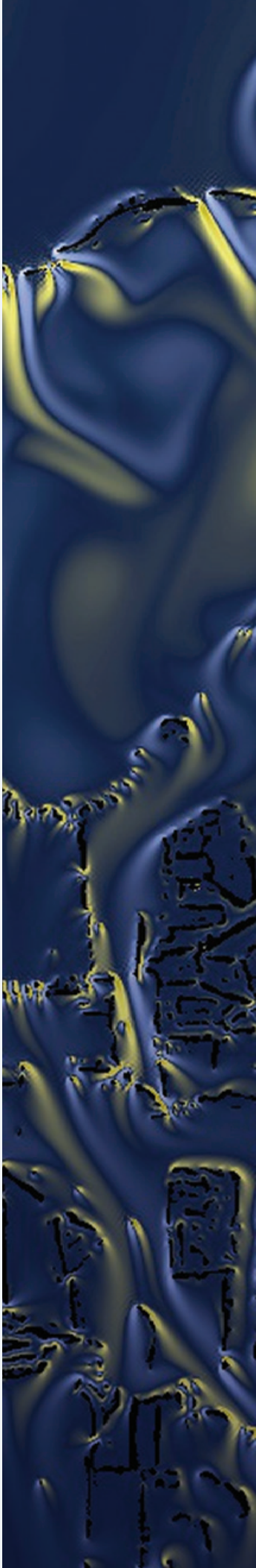
PI Norman Christ, Columbia University

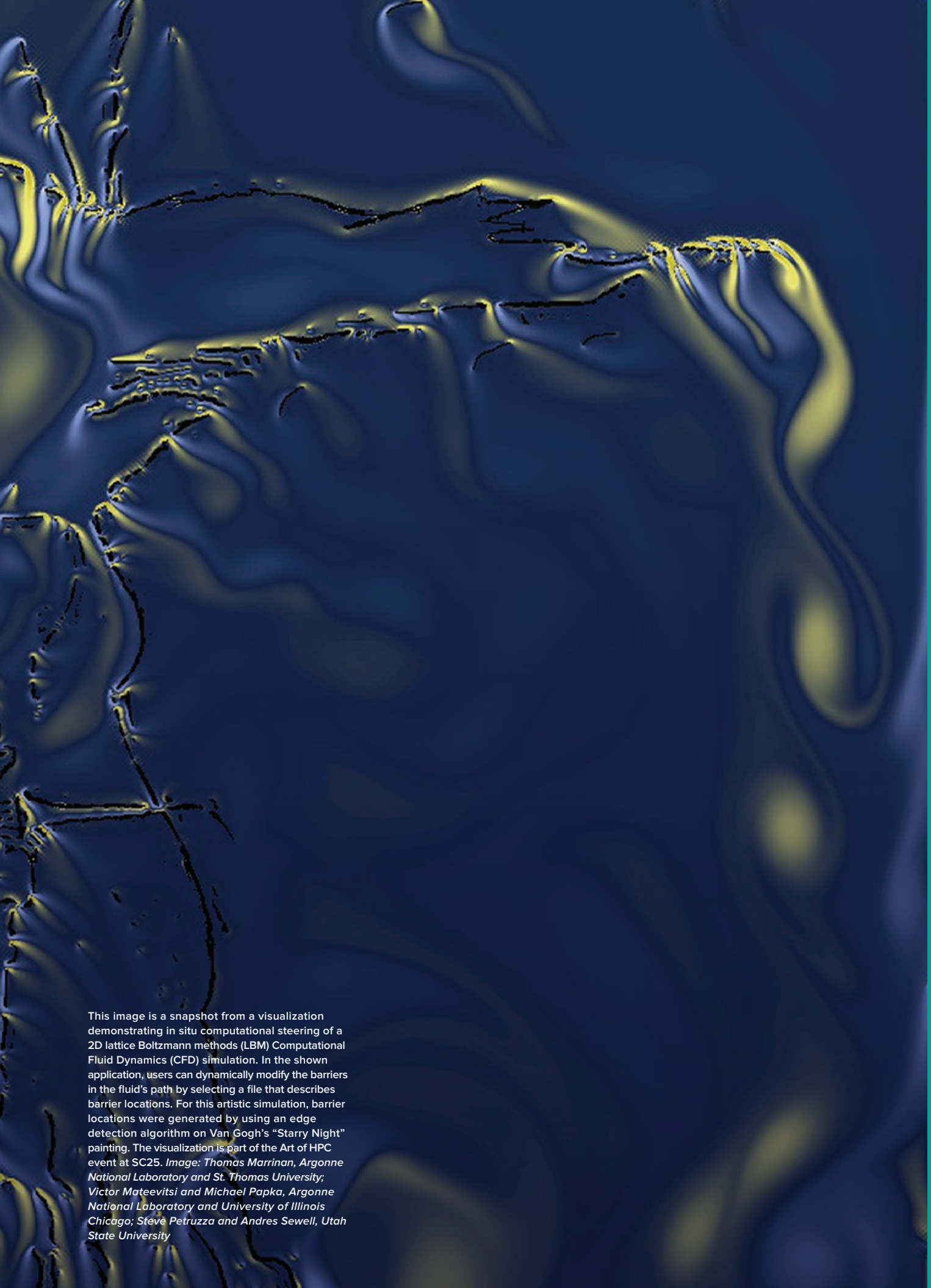
Novel LArTPC Simulation and Reconstruction

PI Zelimir Djurcic, Argonne National Laboratory

Simulation-Based Dark Energy Inference

PI Marco Gatti, University of Pennsylvania





This image is a snapshot from a visualization demonstrating in situ computational steering of a 2D lattice Boltzmann methods (LBM) Computational Fluid Dynamics (CFD) simulation. In the shown application, users can dynamically modify the barriers in the fluid's path by selecting a file that describes barrier locations. For this artistic simulation, barrier locations were generated by using an edge detection algorithm on Van Gogh's "Starry Night" painting. The visualization is part of the Art of HPC event at SC25. *Image: Thomas Marrinan, Argonne National Laboratory and St. Thomas University; Victor Mateevitsi and Michael Papka, Argonne National Laboratory and University of Illinois Chicago; Steve Petruzza and Andres Sewell, Utah State University*

About the Argonne Leadership Computing Facility

Argonne's Leadership Computing Facility Division operates the Argonne Leadership Computing Facility (ALCF) as part of the U.S. Department of Energy's effort to provide leadership-class computing resources to the scientific community. The ALCF is supported by the DOE Office of Science, Advanced Scientific Computing Research (ASCR) program.

About Argonne National Laboratory

Argonne is a U.S. Department of Energy Laboratory managed by UChicago Argonne, LLC, under contract DE-AC02-06CH11357. The Laboratory's main facility is outside of Chicago, at 9700 South Cass Avenue, Argonne, Illinois 60439. For information about Argonne and its pioneering science and technology programs, visit www.anl.gov.

Availability of this Report (ANL/ALCF-25/4)

Online Access: U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available for free via DOE's SciTech Connect (http://www.osti.gov/scitech/).

Reports not in digital format may be purchased by the public from the National Technical Information Service (NTIS):

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd.
Alexandra, VA 22312

phone | 800.553.NTIS (6847) or 703.605.6000
fax | 703.605.6900
orders@ntis.gov
www.ntis.gov

Reports not in digital format are available to DOE and DOE contractors from the Office of Scientific and Technical Information (OSTI):

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062

phone | 865.576.8401
fax | 865.576.5728
reports@adonis.osti.gov
www.osti.gov

ALCF Leadership: Michael E. Papka (Division Director), Bill Allcock/Rick Wagner (Director of Operations), Susan Coghlan (ALCF-3 Project Director), Kevin Harms (ALCF-4 Technical Director), Kalyan Kumaran (Director of Technology), Jini Ramprakash (Deputy Division Director, ALCF-4 Project Director), and Katherine Riley (Director of Science)

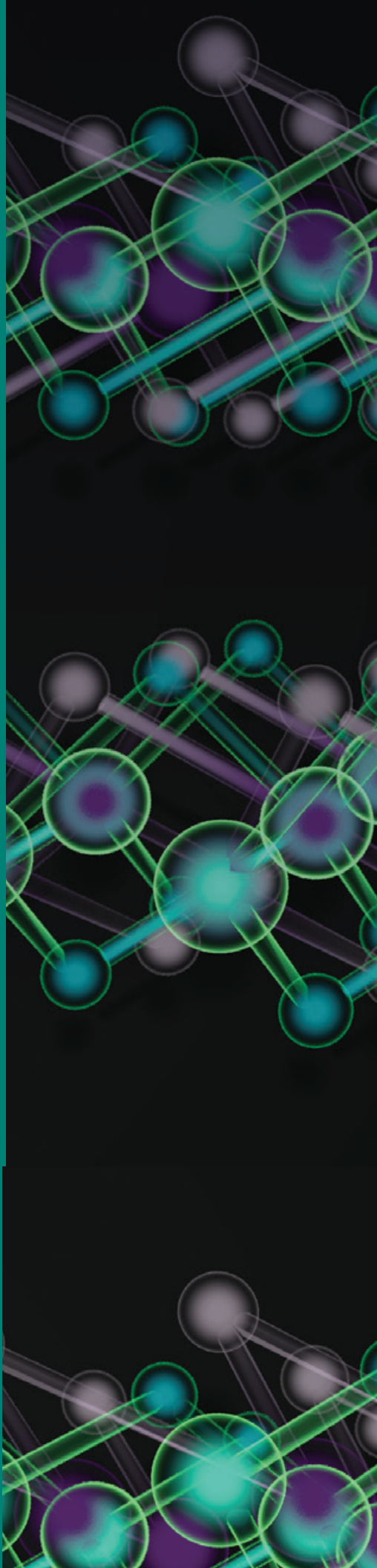
Editorial Team: Beth Cerny, Jim Collins, Nils Heinonen, Logan Ludwig, Rachel Taub, and Laura Wolf

Design and production: Sandbox Studio, Chicago

Argonne NATIONAL LABORATORY

Argonne Leadership Computing Facility

CONTACT
media@alcf.anl.gov
alcf.anl.gov



U.S. DEPARTMENT of ENERGY

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.