
Argonne
NATIONAL LABORATORY
Argonne Leadership
Computing Facility
Science
2024

Researchers used the ALCF's Polaris supercomputer to perform GPU-enabled weather simulations at cloud-resolving (1 km) spatial resolution for the month of September 2017. Covering the entire North American continent and approximately 3 billion grid cells, this visualization shows water vapor (red-blue) and cloud water (white-grayscale) fields, with wind vectors highlighting the jet stream pattern.
Image: ALCF Visualization and Data Analytics Team; Rao Kotamarthi and Gökhan Sever, Argonne National Laboratory

CONTENTS

MICHAEL E. PAPKA
ALCF Director

BILL ALLCOCK
ALCF Director of Operations

SUSAN COGHLAN
ALCF-X Project Director

KALYAN KUMARAN
ALCF Director of Technology

JINI RAMPRAKASH
ALCF Deputy Director

KATHERINE RILEY
ALCF Director of Science

MESSAGE FROM ALCF LEADERSHIP

The fully assembled Aurora is now generating data that has the potential to revolutionize many scientific fields. Having surpassed the exascale barrier this year and ranked among the fastest machines in the world for traditional computing tasks, Aurora is also unmatched as the world's fastest artificial intelligence (AI) system for open science.

Aurora Early Science projects range from using machine learning to discover singlet fission materials to tackling biochemical challenges like biofuels to understanding extreme-scale cosmological hydrodynamics. All of these projects have seen significant gains on Aurora, with equally far-reaching expectations for discoveries in fields as diverse as materials science, drug discovery, and neutrino physics.

In other facility updates, with the retirement of Theta and ThetaGPU in December 2023, ALCF decoupled and reconfigured ThetaGPU to become a new system named Sophia, which went into production in July with a focus on running JupyterHub instances and large language models. ALCF's GPU-accelerated system, Polaris, continues to support projects from INCITE, ALCC, and other programs and serves as a development platform for merging ALCF resources with large-scale experimental facilities. Lastly, the ALCF's AI Testbed of next-generation AI accelerator technologies continues to enable a range of exploratory research campaigns.

With its powerful computing resources and strong expertise in using AI in science, ALCF continues to grow its scientific user community. Building on a decade of hosting training workshops and various outreach activities to teach young researchers how to use Department of Energy (DOE) supercomputers, ALCF's training program in the AI space has taken off—attracting over 600 participants in the past three years alone. This year, ALCF's new Lighthouse Initiative to build a national network of university partnerships kicked off a pilot in collaboration with the University of Chicago and the University of Illinois Chicago, with plans to expand to more universities next year.

Along with Aurora and Polaris, the ALCF AI Testbed is enabling the facility to support major U.S. initiatives to expand the nation's AI capabilities for pioneering science. These efforts include the DOE's Frontiers in Artificial Intelligence for Science, Security and Technology (FASST) and the National Science Foundation-led National Artificial Intelligence Research Resource (NAIRR) Pilot launched this year in collaboration with DOE and other partners.

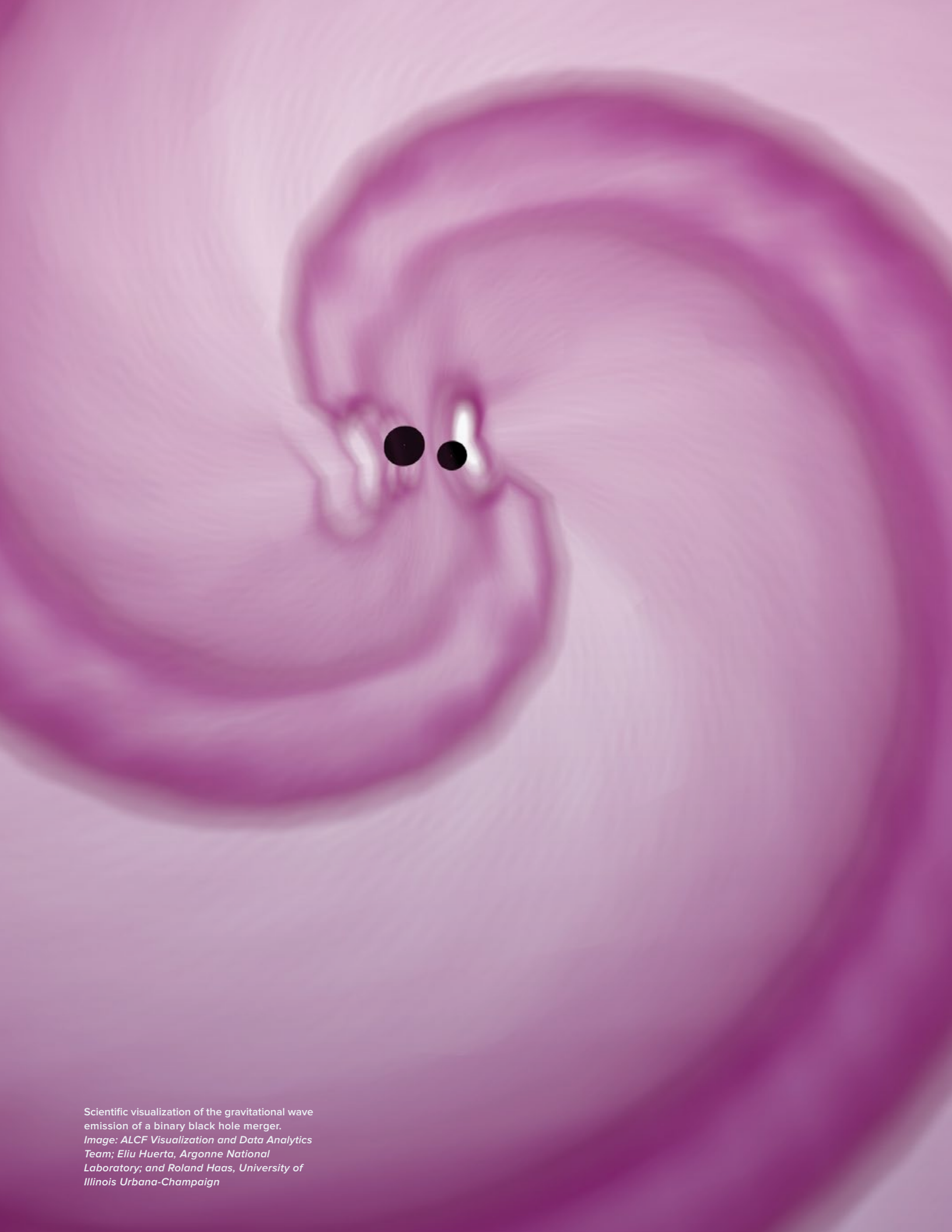
Meanwhile, Argonne's Nexus effort, launched in 2023, is firmly underway to integrate supercomputers and experimental facilities on the scale of the DIII-D National Fusion Facility and CERN's Large Hadron Collider. Nexus builds on established collaborations between ALCF and other DOE computing facilities to help shape DOE program plans for a future Integrated Research Infrastructure (IRI) and to coordinate and align Argonne's IRI research and collaborations—a significant focus being ALCF's coupling with the newly upgraded Advanced Photon Source (APS)—with DOE's efforts in this area. The new tools deployed between ALCF and APS, along with the latest capabilities for on-demand computing and managing complex workflows developed by ALCF and its partners in other large experimental facilities, support the overall IRI program goals.

Finally, we present the science stories—the highlights of this annual publication and a selection of the successes that are fundamental and critical to our mission year after year. These stories are among the most exciting results to emerge from ALCF this year, whether as publications following multi-year campaigns or as work that significantly advances future discoveries.

As we conclude, we want to acknowledge all the people who make the ALCF an incredible place to work and do world-changing work: our expert staff, program officers, and especially our user community, which continues to challenge us to find new ways to excel in HPC for scientific discovery.

ARGONNE LEADERSHIP COMPUTING FACILITY

The ALCF enables breakthroughs in science and engineering by providing supercomputing and AI resources to the research community.



Scientific visualization of the gravitational wave emission of a binary black hole merger.
Image: ALCF Visualization and Data Analytics Team; Eliu Huerta, Argonne National Laboratory; and Roland Haas, University of Illinois Urbana-Champaign

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 resources—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 and support services, 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 pursue discoveries 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 pursue 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 applies broad expertise 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 experience with a wide range of numerical methods, programming models, and computational approaches, staff scientists and performance engineers help researchers optimize the performance and productivity of simulation, data, and learning applications on ALCF systems.

Technology
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.

ALCF computer scientists, performance engineers, and software engineers develop and optimize new tools and capabilities to facilitate 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, the ALCF team is actively involved in programming language standardization efforts and contributes to cross-platform libraries to further enable the portability of HPC applications.

Outreach
ALCF staff members organize and participate in training events that prepare researchers for efficient use of leadership computing systems. They also participate in a wide variety of educational activities aimed at cultivating a diverse and 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.

Supercomputing Resources

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	SOPHIA
Purpose	Science Campaigns	Science Campaigns	Science Campaigns
Architecture	HPE Cray EX	HPE Apollo 6500 Gen10+	NVIDIA DGX A100
Peak Performance	2 EF	34 PF; 44 PF of Tensor Core FP64 performance	3.9 PF
Processors per Node	2 Intel Xeon CPU Max Series processors	3rd Gen AMD EPYC	2 AMD EPYC 7742
GPU per Node	6 Intel Data Center GPU Max Series	4 NVIDIA A100 Tensor Core	8 NVIDIA A100 Tensor Core
Nodes	10,624 CPUs: 21,248 GPUs: 63,744	560	24
Cores	9,264,128	17,920	3,072
Memory	20.4 PB	280 TB (DDR4); 87.5 TB (HBM)	26 TB (DDR4); 8.32 TB (GPU)
Interconnect	HPE Slingshot 11 with Dragonfly Configuration	HPE Slingshot 11 with Dragonfly Configuration	NVIDIA HDR InfiniBand
Racks	166	40	7

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	SAMBANOVA CARDINAL SN30	GROQRACK	GRAPHCORE BOW POD-64	HABANA GAUDI-1
System Size	2 Nodes (Each with a Wafer-Scale Engine) Including MemoryX and SwarmX	64 Accelerators (8 Nodes and 8 Accelerators per Node)	72 Accelerators (9 Nodes and 8 Accelerators per Node)	64 Accelerators (4 Nodes and 16 Accelerators per Node)	16 Accelerators (2 Nodes and 8 Accelerators per Node)
Compute Units per Accelerator	850,000 Cores	1,280 Programmable Compute Units	5,120 Vector ALUs	1,472 Independent Processing Units	8 TPC + GEMM Engine
Single Accelerator Performance (TFlops)	> 5,780 (FP16)	>660 (BF16)	>188 (FP16) >750 (INT8)	>250 (FP16)	>150 (FP16)
Software Stack Support	Cerebras SDK, TensorFlow, PyTorch	SambaFlow, PyTorch	GroqWare SDK, ONNX	PopART, TensorFlow, PyTorch, ONNX	SynapseAI, TensorFlow, PyTorch
Interconnect	Ethernet-based	Ethernet-based	RealScale™	IPU Link	Ethernet-based

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	EAGLE	GRAND	SWIFT	TAPE STORAGE
File System	Lustre	Lustre	Lustre	—
Storage System	HPE ClusterStor E1000	HPE ClusterStor E1000	All NVMe Flash Storage Array	LTO6 and LTO8 Tape Technology
Usable Capacity	100 PB	100 PB	123 TB	300 PB
Sustained Data Transfer Rate	650 GB/s	650 GB/s	48 GB/s	—
Disk Drives	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 technologies that make up the JLSE.

Florentia: Test and development system equipped with early versions of the Sapphire Rapids CPUs and Ponte Vecchio GPUs that power Aurora

Arcticus, DevEP, Iris: Intel discrete and integrated GPU testbeds to support the development, optimization, and scaling of applications and software for Aurora

Aurora Software Development Kit: Frequently updated version of the publicly available Intel oneAPI toolkit for Aurora development

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

Presque: Intel DAOS nodes for testing the Aurora storage system

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

NVIDIA and AMD GPUs: Clusters of NVIDIA V100, A100, and A40 GPUs, and AMD MI50 and MI100 GPUs for preparing applications for heterogeneous computing architectures

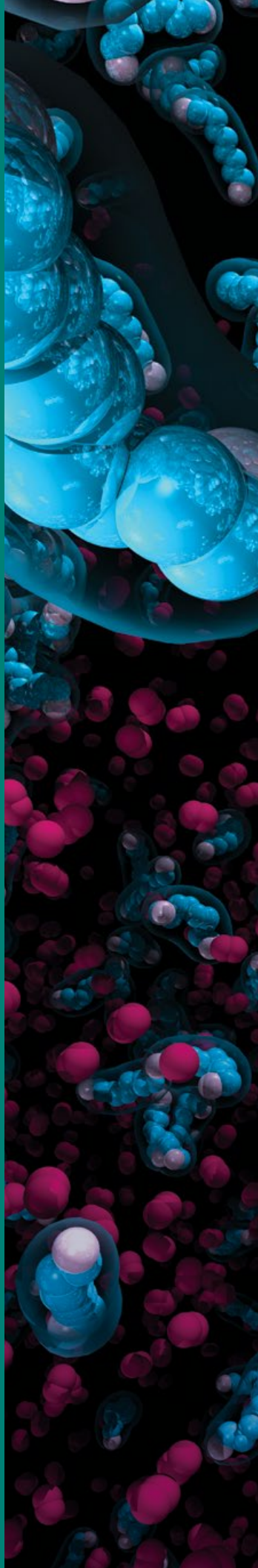
NVIDIA Bluefield-2 DPU SmartNICs: Platform used for confidential computing, MPICH offloading, and APS data transfer acceleration

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 supercomputing and AI resources and expertise, the ALCF is driving pioneering research at the intersection of simulation, large-scale data analytics, and machine learning.



Fuel molecules cluster when they are heated to extreme temperatures and then cooled. Scientists study the cluster size distributions and the number of clusters in the system as smaller liquid drops form and coalesce over time. Understanding this clustering behavior will help reveal the underlying dynamics of nano-sized fuel droplets, potentially aiding in the design of more efficient high-pressure combustion systems, such as diesel engines and rockets. *Image: ALCF Visualization and Analytics Team; Subramanian Sankaranarayanan, Argonne National Laboratory; Suman Chakraborty, University of Illinois Chicago*

Aurora Breaks the Exascale Barrier

The ALCF supercomputer officially surpassed the exascale threshold while also setting a new high mark in AI performance.

In a new set of benchmark results unveiled at the 2024 ISC High Performance conference in May, the ALCF's Aurora supercomputer cemented its status as one of the most powerful supercomputers ever built.

Aurora registered 1.012 exaflops using 9,234 nodes (87 percent of the system's 10,624 nodes) in its latest run for the High Performance LINPACK (HPL) benchmark. After making its Top500 debut in November 2023, the Argonne system now joins OLCF's Frontier as the world's second exascale machine on the Top500 list. Aurora also earned the top spot in a measure of AI performance, achieving 10.6 exaflops with 9,500 nodes on the HPL-MxP mixed-precision benchmark.

In addition, Aurora reached 24,250 GTEPS (giga-traversed edges per second) using 4,096 nodes on the Graph500 list—a benchmark for data-intensive applications. The Argonne system also made its debut with the High-Performance Conjugate Gradients (HPCG) Benchmark, achieving 5,612.6 teraflops with 4,096 nodes. Finally, Aurora's storage system, DAOS, retained the top spot on the IO500 production list, a semi-annual ranking of HPC storage performance.

Built by Intel and Hewlett Packard Enterprise (HPE), Aurora's first-of-its-kind architecture includes new technologies being deployed at an unprecedented scale. The supercomputer's 63,744 graphics processing units (GPUs) make it the world's largest GPU-powered system yet. Aurora also has more endpoints in its HPE Slingshot interconnect than any other system to date.

While the latest benchmark results illustrate the system's powerful capabilities, enabling groundbreaking science is the ultimate goal. Throughout the process of stress-testing and preparing Aurora for production in 2025, teams participating in the ALCF's Aurora Early Science Program and DOE's Exascale Computing Project have been working to scale and optimize scientific codes and software for the system, demonstrating strong early performance gains. In the following pages, we take a look at some of the initial research campaigns being carried out on Aurora.

The ALCF is readying Aurora for production use in 2025.

ISC 2024 Performance Benchmark Results

At the 2024 ISC High Performance conference in Hamburg, Germany, in May, Aurora achieved early performance numbers using only a portion of its nodes.

HPL-MxP #1 Ranked

10.6 exaflops	AI workloads
9,500 nodes	89% of Aurora's total nodes

HPL #2 Ranked

1.012 exaflops	peak computational power
9,234 nodes	87% of Aurora's total nodes

IO500 #1 Ranked

10,066 GiB/s	bandwidth
642 servers	63% of Aurora's total servers

HPCG #3 Ranked

5,612.6 teraflops	real-world application performance
4,096 nodes	39% of Aurora's total nodes

Graph500 #5 Ranked

24,250 GTEPS	data-intensive workloads
4,096 nodes	39% of Aurora's total nodes

Pioneering Science at Exascale

ALCF, Intel, and HPE were recognized for Aurora's performance benchmark results at the 2024 ISC High Performance conference.

Large-Scale Connectomics: AI-Guided Brain Mapping

The structure of the human brain is enormously complex and not well understood. Its 80 billion neurons, each connected to as many as 10,000 other neurons, support activities from sustaining vital life processes to defining who we are. Using high-resolution electron microscopy images of brain tissue, computer vision and machine learning techniques operating at exascale can reveal the morphology and connectivity of neurons in brain tissue samples, helping to advance our understanding of the structure and function of mammalian brains.

With the arrival of Aurora, scientists are leveraging innovations in imaging, supercomputing, and AI to advance connectomics research, which is focused on mapping the neural connections in the brain. The team's techniques will enable computing to scale from cubic millimeters of brain tissue today, to a cubic centimeter whole mouse brain, and to larger volumes of brain tissue in the future.

CosmicTagger: Deep Learning-Enhanced High-Energy Particle Physics at Scale

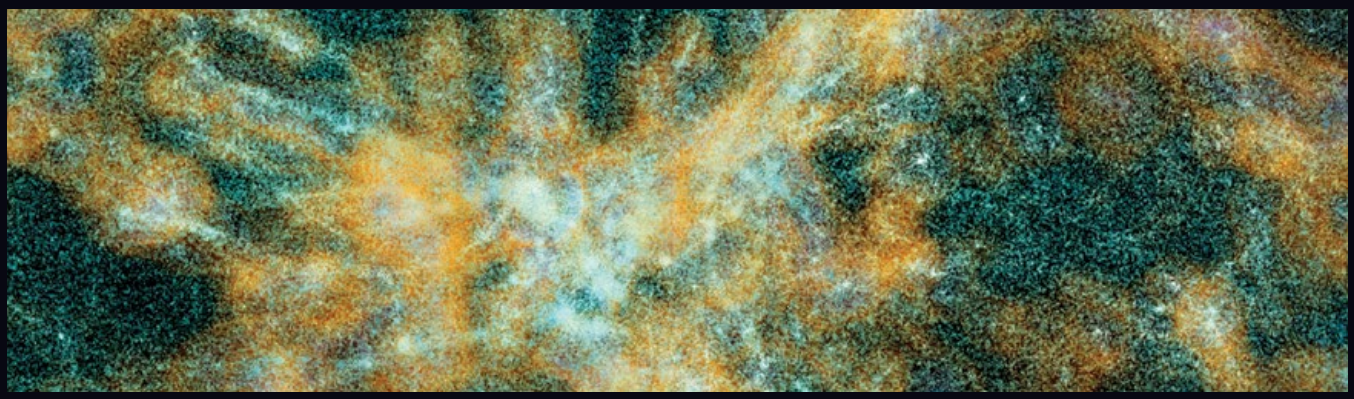
Researchers in high-energy particle physics are working to develop improved methods for detecting neutrino interactions and distinguishing them from other cosmic particles and background noise in a crowded environment. Deep learning methods, such as the CosmicTagger application, offer powerful tools for advancing neutrino physics research. Now running on Aurora, CosmicTagger is a deep learning-guided computer vision model featuring high-resolution imaging data and corresponding segmentation labels originating from high-energy neutrino physics experiments.

A key benchmark for HPC systems, CosmicTagger aims to differentiate and classify each pixel to separate cosmic pixels, background pixels, and neutrino pixels in a neutrino dataset. The application, which is enabling substantially improved background particle rejection compared to classical techniques, will aid in the analysis of large-scale datasets generated by upcoming neutrino experiments.

Accelerating Drug Design and Discovery with Machine Learning

High-throughput screening of massive datasets of chemical compounds to identify therapeutically advantageous properties is a promising step in the treatment of diseases like cancer, as well as for response to epidemics like SARS-CoV-2. However, traditional structural approaches for assessing binding affinity, such as free energy methods or molecular docking, pose significant computational bottlenecks when dealing with quantities of data of this magnitude.

Using Aurora, researchers are leveraging machine learning to screen extensive compound databases for molecular properties that could prove useful for developing new drugs. Aurora's computing power and AI capabilities will make it possible to screen 40 to 60 billion candidate compounds per hour for potential synthesis. A key future direction for the team's workflow involves integrating de novo drug design, enabling the researchers to scale their efforts to explore the limits of synthesizable compounds within the chemical space.



As part of the Aurora Early Science Program, a team is using Aurora to perform extreme-scale cosmological simulations that will advance our understanding of the universe. *Image: ALCF Visualization and Data Analytics Team; HACC Collaboration*

GAMESS: Using Quantum Simulations to Unlock the Secrets of Molecular Science

Researchers in computational chemistry require the means to carry out demanding tasks like computing the energies and reaction pathways of catalysis processes. Designed to help solve such problems using high-fidelity quantum simulations, GAMESS, or General Atomic and Molecular Electronic Structure System, is a general-purpose electronic structure code for computational chemistry. The code has been rewritten for next-generation exascale systems, including Aurora.

Using a well-defined heterogeneous catalysis problem to prepare for Aurora, GAMESS has demonstrated advanced capabilities for modeling complex physical systems that involve chemical interactions with many thousands of atoms.

HACC: Probing the Universe with Extreme-Scale Cosmological Simulations

Cosmology currently poses some of the deepest questions in physics about the nature of gravity and matter. The analysis of observations from large sky surveys requires detailed simulations of structure formation in the universe. The simulations must cover vast length scales, from small dwarf galaxies to galaxy clusters (the largest bound objects in the universe). At the same time, the physical modeling must be sufficiently accurate on the relevant scales—gravity dominating on large scales, and baryonic physics becoming important on small scales.

The Hardware/Hybrid Accelerated Cosmology Code (HACC) is a cosmological N-body and hydrodynamics simulation code designed to run at extreme scales. HACC computes the complicated emergence of structure in the universe across cosmological history. The core of the code's functionality consists of gravitational calculations along with the more recent addition of gas dynamics and astrophysical subgrid models.

HACC simulations have been performed on Aurora in runs using as many as 1,920 nodes, and visualizations of results generated on the system illustrate the large-scale structure of the universe. Deploying the HACC code at exascale will help deepen our understanding of the structure of the universe and its underlying physics.

NWChemEx: High-Performance Computational Chemistry at the Quantum Level

Producing high-quality biofuels in a sustainable and economically competitive way presents significant technical challenges. Two such challenges are designing feedstock to efficiently produce biomass and developing new catalysts to efficiently convert biomass-derived intermediates into biofuels.

The NWChemEx project, which includes quantum chemical and molecular dynamics functionality, has the potential to accelerate the development of stress-resistant biomass feedstock and energy-efficient catalytic processes for producing biofuels. Additionally, it aims to address a variety of other challenges at the forefront of molecular modeling, including development of next-generation batteries, the design of new functional materials, and the simulation of combustive chemical processes.

Aurora Early Science Program Projects

The Aurora Early Science Program is designed to prepare key scientific codes and software for the scale and architecture of the ALCF's exascale supercomputer, paving the way for other production applications to run on the system.

Simulation Projects

Extending Moore's Law Computing with Quantum Monte Carlo

PI Anouar Benali, Argonne National Laboratory
SW QMCPACK

Extreme-Scale Cosmological Hydrodynamics

PI Katrin Heitmann, Argonne National Laboratory
SW CRK-HACC, Thrust

Extreme-Scale Unstructured Adaptive CFD

PI Kenneth Jansen, University of Colorado Boulder
SW PHASTA

High-Fidelity Simulation of Fusion Reactor Boundary Plasmas

PI CS Chang, Princeton Plasma Physics Laboratory
SW XGC, Kokkos, Cabana, ADIOS2

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

PI Theresa Windus, Iowa State University and Ames Laboratory
SW NWChemEx

Data Projects

Dark Sky Mining

PI Salman Habib, Argonne National Laboratory
SW Custom analyses, containers, TensorFlow, hyperparameter optimization, HACC, CosmoTools, JAX, Numba/JIT

Data Analytics and Machine Learning for Exascale Computational Fluid Dynamics

PI Kenneth Jansen, University of Colorado Boulder
SW SmartSim, RedisAI, PHASTA, VTK, ParaView

Exascale Computational Catalysis

PI David Bross, Argonne National Laboratory
SW NRRAO, RMG, PostgreSQL, Fitpy, KinBot, Sella, Balsam, NWChemEx

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

PI Amanda Randles, Duke University
SW HARVEY, SENSEI, VTK, Kokkos

Simulating and Learning in the ATLAS Detector at the Exascale

PI Walter Hopkins, Argonne National Laboratory
SW AthenaMT, Root, workflows, containers, TensorFlow, Kokkos

Learning Projects

Accelerated Deep Learning Discovery in Fusion Energy Science

PI William Tang, Princeton Plasma Physics Laboratory
SW FusionDL/FRNN, PyTorch, TensorFlow, DeepHyper

Enabling Connectomics at Exascale to Facilitate Discoveries in Neuroscience

PI Nicola Ferrier, Argonne National Laboratory
SW TensorFlow, Horovod, Flood Filling Network

Machine Learning for Lattice Quantum Chromodynamics

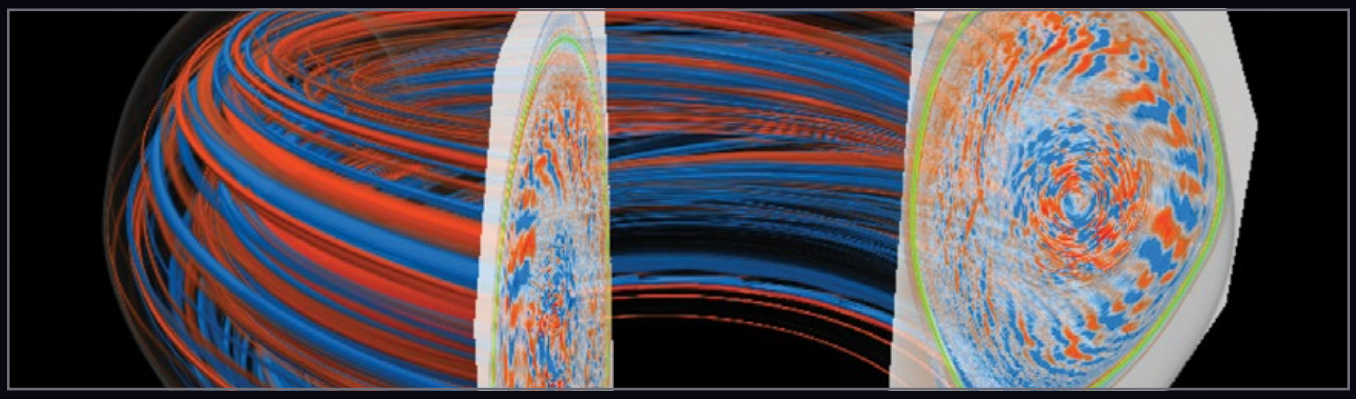
PI William Detmold, Massachusetts Institute of Technology
SW MLHMC, USQCD libraries, PyTorch, TensorFlow, HDF5, MongoDB

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

PI Noa Marom, Carnegie Mellon University
SW Balsam, SISSO, Bayesian optimization, BerkeleyGW, FHI-aims

Virtual Drug Response Prediction

PI Rick Stevens, Argonne National Laboratory
SW CANDLE, Keras, TensorFlow



Researchers are using Aurora to carry out large-scale simulations that will enhance understanding of how to control the behavior of magnetically confined fusion plasmas. *Image: Princeton Plasma Physics Laboratory*

Early simulations using NWChemEx, a ground-up rewrite of the ab initio computational chemistry software package NWChem, are being run on Aurora across hundreds of nodes.

OpenMC: High-Fidelity Monte Carlo Neutron and Photon Transport Simulations

Researchers aim to harness the power of exascale computing to model the entire core of a nuclear reactor, generating virtual reactor simulation datasets with high-fidelity, coupled physics models for reactor phenomena that are truly predictive. To this end they will employ OpenMC, a Monte Carlo neutron and photon transport simulation code that simulates the stochastic motion of neutral particles. Models used in OpenMC simulations can range in complexity from a simple slab of radiation-shielding material to a full-scale nuclear reactor.

Originally written for CPU-based computing systems, OpenMC has been rewritten for GPU-powered machines, including Aurora. The extreme performance gains OpenMC has achieved on GPUs are finally bringing within reach a much larger class of problems that historically were deemed too expensive to simulate using Monte Carlo methods.

XGC: Large-Scale Multiphysics Magnetic Fusion Reactor Simulations

DOE supercomputers play a key role in fusion research, which depends heavily on computationally intensive, high-fidelity simulations. To advance fusion research on Aurora, scientists are preparing the X-Point Gyrokinetic Code (XGC) to perform large-scale simulations of fusion edge plasmas. XGC is being developed in tandem with the Whole Device Model Application project—which aims to build a high-fidelity model of magnetically confined fusion plasmas to help plan research at the ITER fusion experiment in France.

Specializing in edge physics and realistic geometry, XGC is capable of solving boundary multiscale plasma problems across the magnetic separatrix (that is, the boundary between the magnetically confined and unconfined plasmas) and in contact with a material wall called a divertor, using first-principles-based kinetic equations. Aurora will enable a much larger and more realistic range of dimensionless plasma parameters than ever before, with the core and the edge plasma strongly coupled at a fundamental kinetic level based on the gyrokinetic equations.

Planning For Argonne's Next-Generation Exascale System

Q3 2024

Published draft technical specifications/ requirements

Technical design review

Q2 2025

CD-1 review

Q3 2025

RFP release

Q4 2025

Vendor selection

Q4 2028

System delivery

Looking beyond Aurora, the ALCF continued the efforts to prepare for its next-generation leadership supercomputer. Known as ALCF-4, the undertaking is being led by project director Jini Ramprakash and technical director Kevin Harms. The team released the draft technical specifications for the system in June, soliciting comments and questions from the HPC community and potential vendors. They will use this input to prepare the request for proposals, which will be issued in 2025.

The ALCF-4 team is targeting 2028–2029 for the deployment of the facility's next production system. The project's goals include enabling a significant improvement in application performance over Aurora, continuing to support traditional HPC workloads alongside AI and data-intensive computations, and investigating the potential to accelerate the deployment and realization of new technologies.

Revolutionizing AI for Science

The ALCF's powerful computing resources and team of experts continue to drive innovations in the use of AI for scientific research.

As AI continues to permeate many aspects of our everyday lives, it has become a national priority to develop safe and trustworthy AI capabilities to advance scientific discovery, economic prosperity, and national security.

Two major U.S. initiatives—Frontiers in Artificial Intelligence for Science, Security and Technology (FASST) and the National Artificial Intelligence Research Resource (NAIRR) Pilot—were unveiled in 2024 to expand and enhance the nation's AI capabilities.

In July, DOE detailed the roadmap for its FASST initiative, which is focused on solidifying U.S. leadership in safe and trustworthy AI systems to support its core missions in science, energy, and security. Meanwhile, the NAIRR Pilot, led by the National Science Foundation in collaboration with DOE and several partners, was launched to advance the development of a shared research infrastructure that will strengthen and democratize access to critical resources necessary to power responsible AI discovery and innovation.

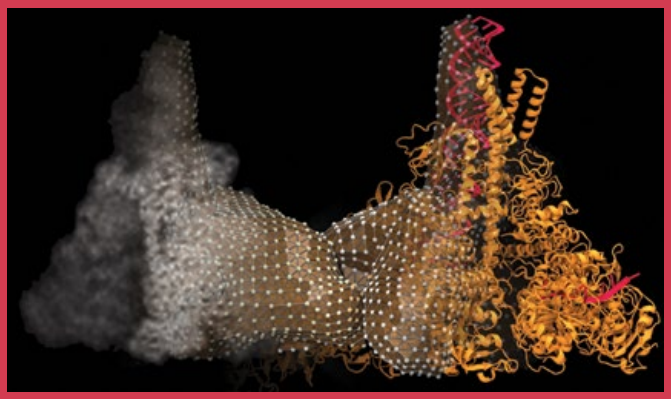
With powerful AI computing resources and extensive staff expertise, the ALCF is playing a key role in supporting the FASST and NAIRR efforts. As a leader in the use of AI for science, the facility continues to drive progress by deploying state-of-the-art computing systems including the ALCF AI Testbed and the Aurora supercomputer; supporting several innovative AI-driven research projects involving large language models (LLMs), inference, data analysis, and other techniques; and hosting outreach activities that are instrumental in advancing the nation's goal to grow an AI-ready workforce.

Building a Computing Ecosystem for AI-Driven Research

For the past decade, the ALCF has been building a computing environment that supports AI and data analysis workloads alongside more traditional modeling and simulation-based research campaigns. To that end, the facility has hosted annual Simulation, Data, and Learning Workshops to grow a new community of scientists who can employ AI and data analysis methods at a scale that requires DOE supercomputers. In addition, ALCF allocation programs, including the Aurora Early Science Program and INCITE, have been expanded to seek proposals for projects focused on using AI and data science approaches.

Argonne's Rick Stevens discusses the future of AI at the Fall 2024 HPC/AI User Forum.

Advanced Capabilities for AI Workloads



An Argonne-led team developed a novel workflow that integrates AI methods, cryo-EM data, fluctuating finite element analysis (FFEA), and all-atom molecular dynamics (AAMD) simulations to efficiently study the SARS-CoV-2 replication transcription complex. *Image: Argonne National Laboratory, University of Illinois Urbana-Champaign*

ALCF supercomputing and AI resources are equipped with unique hardware and software environments that allow them to perform a variety of AI tasks and workloads efficiently.

AI Model Training: Using large datasets to "teach" an AI model to detect patterns and make accurate, trustworthy predictions.

Inference: Employing a trained AI model to make predictions on new data.

Large Language Models (LLMs): AI models that are trained on large amounts of text data to understand, generate, and predict text-based content.

Computer Vision Models: AI models that are trained to understand and analyze visual data for tasks such as image classification and object recognition.

Foundation Models: Similar to LLMs, these AI models are trained on diverse datasets to perform a broad set of processing tasks. Foundation models, however, can serve as a starting point for developing more specialized AI models for specific domains or applications.

In 2022, the facility deployed the ALCF AI Testbed, a growing collection of some of the world's most advanced AI accelerators available for open science. Designed to enable researchers to pursue innovative AI-driven research campaigns, the testbed systems are also helping the facility to gain a better understanding of how novel AI technologies can play a role in boosting the capabilities and efficiency of next-generation supercomputers.

The ALCF's HPC systems—Aurora, Polaris, and Sophia—are all equipped with powerful GPU capabilities that support large-scale AI and data analysis tasks. Notably, the Aurora exascale supercomputer took first place on the HPL-MxP benchmark in May at the 2024 ISC High Performance conference, registering a record 10.6 exaflops of AI performance with only 89 percent of the system's total nodes.

Supporting Pioneering AI Research

These efforts have laid the groundwork for several innovative AI research projects. Scientists are employing AI and machine learning techniques on ALCF resources to accelerate research in many fields, including materials design, drug discovery, fusion energy science, and neutrino physics.

As part of a study that leveraged the ALCF AI Testbed and Polaris, for example, an Argonne-led team developed the first genome-scale language models. Known as GenSLMs, the team's models were built to study the evolution of SARS-CoV-2 (the virus that causes COVID-19). Their research demonstrated how LLMs can help scientists identify and classify new variants of SARS-CoV-2 and other viruses. The team's work was recognized with the prestigious Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research in 2022.

The ALCF AI Testbed includes novel accelerators from Cerebras, Graphcore, Groq, and SambaNova.

Experimental data analysis is another area getting a boost from the ALCF AI Testbed. Researchers from the Advanced Photon Source (APS) are exploring how different accelerators can enable fast, scalable AI model training and inference to accelerate the analysis of x-ray imaging data. Rapid data analysis methods are becoming increasingly important for the APS and other experimental facilities as data generation rates continue to grow.

In 2024, the ALCF began supporting the first round of NAIRR projects, with three teams awarded time on the AI Testbed. The NAIRR projects are using the ALCF's AI systems to develop surrogate models to speed up large-scale scientific simulations, employ federated learning to advance the use of trustworthy AI models for science, and generate a benchmark hydrologic dataset and AI model to enhance climate impact studies.

When Aurora enters production mode, several teams will use the system for large-scale AI research campaigns, including Argonne's AuroraGPT project, which is working to build a science-focused LLM. Researchers from the Aurora Early Science Program will use machine learning at exascale to pursue the discovery of new solar materials, explorations into dark matter and dark energy, and advances in connectomics. Additional projects are targeting Aurora to develop foundation models for the design of battery materials and for predictive molecular epidemiology.

Growing an AI-Ready Workforce

The ALCF has long been committed to building the next-generation STEM workforce by leading and contributing to various training and educational programs for students and early career researchers.

In the AI space, the facility launched the "Intro to AI-Driven Science on Supercomputers" series in 2021 to provide hands-on training to undergraduate and graduate students enrolled at U.S. universities and community colleges. Since its inception, the series has hosted over 600 attendees for an immersive learning experience focused on the fundamentals of using AI and supercomputers.

Beyond student outreach, the ALCF also has a robust training program aimed at assisting and expanding its user community. The facility hosts hands-on workshops and GPU hackathons to help researchers improve the performance of HPC and AI applications and workloads on its computing resources.

Over the past two years, the ALCF has hosted training workshops focused on each of the ALCF AI Testbed systems. To help researchers get started with the novel accelerators, the facility partners with the AI vendors to host two-day events that introduce attendees to system hardware and software, cover best practices, and provide hands-on guidance using the machines. The ALCF team also organizes tutorials on programming the AI accelerators for science at the annual Supercomputing conference.

ALCF Launches Lighthouse Initiative

By forging partnerships with academic institutions, the initiative aims to grow the ALCF user community and the next generation of computing professionals.

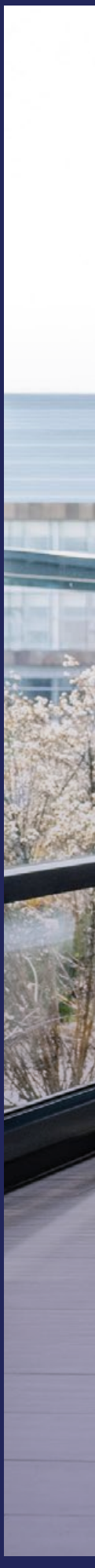
The ALCF kicked off its Lighthouse Initiative this year to build enduring partnerships with colleges and universities across the nation.

Designed to expand the ALCF's user base and foster connections with students, the initiative got underway with an initial pilot phase in partnership with the University of Chicago and the University of Illinois Chicago (UIC). The ALCF is working with the University of Chicago's Research Computing Center (RCC) and UIC's Advanced Cyberinfrastructure for Education and Research (ACER) group to prepare the universities' researchers for the ALCF's larger-scale systems, establish collaborative frameworks, and define a path forward for expanding such partnerships with other universities.

The initial partnerships will focus on familiarizing RCC and ACER staff with ALCF's comprehensive resources, which include the Polaris and Aurora supercomputers, ALCF AI Testbed systems, large-scale data storage and sharing capabilities, and staff with wide-ranging expertise in HPC. With support from the ALCF, RCC and ACER staff members will work with university researchers on projects that stand to benefit from the facility's powerful computing systems.

The ALCF will provide access to its computing resources to enable RCC and ACER researchers to port applications and evaluate the suitability of using such systems for their projects. Support from the ALCF will include introductory onboarding workshops and a dedicated point of contact to help new users maximize their time on the facility's systems. To realize its workforce development goals, the ALCF Lighthouse Initiative will also provide internship opportunities to students from partner institutions.

The ALCF aims to extend the pilot phase of the Lighthouse Initiative to additional universities while planning to broaden its reach to even more institutions in the future. As the program grows, there will be a strong focus on increasing access to researchers and students who have less exposure to large HPC facilities like the ALCF.



ALCF's Paige Kinsley (front, right) discusses the facility's Lighthouse Initiative with a team from the University of Chicago's Research Computing Center.

Integrating Supercomputers with Experiments

The ALCF continues to build on its long history of developing tools and capabilities to accelerate data-intensive science via an Integrated Research Infrastructure.

With the volume of scientific data growing at an exponential rate, the ALCF continues to play a key role in supporting DOE's vision to build an Integrated Research Infrastructure (IRI). The development of an IRI aims to accelerate discoveries by creating an environment that seamlessly connects data-intensive experimental and observational facilities with DOE's world-class supercomputing, AI, and data resources.

While the ALCF has been working in this space for over a decade, the efforts to build an IRI have gained significant momentum in recent years. DOE's Advanced Scientific Computing Research (ASCR) program began making IRI a more formal initiative with the creation of the IRI Task Force in 2020 and the IRI Blueprint Activity in 2023. ASCR is now standing up an official IRI program to invest in foundational infrastructure, coordinate existing IRI projects, and deploy an IRI Science Testbed across ASCR facilities. ALCF's Thomas Uram is serving as vice chair of ASCR's IRI program, and ALCF staff are contributing to each of the program's technical subcommittees (Outreach and Engagement, TRUSTID, and Interfaces).

The DOE IRI program has identified three initial Pathfinder projects to help propel the development of IRI-enabling technologies. These projects, in partnership with the DIII-D National Fusion Facility, DOE light sources, and the Earth System Grid Federation (ESGF), are building on established collaborations with the ALCF and its sister computing facilities to provide requirements and feedback on initial IRI plans and implementations. The DOE program plans to add additional IRI Pathfinder projects in the future.

To further advance IRI-related research, the science community has gathered at several computing events, including ESnet's Confab, Supercomputing, Monterey Data Conference, and Smoky Mountains Computational Sciences and Engineering Conference, to share and discuss successes and challenges in developing IRI capabilities.

At Argonne, the lab launched its Nexus effort in 2023, providing a vehicle to coordinate and align all of its IRI research and collaborations with DOE's broader efforts. Prior to establishing Nexus, the ALCF initiated several successful collaborations that demonstrated the efficacy of integrating its supercomputers with experiments for rapid data analysis. Coupling ALCF supercomputers with Argonne's Advanced Photon Source (APS) has been a significant focus of the lab's IRI research, but the work has also included partnerships with DIII-D, ESGF, CERN's Large Hadron Collider, and other experimental facilities.

Argonne's Advanced Photon Source is a DOE Office of Science user facility that provides ultra-bright, high-energy x-ray beams for research across almost all scientific disciplines.

These collaborations have led to the creation of new tools and capabilities for on-demand computing and managing complex workflows, giving the lab valuable experience to support the DOE IRI program. Argonne also operates several resources and services that are key to realizing the IRI vision, including:

The ALCF's Polaris and Aurora systems are powerful supercomputers with advanced capabilities for simulation, AI, and data analysis.

The ALCF AI Testbed provides researchers with access to novel AI accelerators for data-intensive tasks and AI workloads, including training, inference, large language models, and computer vision models.

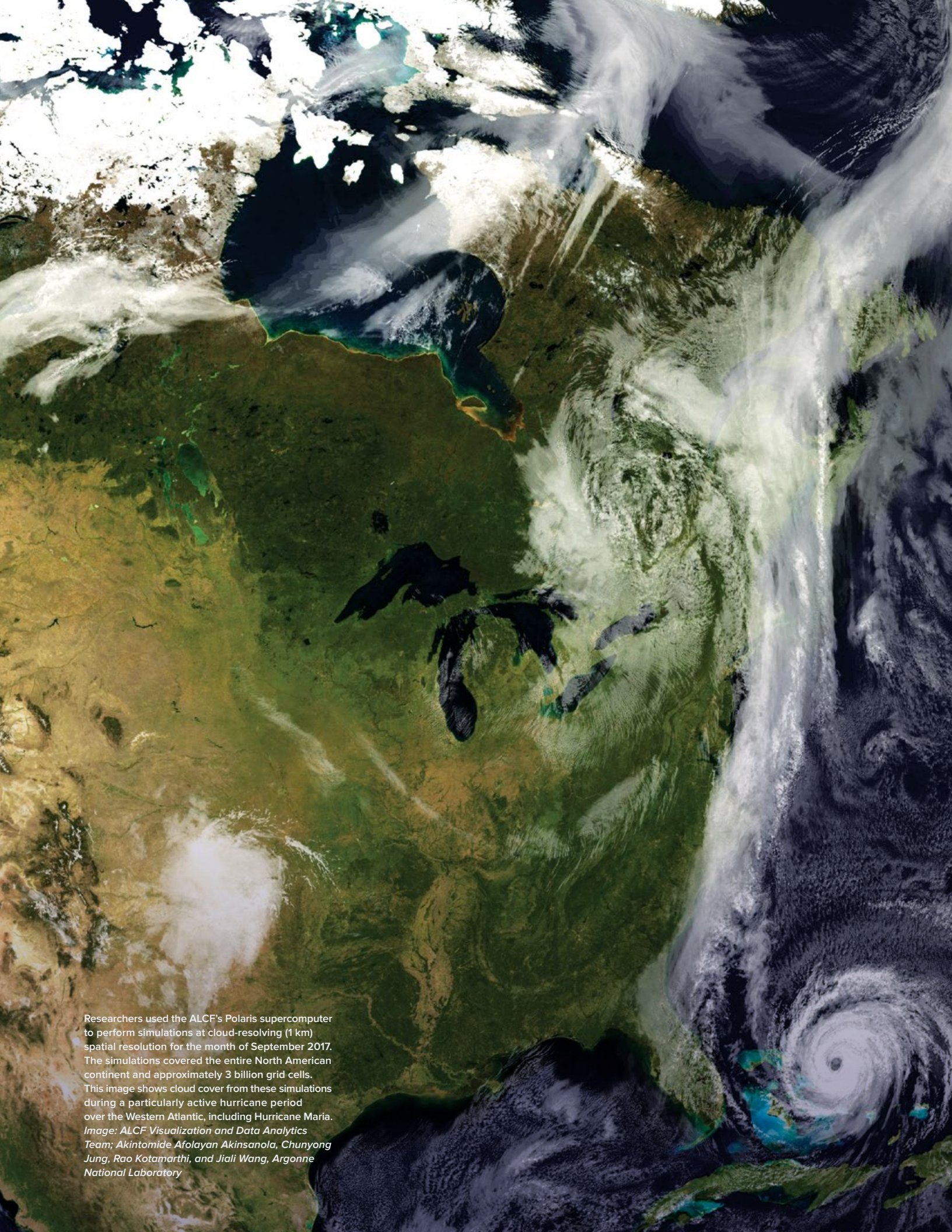
The ALCF Community Data Co-Op (ACDC) provides large-scale data storage capabilities, offering a portal that makes it easy to share data with external collaborators across the globe.

Globus, a research automation platform created by researchers at Argonne and the University of Chicago, is a not-for-profit service used to manage high-speed data transfers, computing workflows, data collection, and other tasks for experiments.

As an example of recent progress, the longstanding ALCF-APS collaboration has made significant strides in addressing the challenges of providing instant access to DOE supercomputers. The team developed and launched "service accounts," which give large experimental teams secure, shared access to ALCF supercomputers without requiring each team member to have an active ALCF user account. To overcome the wait times involved with the traditional queue system, the ALCF team implemented "on-demand" and "preemptable" queues on Polaris to ensure time-sensitive jobs have immediate access to ALCF supercomputing resources. These advancements have been successfully tested through fully automated, end-to-end runs using APS experimental data with no humans in the loop. The team is using these innovations to enable experiment-time analysis at several beamlines as operations begin to resume with the upgraded APS. As IRI work continues to progress across the DOE ecosystem, these capabilities can be extended to additional light sources as well as other data-intensive research facilities.

SCIENCE

The ALCF is accelerating scientific discoveries in many disciplines, ranging from biology and engineering to physics and materials science.



Researchers used the ALCF's Polaris supercomputer to perform simulations at cloud-resolving (1 km) spatial resolution for the month of September 2017. The simulations covered the entire North American continent and approximately 3 billion grid cells. This image shows cloud cover from these simulations during a particularly active hurricane period over the Western Atlantic, including Hurricane Maria. *Image: ALCF Visualization and Data Analytics Team; Akintomide Afolayan Akinsanola, Chunyong Jung, Rao Kotamarthi, and Jiali Wang, Argonne National Laboratory*

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 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

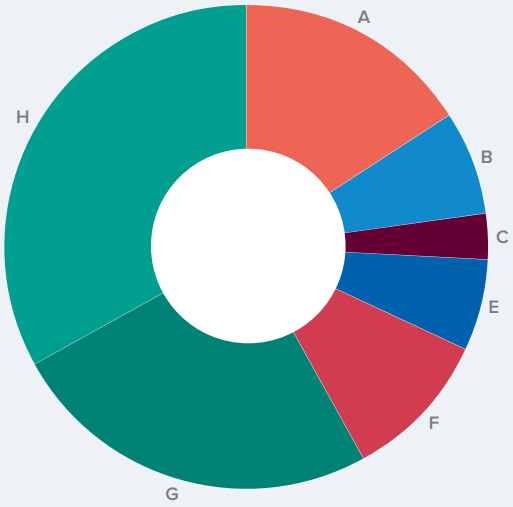
As part of the process of bringing a new supercomputer into production, the ALCF hosts its Early Science Program (ESP) to prepare applications for the architecture and scale of a new system. ESP projects represent a typical system workload at the ALCF and cover key scientific areas and numerical methods.

INCITE/ALCC BY DOMAIN

2024 INCITE

2M NODE HOURS

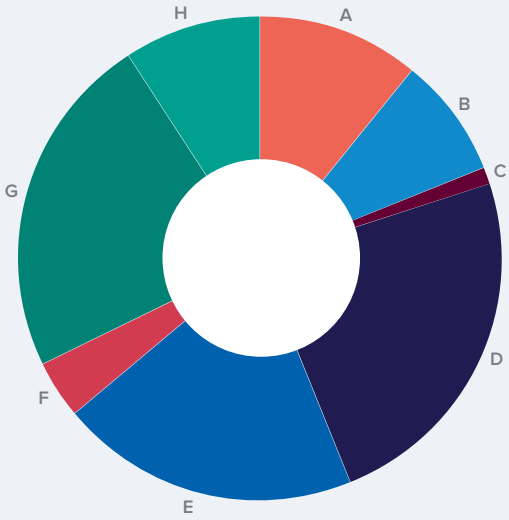
A	Biological Sciences	16 %
B	Chemistry	7
C	Computer Science	3
D	Earth Science	–
E	Energy Technologies	6
F	Engineering	10
G	Materials Science	25
H	Physics	33



2024 ALCC

1.3M NODE HOURS

A	Biological Sciences	11 %
B	Chemistry	8
C	Computer Science	1
D	Earth Science	24
E	Energy Technologies	20
F	Engineering	4
G	Materials Science	23
H	Physics	9



ALCC data are from calendar year 2024.

2024 Science Highlights

The ALCF user community continues to push the boundaries of scientific computing, producing groundbreaking studies in areas ranging from drug discovery to quantum computing.

ALCF users employ simulations, AI-driven methods, and large-scale data analyses to conduct pioneering investigations and achieve scientific breakthroughs.

From detailed atomic-level simulations to expansive cosmological studies, researchers leverage ALCF systems to probe extremely complex physical systems and processes that are too small or large, costly, or dangerous to study in a laboratory.

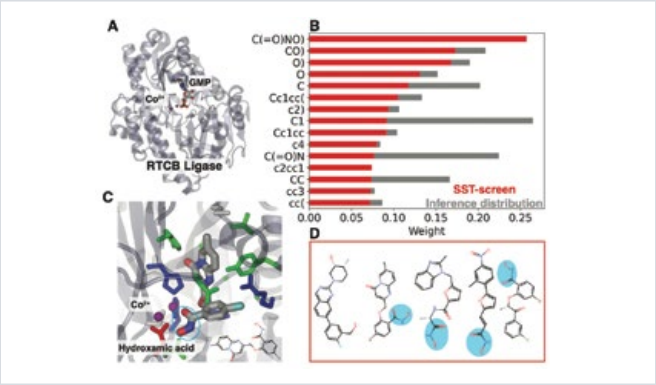
Year after year, ALCF users produce noteworthy results, whether they are developing and demonstrating novel computational methods or publishing papers in high-impact scientific journals.

In the following pages, we present a selection of science highlights from projects supported by ALCF computing resources. This year's highlights range from mapping the structure of the human brain to simulating millions of images of the cosmos to prepare for future observations from NASA's Nancy Grace Roman Space Telescope and the NSF-DOE Vera C. Rubin Observatory.

ALCF users also employed AI techniques to accelerate discoveries in the chemical compound space; developed a transformer model to enhance weather forecasting; advanced efforts to integrate supercomputing resources with data-intensive experimental facilities; and conducted large-scale simulations to gain a deeper understanding of hypersonic turbulent flows, plasma physics, and other complex phenomena.

Accelerating Drug Discovery with Deep Learning

PI Rick Stevens, Argonne National Laboratory
Archit Vasan, Argonne National Laboratory
AWARD Aurora ESP
SYSTEM Aurora



Results of a screening run performed on Aurora to target the oncoprotein RTCB ligase, illustrating SST's ability to identify compounds with key chemical motifs for binding. The compounds identified share a hydroxamic acid group common to chelate metal ions such as that in the active site of RTCB. *Image: Alexander Brace, Ozan Gokdemir, and Archit Vasan, Argonne National Laboratory*

High-throughput screening of compound datasets to identify advantageous properties represents a promising direction in drug discovery. However, traditional approaches for molecular assessment fail at the massive scales necessary for drug compound databases. To overcome this difficulty, Argonne researchers have leveraged AI to quickly and accurately identify candidate drug molecules.

CHALLENGE Traditional structural approaches for assessing the strength of the interaction between two or more molecules, a property called binding affinity, become inefficient when used to analyze data on the massive scales of drug compound databases. Such approaches include a computational simulation technique known as molecular docking. To address this, researchers have developed a docking surrogate called the Simple SMILES transformer (SST), which learns molecular features from the SMILES (Simplified Molecular Input Line Entry System) representation of compounds and approximates their binding affinity.

APPROACH SMILES data are first tokenized and then fed into a transformer model to generate vector embeddings for each molecule, effectively capturing the essential information. These extracted embeddings are subsequently fed into a regression model to predict binding affinity.

Leveraging ALCF resources, the researchers devised a workflow to scale model training and inference across multiple supercomputer nodes. To evaluate the performance and accuracy of the workflow, the team conducted experiments using molecular docking binding affinity data, comparing SST with another state-of-the-art docking surrogate.

RESULTS Drug screening inference scaled to 128 nodes on Aurora, screening approximately 11 billion drug molecules per hour, and was then increased to 256 nodes on Aurora, screening approximately 22 billion drug molecules per hour. Assuming linear scaling, researchers could expect about a trillion compounds screened per hour across the entire Aurora system.

SST showed comparable accuracy to state-of-the-art surrogate models, affirming the capability of SST to learn molecular information directly from language-based data. One significant advantage of the SST approach is its notably faster tokenization preprocessing compared to alternative preprocessing methods such as generating molecular descriptors. Furthermore, SST predictions emphasize several molecular motifs that have previously been confirmed to interact with residues in their target binding pockets.

IMPACT The team's approach presents an efficient means for screening large-scale compound databases for molecular properties useful in targeting cancer and other diseases. Aurora system capabilities will enable the screening of hundreds of billions of compounds for potential synthesis. The researchers aim to integrate de novo drug design, enabling them to explore the limits of synthesizable compounds within chemical space.

PUBLICATIONS

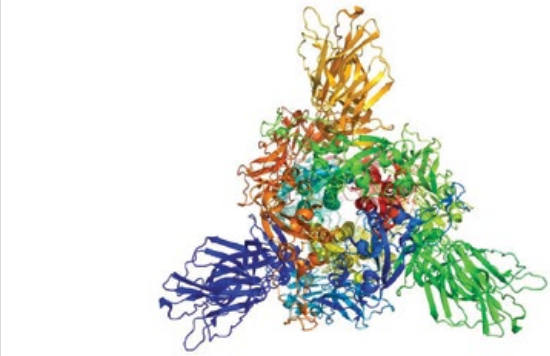
Vasan, A., T. Brettin, R. Stevens, A. Ramanthan, and V. Vishwanath. "Scalable Lead Prediction with Transformers using HPC Resources," *SC-W '23: Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network Storage, and Analysis* (November 2023), ACM. https://doi.org/10.1145/3624062.3624081

Vasan, A., O. Gokdemir, A. Brace, A, Ramanathan, T. Brettin, R Stevens, and V. Vishwanath. "High Performance Binding Affinity Prediction with a Transformer-Based Surrogate Model," *2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (May 2024), IEEE. https://doi.org/10.1109/IPDPSW63119.2024.00114

Biological Sciences | Simulation, Data, Learning

APACE: AlphaFold2 and Advanced Computing as a Service for Accelerated Discovery in Biophysics

PI Eliu Huerta, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



As one of the most important therapeutic targets for drug discovery, spike protein is subject to many drug modality developments. This figure of a SARS-CoV-2 spike protein shows trimeric receptor binding domain in a prefusion closed conformation. *Image: Eliu Huerta and Hyun Park, Argonne National Laboratory and University of Illinois Urbana-Champaign*

AlphaFold2 is a foundation AI model that is used to study proteins' genetic code and predict their 3D shape accurately. This model has led to transformational discoveries in medicine and biology. A team of scientists from Argonne and the University of Illinois Urbana-Champaign have developed APACE, a framework that optimizes AlphaFold2 to run at scale on supercomputers, significantly reducing time-to-solution.

CHALLENGE Predicting protein structures is a computationally intensive endeavor. There are 20 known different amino acids that may be used to build proteins, and the number of possible combinations of amino acids in a protein is given by 20^n, where n is the number of amino acids in a protein. Given this magnitude of possible combinations, AlphaFold2 generally takes days or weeks to accurately predict how proteins fold and function.

APPROACH To address these challenges, APACE leveraged solid-state drive data storage and Infinite Memory Engine data staging to optimally handle AlphaFold2's 2.6 terabyte database. APACE also introduced a new approach to accelerate the serial method used for computing the sequence to structure predictions. Most importantly, the team designed APACE to be readily accessible and usable by a wide range of users. APACE provides the original science capabilities of AlphaFold2, and parallelizes CPU- and GPU-intensive computations to run at scale in supercomputing environments. To test this framework, the researchers deployed APACE on two supercomputers: ALCF's Polaris and the National Center for Supercomputing Applications' Delta. In both instances, the team measured how well APACE performed in predicting the structure of four example proteins.

RESULTS Running APACE on as many as 300 ensembles distributed across 300 NVIDIA A100 GPUs, the team found that their framework is up to two orders of magnitude faster than AlphaFold2's regular implementations. Specifically, they modeled the 7MEZ protein, which plays a role in various cellular processes, including immune cell activation, chemotaxis (cell movement in response to chemical signals), and inflammation, using 300 GPUs in Polaris and Delta, reducing time-to-solution from 13 days to 2 hours.

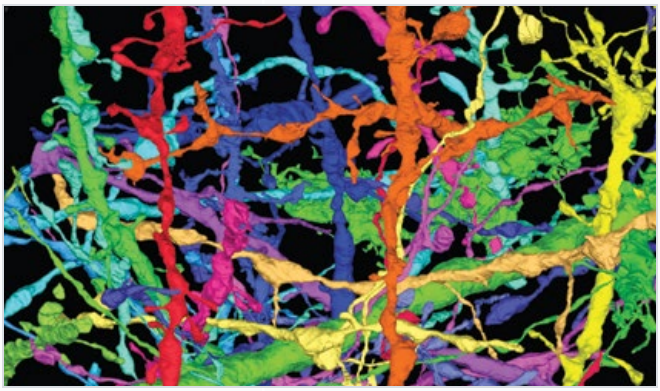
IMPACT APACE has the potential to accelerate innovation and streamline new discoveries using AlphaFold2. This has promising applications for pharmaceutical research, as it will drastically reduce the time required to screen potential candidate compounds and to develop drugs tailored toward specific viral strains. It may also be combined with robotics labs to further automate scientific discovery in biophysics.

PUBLICATIONS
Park, H., P. Patel, R. Haas, and E. A. Huerta. "APACE: AlphaFold2 and Advanced Computing as a Service for Accelerated Discovery in Biophysics," *Proceedings of the National Academy of Sciences* (June 2024), National Academy of Sciences. https://doi.org/10.1073/pnas.2311888121

Biological Sciences | Data, Learning

Brain Reconstruction at Scale: Large-Scale Connectomics with Flood-Filling Networks

PI Nicola Ferrier, Argonne National Laboratory
AWARD Aurora ESP
SYSTEM Aurora



A subset of neurons reconstructed on Aurora using the FFN convolutional neural network, from a sample of human brain tissue, based on electron microscopy images collected at Harvard. *Image: Lichtman Lab, Harvard University*

The structure of the human brain is enormously complex and not well understood. Its 80 billion neurons, each connected to as many as 10,000 other neurons, support activities that range from sustaining vital life processes to defining who we are. Using high-resolution electron microscopy images of brain tissue, computer vision and machine learning techniques operating at the exascale can reveal the morphology and connectivity of neurons in brain tissue samples, informing future studies of the structure and function of mammalian brains.

CHALLENGE Connectomics stresses many boundaries: high-throughput electron microscopy technology operating at nanometer resolution; tens of thousands of images, each with tens of gigapixels; accuracy sufficient to capture minuscule synaptic detail; computer vision methods to align corresponding structures across large images; and deep learning networks that can trace narrow axons and dendrites over large distances. Multiple applications contribute to the 3D reconstruction of neurons; the most demanding of them perform image alignment and segmentation.

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 Using the aligned image stack, a convolutional neural network model trained to identify neuron bodies and membranes helps reconstruct the 3D shape of neurons. The flood-filling network code, developed at Google and adapted to run on ALCF systems, traces individual neurons over long distances, enabling analysis at the level of synapses.

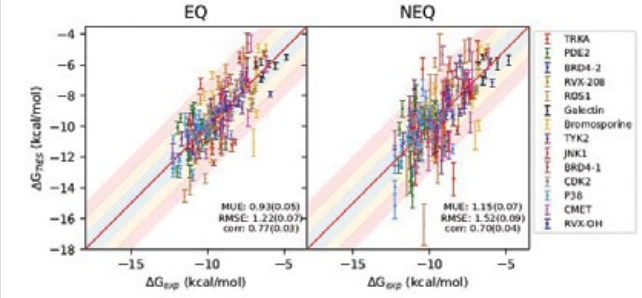
RESULTS Deep learning models for reconstruction have been trained on Aurora on up to 512 nodes, demonstrating performance increases up to 40 percent. Reconstructions have been run with these models on up to 1024 nodes on Aurora, with multiple inference processes per GPU, to produce a segmentation of a teravoxel of data. Projecting from these runs to the full machine, the researchers anticipate being able to segment a petavoxel dataset in a few days running on Aurora.

IMPACT Connectomics is leveraging innovations in imaging, supercomputing, and AI to improve our understanding of how neurons are arranged and connected. The techniques developed will enable computing to scale from cubic millimeters of brain tissue today, to a cubic centimeter whole mouse brain in the future, and to larger volumes of human brain tissue. As imaging technology advances, computing will need to achieve high performance on post-exascale machines to avoid becoming the bottleneck. The work done to prepare this project for exascale will benefit exascale users in general; the electron microscopy algorithms under development, for example, promise broad application to x-ray data, especially with the recent upgrade to Argonne's Advanced Photon Source.

Biological Sciences | Simulation, Data, Learning

COMbining deep-learning with Physics-Based affinIty estimatiOn 3 (COMPBIO3)

PI Peter Coveney, University College London
AWARD INCITE
SYSTEM Theta, Polaris



Correlation plots for absolute binding free energies from the equilibrium as well as nonequilibrium (600 ps) alchemical methods against experimental data. The solid red line denotes the perfect correlation, whereas the blue, orange and red shaded regions represent ±1, ±1-2 and ±2-3 kcal/mol ranges. *Image: University College London*

Over the past decade, scientists have made great strides in drug discovery by using advanced computer methods to understand how drug molecules interact with target proteins. Building on years of work using DOE supercomputers, a team led by researchers from University College London developed an ensemble simulation method to enable accurate and reliable predictions of complex protein-ligand binding affinities.

CHALLENGE To make robust predictions in molecular dynamics, researchers use ensemble-based enhanced sampling methods on massively parallel computers to manage the simulation complexities. Structure-based methods are widespread, but the limited value of virtual screening methods makes binding affinity calculation more promising. Pharmaceutical companies have started using relative binding free energy (RBFE) calculations, but these are limited as they require compounds and structures be similar to each other. Alchemical absolute binding free energy (ABFE) predictions are theoretically more exact. However, they are often computationally expensive, slow, hard to implement and automate, and unreliable when incorrectly implemented. Therefore, alchemical ABFE predictions are rarely performed in large-scale studies. Prior studies using this method also lack systematic assessments of the parameters involved in the complex setup those calculations.

APPROACH In their latest study, the team applied ensemble alchemical ABFE methods to a large dataset of 219 protein-ligand complexes comprising 31 different proteins and 186 compounds with a variety of scaffolds. This is the largest dataset ever used to validate an alchemical ABFE methodology. The team developed two methods for ABFE predictions. The equilibrium (EQ) approach involved performing fixed intermediate alchemical states. The nonequilibrium (NEQ) approach relied on rapidly varying the alchemical state of the system to drive it out of equilibrium. To address the substantial differences between the two end-states in an alchemical ABFE calculation, the team applied their previously developed TIES approach to control for uncertainties. The simulations were performed on ALCF and OLCF supercomputers.

RESULTS The team found that their ensemble alchemical ABFE approach generated statistically robust results with high accuracy. They compared the EQ and NEQ approaches and found that the EQ method was more accurate, faster, computationally cost effective, and required a simple protocol, making it preferable for large scale and blind applications. Based on their findings, the team developed a definitive protocol to perform ABFE calculations optimally.

IMPACT Through this project, the team developed protocols that enable thousands of alchemical ABFE calculations to be performed within two hours on exascale computers. This highly accurate, reliable, and large-scale prediction of protein-ligand binding affinities can play an important role in the drug discovery process.

PUBLICATIONS

Bhati, A. P., S. Wan, and P. V. Coveney. "Equilibrium and Non-equilibrium Ensemble Methods for Accurate, Precise and Reproducible Absolute Binding Free Energy Calculations," *ChemRxiv* (July 2024). https://doi.org/10.26434/chemrxiv-2024-sslzp-v2

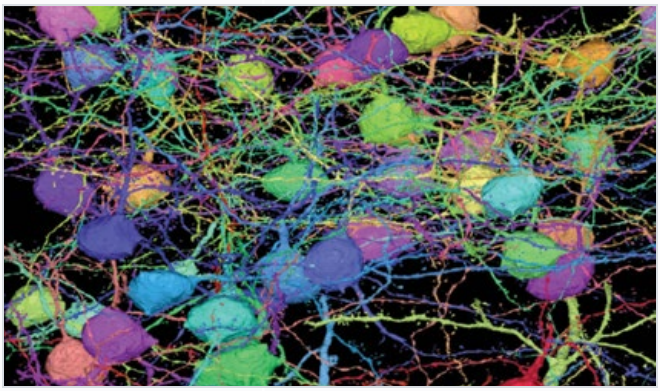
Wan, S., A. P. Bhati, and P. V. Coveney. "Comparison of Equilibrium and Nonequilibrium Approaches for Relative Binding Free Energy Predictions," *Journal of Chemical Theory and Computation* (October 2023), American Chemical Society. https://doi.org/10.1021/acs.jctc.3c00842

Wan, S., A. P. Bhati, A. D. Wade, and P. V. Coveney. "Ensemble-Based Approaches Ensure Reliability and Reproducibility," *Journal of Chemical Information and Modeling* (November 2023), American Chemical Society. https://doi.org/10.1021/acs.jcim.3c01654

Biological Sciences | Learning

Isochronic Development of Cortical Synapses in Primates and Mice

PI Gregg Wildenberg, Argonne National Laboratory and University of Chicago
AWARD Director's Discretionary
SYSTEM Theta, Cooley



3D rendering of neurons made from automatic segmentation of vEM data using ALCF computing resources. *Image: Hanyu Li, Argonne National Laboratory*

Traditionally, brain development was understood as happening faster in mice than in other, longer-living mammals such as primates and humans. However, researchers from Argonne and the University of Chicago discovered that mouse and primate brains mature at similar rates after birth, suggesting genetic events, rather than experience, drive these developments.

CHALLENGE Synaptic and neuronal development in mammals has been well studied, but the approaches used each have varying false positive and negative rates. Consequently, quantitatively measuring developmental trajectories across different mammals remains difficult. This has likely contributed to varying results over the nature of synaptic development in model systems. To address this issue, the team found that they would need to map the brain's neurons using methods that work across species and could directly visualize changes in synaptic connections. This procedure would require the use of supercomputers to collect, analyze, and process the significant volume of generated data.

APPROACH The team conducted a complete evaluation of synaptic development between macaques and mice to observe any differences in the rate and timing of post-natal synaptic development. They used large volume serial electron microscopy (vEM), a gold standard for detailing neuronal connections, to compare development timelines. Using a software pipeline and flood-filling network architecture designed in previous research, the team reconstructed excitatory and inhibitory connections onto excitatory neurons from multiple cortical regions, layers, and across multiple time points and animals. The team used the ALCF's Cooley and Theta systems to process the images through 2D montage aging and to create 3D EM volumes. They used AI-based image algorithms to segment the data, tracing individual neurons and identifying their connections. This process allowed the team to trace the brain cells at a nanoscale level, looking at every neuron and counting every synapse for each animal.

RESULTS In a paper published in *Nature Communications*, the team showed that synapse development in mice and primates follows the same absolute timescale, regardless of species, challenging what other models have typically predicted for long-lived, highly cognitive species like primates. Their results suggest a universal developmental clock across mammalian species and their cortical areas. They also found key differences in excitatory and inhibitory synapse development between the species, highlighting a potentially important distinction in lifelong brain development.

IMPACT This research challenges assumptions about aging and brain development. It calls into question research results that have used young mouse brain data to understand various human conditions. Future studies that apply the team's analyses could lead to new insights into brain evolution that help advance the development of better pharmaceutical treatments for human neurological diseases.

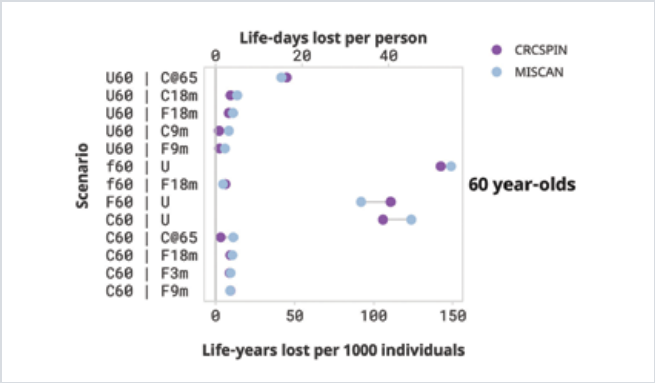
PUBLICATIONS

Wildenberg, G., H. Li, V. Sampathkumar, A. Sorokina, and N. Kasthuri. "Isochronic Development of Cortical Synapses in Primates and Mice," *Nature Communications* (December 2023), Springer Nature. https://doi.org/10.1038/s41467-023-43088-3

Biological Sciences | Simulation

Probabilistic Comparative Modeling of Colorectal Cancer Screening Strategies

PI Jonathan Ozik, Argonne National Laboratory
AWARD ALCC
SYSTEM Theta



Screening benefits lost due to disruptions by cohort and scenario for two microsimulation models and for unscreened (U60), colonoscopy screening-adherent (C60), and fecal immunochemical test screening-adherent (F60) 60-year-olds. *Image: Pedro Nascimento de Lima, RAND Corporation*

The COVID-19 pandemic has had far-reaching health repercussions worldwide. One notable impact has been a sharp decline in cancer screening rates, including for colorectal cancer (CRC), which remains the second-leading cause of cancer deaths in the United States. To investigate the effects of these screening disruptions, a multi-institutional team of researchers leveraged ALCF supercomputers to run CRC models to estimate their impact on long-term cancer outcomes.

CHALLENGE Despite cancer screening reopening efforts, CRC screening has not yet returned to pre-pandemic levels. The pandemic continues to affect CRC screening and diagnosis through staff shortages that reduce capacity at gastroenterology clinics and patient hesitancy to seek care. The pandemic may also further exacerbate existing disparities related to screening. The burden of unemployment and loss of access to healthcare varies across different racial and ethnic groups, which could contribute to widening disparities in cancer outcomes.

APPROACH With help from ALCF computing resources, a team of researchers from Argonne National Laboratory, RAND Corporation, Erasmus Medical Center, Fred Hutchinson Cancer Center, and Memorial Sloan Kettering Cancer Center used two independently developed microsimulation models—CRC-SPIN and MISCAN-Colon—to estimate the effects of pandemic-induced disruptions in colonoscopy screening for eight pre-pandemic average-CRC risk population cohorts. The team leveraged the ALCF's Theta supercomputer to calibrate the CRC-SPIN model using the Incremental Mixture Approximate Bayesian Computation (IMABC) method. Each Theta node could run 64 concurrent CRC-SPIN models, with jobs consisting of large, space-filling parameter samples and longer iterative parameter space sampling. The researchers evaluated three channels through which screening was disrupted: delays in screening, regimen switching, and screening discontinuation. The impact on long-term CRC outcomes was measured by comparing the number of life-years lost due to screening disruptions with a no-disruption scenario.

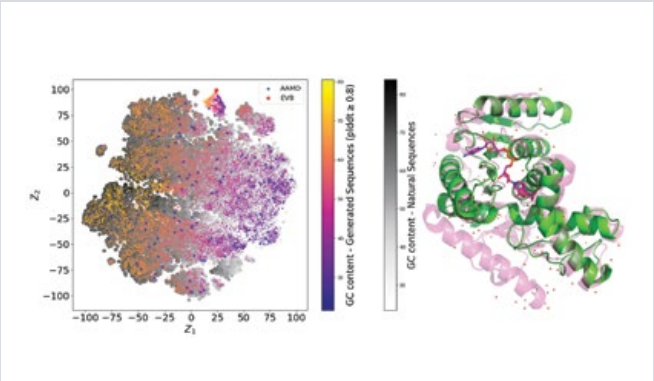
RESULTS The team examined a total of 25 scenarios based on different population cohorts (e.g., 50-, 60-, and 70-year-olds who did or did not adhere to screening) that post-pandemic experienced no disruptions, some delays, or discontinued screening. While short-term delays in screening of 3–18 months are predicted to result in minor decreases in life expectancy, discontinuing screening resulted in much more significant decreases. The team's findings demonstrate that unequal recovery of screening following the pandemic can further widen disparities. The worst-case scenario considered was that of 50-year-olds who postponed screening until the age of 65 when they became Medicare eligible, whereas other disruption scenarios for this group are predicted to have minor effects.

IMPACT The team's research highlights the potential harm caused by disruptions in cancer screening due to the COVID-19 pandemic. By analyzing different age groups and screening statuses, their study underscores how discontinuing screening could reduce life expectancy, emphasizing the importance of ensuring equitable recovery to screening to prevent further disparities.

PUBLICATIONS
Nascimento de Lima, P., R. van den Puttelaar, A.I. Hahn, M. Harlass, N. Collier, J. Ozik, A.G. Zauber, I. Lansdorp-Vogelaar, and C.M. Rutter. "Projected Long-Term Effects of Colorectal Cancer Screening Disruptions Following the COVID-19 Pandemic," *eLife* (May 2023), eLife Sciences Publications, Ltd. https://doi.org/10.7554/eLife.85264

Protein Generation via Genome-Scale Language Models with Biophysical Scoring

PI Arvind Ramanathan, Argonne National Laboratory
Venkatram Vishwanath, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



The team's workflow shows that embeddings learned by the model can recapture not only sequences in the training data, but also generate novel sequences that are out of distribution (gray dotted rectangle). *Image: G. Dharuman et. al., Proc. SC 23 Workshops Int. Conf. High Performance Computing (2023), ACM.*

Recent developments in machine learning, particularly large language models (LLMs), have created wide-ranging opportunities to drive innovations in biological research. To advance the use of AI for protein design, an Argonne-led research team used ALCF computing resources to develop a workflow that integrates LLMs with multi-level scoring systems, enabling the generation of viable protein variants.

CHALLENGE Designing new proteins with specific properties is particularly challenging due to downstream interactions that may impact a design's viability. Current protein engineering approaches require exploring a complex and large design space in which millions of mutations must be evaluated. While leveraging LLMs helps address this challenge, current generative AI models suffer from "hallucinations," which generate invalid sequences. The researchers found that a workflow for addressing these hallucinations would require multiple levels of evaluation for the stability and activity of the proteins. To make the application cohesive, the workflow would require a steering logic to coordinate this complex process.

APPROACH Leveraging the ALCF's Polaris supercomputer, the team developed a new AI-driven workflow that generates novel gene sequences with downstream evaluation through domain-specific scoring and computationally intensive simulations that can mimic laboratory measurements. The workflow enables scoring at two levels: fast scoring from folding metrics and slow scoring from all-atom molecular dynamics simulations. This approach ensured maximum utilization of available computing resources. The researchers developed the workflow so that the application would be cohesive, weaving inference with LLMs through generative and evaluative sequences. They employed the open-source Colmena and Parsl libraries to steer the application and coordinate the execution of simulations across thousands of GPUs.

RESULTS The approach was tested using a pre-trained genome-scale language model (GenSLM) to generate new sequences of malate dehydrogenase. The team previously developed GenSLMs to capture long-range sequence context and generate accurate SARS-CoV-2 sequences. In test runs on Polaris, the workflow generated sequences that captured sufficient diversity while demonstrating better catalytic activity, suggesting biophysical rewards can implicitly guide productive enzyme designs. Going forward, the researchers believe that incorporating experimental preferences as feedback, combined with techniques like reinforcement learning, would further optimize the generation of new sequences.

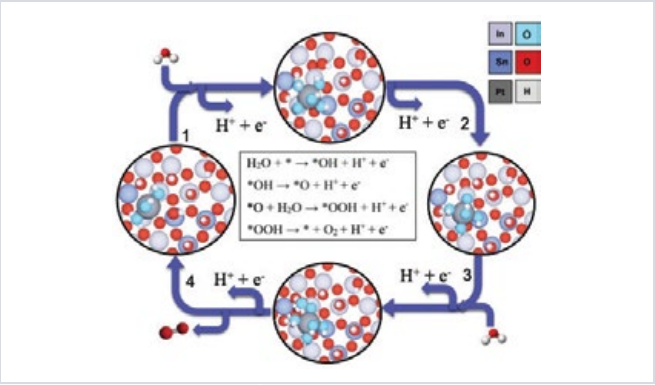
IMPACT This workflow marks the first known end-to-end system for identifying novel proteins or gene sequences that relies on computational catalytic calculations to mimic experimental data. This protein engineering process has significant implications for a variety of applied fields, including biomedicine, biomanufacturing, and synthetic biology.

PUBLICATIONS
Dharuman, G., L. Ward, H. Ma, P. V. Setty, O. Gokdemir, S. Foreman, M. Emani, K. Hippe, A. Brace, K. Keipert, T. Gibbs, I. Foster, A. Anandkumar, V. Vishwanath, and A. Ramanathan. "Protein Generation via Genome-scale Language Models with Biophysical Scoring," *Proceedings of the SC '23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis* (2023). Association for Computing Machinery. https://doi.org/10.1145/3624062.3626087

Chemistry | Simulation

Elucidation of Catalytic Site for Water Electrolysis

PI Philippe Sautet, University of California, Los Angeles
AWARD INCITE
SYSTEM Polaris, Theta



The four stages of the oxygen evolution reaction mechanism in water electrolysis. Each releases the equivalent of one H atom; they combine to form H2. Two water molecules are consumed, and one molecule of O2 is released per cycle. The rate determining step (3) involves OH bonding to O adsorbed on Pt.
Image: Philippe Sautet, University of California, Los Angeles

Single-atom catalysts are of great interest due to their high catalytic activity and selectivity, but the nature of their active sites under realistic reaction conditions is not well understood. A team of researchers led by University of California, Los Angeles developed a rigorous methodology for predicting the composition and structure of an active site under reaction conditions. The team applied their method to the electrolysis of water catalyzed by platinum (Pt) atoms.

CHALLENGE Undergirding this work is the notion that a catalytic interface in the steady state is in constant motion enabled by reaction conditions (temperature and pressure of gases in thermal catalysis; or electrochemical potential, solvent, and acidity in electrocatalysis). Due to these dynamics, the interface presents a fluxional ensemble of many states and active sites, each characterized by its specific activity, selectivity, deactivation propensity, and operando spectral signatures. Catalysis, therefore, is a collective ensemble phenomenon largely driven by highly active metastable states.

APPROACH Using ALCF supercomputing resources, the researchers rely on density functional theory calculations and grand canonical basin-hopping to theoretically investigate the active site for the oxygen evolution reaction—including the influence of the electrochemical potential—on a single Pt atom supported on an indium tin oxide surface.

RESULTS As detailed in *The Journal of Physical Chemistry Letters*, the predicted electrochemical potential was in excellent agreement with experimental results, yielding structural insights such as the role of adsorbed hydroxyl groups and the oxidation states of the Pt atom at each step in the oxygen evolution reaction. The team identified the hydroxylation of the oxygen atom adsorbed on Pt as the rate-determining step. The large difference between the non-optimal and optimized active site electrochemical potentials validates the overall approach taken.

IMPACT This work enhances our understanding of single-atom catalysts for the oxygen evolution reaction. Because their predictive method strongly agreed with experimental results, the researchers have gained deep insights into the structure and composition of active sites. Moreover, the hydrogen released by the electrolysis of water can be used to sequester carbon dioxide.

PUBLICATIONS
Kumari, S., and P. Sautet. "Elucidation of the Active Site for the Oxygen Evolution Reaction on a Single Pt Atom Supported on Indium Tin Oxide," *The Journal of Physical Chemistry Letters* (March 2023), ACS Publications. https://doi.org/10.1021/acs.jpclett.3c00160

Cheng, D., A. N. Alexandrova, and P. Sautet. "H-Induced Restructuring on Cu(111) Triggers CO Electroreduction in an Acidic Electrolyte," The Journal of Physical *Chemistry Letters* (January 2024), *ACS Publications*. https://doi.org/10.1021/acs.jpclett.3c03202

Poths, P., S. Vargas, P. Sautet, and A. N. Alexandrova. "Thermodynamic Equilibrium versus Kinetic Trapping: Thermalization of Cluster Catalyst Ensembles Can Extend Beyond Reaction Time Scales," *ACS Catalysis* (March 2024), ACS Publications. https://doi.org/10.1021/acscatal.3c06154

Chemistry | Simulation, Data

Exascale Catalytic Chemistry

PI David Bross, Argonne National Laboratory
AWARD Aurora ESP
SYSTEM Aurora, Polaris

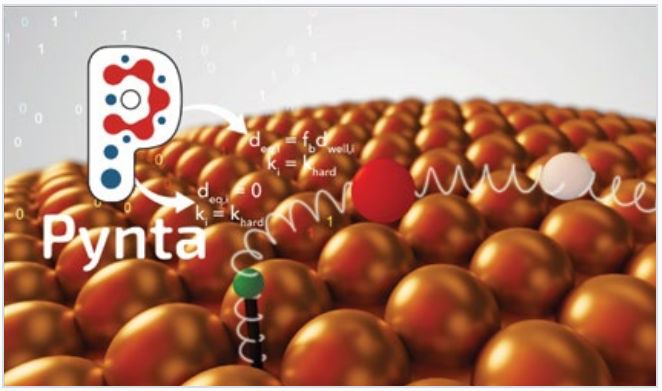


Figure of representative molecules calculated on the surface of Cu(111). *Image: David Bross, Argonne National Laboratory*

Catalysts are at the heart of industrial chemistry, and industrial chemistry is at the heart of the production of cleaner, more abundant energy—making atomic- and molecular-level catalysis research vitally important. To make easier the process of identifying how to efficiently produce high yields of valuable chemicals using exascale computing power, a team of researchers developed Pynta, a workflow software that automates necessary calculations.

CHALLENGE This work aims to facilitate and significantly speed up the quantitative description of crucial coupled heterogeneous catalyst/gas-phase chemical systems. Given all possible catalysts and conditions of interest, it is impractical to optimize most catalytic systems and processes experimentally. Automatically generated microkinetic models can be used to efficiently consider many catalysts and conditions but require accurate estimation of many thermochemical and kinetic parameters, whose manual calculation is tedious and error-prone. Pynta automates the calculation of surface and gas-surface reaction parameters.

APPROACH Pynta was developed to fully harness ALCF supercomputing resources. The researchers work closely with Argonne and Intel experts to ease bottlenecks and finetune performance issues. Pynta takes the reactants, products, and atom maps for given reactions; generates sets of initial guesses for all species and saddle points; runs all optimizations, frequency, and IRC calculations; and computes the associated thermochemistry and rate coefficients. It can consider all unique low-energy adsorption configurations for both adsorbates and saddle points, allowing it to handle high-index surfaces and bidentate species. Pynta implements a new saddle point guess generation method called harmonically forced saddle point searching (HFSP). HFSP defines harmonic potentials based on the optimized adsorbate geometries and which bonds are breaking and forming that allow initial placements to be optimized using the GFN1-xTB semiempirical method to create reliable saddle point guesses. This method is reaction class agnostic and fast, allowing Pynta to consider all possible adsorbate site placements efficiently.

RESULTS As detailed in *Journal of Chemical Information and Modeling*, the researchers demonstrated Pynta on 11 diverse reactions involving monodenate, bidentate, and gas-phase species; many distinct reaction classes; and both a low- and a high-index facet of Cu. Their results suggest that it is very important to consider reactions between adsorbates adsorbed in all unique configurations for interadsorbate group transfers and reactions on high index surfaces.

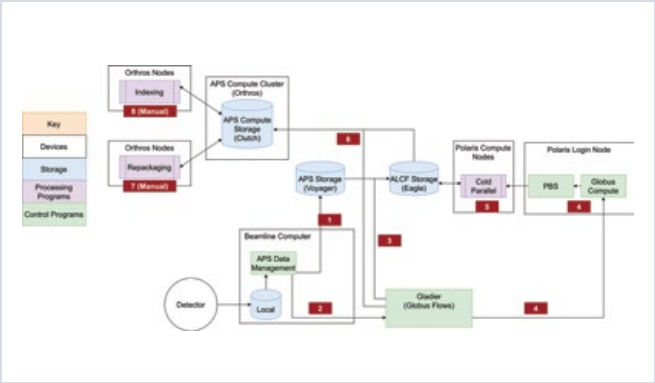
IMPACT Running the team's automated workflows coupled with accurate quantum chemical calculations on exascale systems will enable researchers to examine catalytic systems thoroughly and systematically. This work has the potential to help enable revolutionary advances in predictive catalysis crucial to addressing DOE grand challenges, including both energy storage and chemical transformations.

PUBLICATIONS
Johnson, M. S., M. Gierada, E. D. Hermes, D. H. Bross, K. Sargsyan, H. N. Najm, and J. Zádor. "Pynta–An Automated Workflow for Calculation of Surface and Gas–Surface Kinetics," *Journal of Chemical Information and Modeling* (August 2023), ACS Publications. https://doi.org/10.1021/acs.jcim.3c00948

Computer Science | Data

Demonstrating Cross-Facility Data Processing at Scale with Laue Microdiffraction

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



An overview of the systems and elements involved in the data processing system. *Image: Michael Prince, Argonne National Laboratory*

The Advanced Photon Source (APS) upgrade will provide x-ray beams 500 times more powerful than previously available, with commensurate increases in the amounts of experimental data generated for processing. To meet the demands of the upgrade, a team of researchers led by Argonne National Laboratory developed high-performance software tools and data infrastructure to integrate APS data processing and analysis in near real-time with ALCF computing resources.

CHALLENGE The co-location of the ALCF and the APS offers an ideal proving ground for methods to closely integrate supercomputers and experiments for experiment-time data analysis. The research team deployed a fully automated pipeline that uses ALCF resources to rapidly process data obtained from x-ray experiments at the APS. To demonstrate the capabilities of the pipeline, the team carried out a study using a technique called Laue microdiffraction, which is employed at the APS and other light sources to analyze materials with crystalline structures but requires significant computational resources.

APPROACH As a focused x-ray beam passes through a material, individual crystallites along the path of the beam diffract at different angles depending on their orientation. To obtain a full 3D map of the structure, the angle and position of each diffracted beam must be resolved. A new coded aperture Laue reconstruction algorithm is used instead of using a time-intensive scan to complete the analysis, necessitating supercomputers.

The automatic pipeline the team built for processing APS data leveraged infrastructure and tools being deployed between APS and ALCF as part of the Argonne Nexus effort. Globus handles much of the cross-facility data management. The APS Data Management System integrates with Globus Gladier/FuncX workflow tools to provide a single end-to-end data pipeline.

RESULTS With ALCF resources, the team demonstrated the on-demand reconstruction of data obtained from the APS beamline, returning reconstructed scans to the APS within 15 minutes of them being sent to the ALCF. Near-ideal scaling for the workflow has been shown on as many as 100 nodes. Continuously using up to 50 nodes on Polaris, it was able to keep up with the data generation rate processing scan, which came in every 1 to 2 minutes throughout 6-to-12 hour runs. The work resulted in a Best Paper Award at the SC23 conference's XLOOP workshop.

IMPACT The team's results carry implications for future software development, engineering, beamline science, and the creation of a broader Integrated Research Infrastructure. The parallelization, optimization, and deployment onto Polaris of the Laue microdiffraction technique has enabled full-scale analysis of microdiffraction data. The automated near-real-time reconstruction of coded aperture Laue datasets will enable users of ALCF supercomputing resources to collect data at speeds ten times faster than is currently possible, thereby accelerating the pace of scientific discovery. Furthermore, the full-scale reconstructions produced with Polaris are being used to improve the underlying beamline technique.

PUBLICATIONS
Prince, M., D. Gürsoy, D. Sheyfer, R. Chard, B. Côté, H. Parraga, B. Frosik, J. Tischler, and N. Schwarz. "Demonstrating Cross-Facility Data Processing at Scale with Laue Microdiffraction," *SC-W 2023: Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis* (November 2023), ACM. https://doi.org/10.1145/3624062.3624613

Efficient Algorithms for Monte Carlo Particle Transport on AI Accelerator Hardware

PI John Tramm, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM AI Testbed

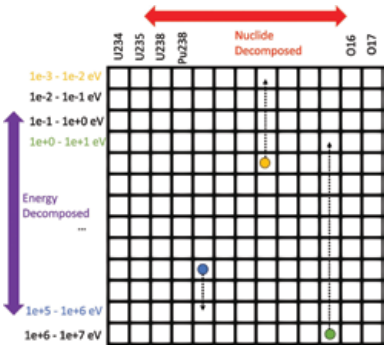


Diagram showing the energy column sorting process. Each square represents a single processing element in the WSE-2 grid, with each dot representing a particle. *Image: John Tramm, Kazutomo Yoshii, and Andrew Siegel, Argonne National Laboratory; Bryce Allen, Argonne National Laboratory and University of Chicago; Leighton Wilson, Cerebras Systems*

In this study, researchers led by Argonne National Laboratory examined the feasibility of performing continuous energy Monte Carlo particle transport on the ALCF AI Testbed's Cerebras Wafer-Scale Engine 2 (WSE-2). The researchers ported a key kernel from the Monte Carlo transport algorithm to the Cerebras Software Language (CSL) programming model and evaluated the performance of the kernel on the Cerebras WSE-2.

CHALLENGE Beyond the challenge of porting the kernel into the low-level CSL programming model, the team proposed and tested various new algorithms to handle the decomposition of neutron cross-sectional data (which is used to generate random samples for particle behavior) into the small local memory domains contained in each of some 750,000 WSE-2 units.

APPROACH The researchers ported a simplified version of the Monte Carlo cross-section lookup kernel (a kernel used by the Monte Carlo neutral particle transport algorithm) using the Cerebras SDK and the Cerebras CSL programming model. Their decomposition and communication scheme involved three stages: (1) the sorting of particles into energy bands within each column of compute cores; (2) an iterative diffusion-based load balancing stage for balancing starting particle loads within each row; and (3) an exchange of particles to allow particles to accumulate nuclide information from each column in the row. All communication patterns had to be developed to avoid any concept of global synchronization or point-to-point message passing, given the limitations of the WSE-2 hardware. Additionally, the team developed an architecture-specific optimization to leverage the capabilities of the WSE-2 and a highly optimized CUDA kernel for testing on an NVIDIA A100 GPU to provide a baseline to contextualize the performance of the WSE-2.

RESULTS A single WSE-2 was found to run 130 times faster than the highly optimized CUDA version of the kernel deployed on a single NVIDIA A100—significantly outpacing expected performance increase, given the difference in transistor counts between the architectures. However, the performance gains came at a cost—namely, increases in both software programming and algorithmic complexity. Considering how AI accelerators such as the WSE-2 were designed almost exclusively around deep learning AI tasks, it is noteworthy that the WSE-2 is already able to exceed performance expectations relative to GPUs—an architecture that has had several decades to mature and is now quite friendly to HPC simulation applications. A follow-up study saw the WSE-2 achieve a 182x speedup over the A100.

IMPACT The team's analysis suggests the potential for a wide variety of complex and irregular simulation methods to be mapped efficiently onto AI accelerators like the Cerebras WSE-2. Monte Carlo simulations themselves offer the potential to fill in crucial gaps in experimental and operational nuclear reactor data.

PUBLICATIONS

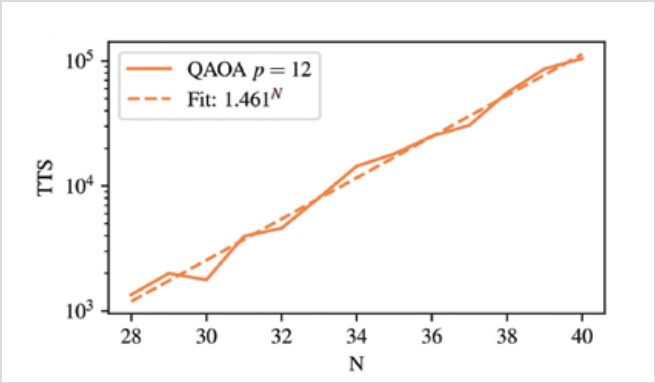
Tramm, J., B. Allen, K. Yoshii, A. Siegel, and L. Wilson. "Efficient Algorithms for Monte Carlo Particle Transport on AI Accelerator Hardware," *Computer Physics Communications* (December 2023), Elsevier BV. https://doi.org/10.1016/j.cpc.2023.109072

Tramm, J., B. Allen, K. Yoshii, and A. Siegel. "Monte Carlo with Single-Cycle Latency: Optimization of a Continuous Energy Cross Section Lookup Kernel for AI Accelerator Hardware," *International Conference on Physics of Reactors* (April 2024), American Nuclear Society, San Francisco, CA.

Computer Science | Simulation

Evidence of Scaling Advantage for the Quantum Approximate Optimization Algorithm on a Classically Intractable Problem

PI Yuri Alexeev, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



Time to solution of QAOA at layers p = 12 vs. size of the problem N. Clear exponential scaling is observed. *Image: Yuri Alexeev, Argonne National Laboratory*

The quantum approximate optimization algorithm (QAOA) is a leading candidate for solving optimization problems on quantum computers. However, the potential of QAOA to tackle classically intractable problems remains unclear. In this work, a team from Argonne, JPMorgan Chase, and Quantinuum leveraged ALCF computing resources and a quantum computer to demonstrate clear evidence of a quantum algorithmic speedup for the QAOA when applied to the low autocorrelation binary sequences (LABS) problem, which has significance in understanding the behavior of physical systems, signal processing, and cryptography.

CHALLENGE While QAOA has been studied extensively, little is known about its potential to provide a scaling advantage over classical solvers. Adding to the complexity of such research, both classical solvers and QAOA are heuristic. As a result, the only way to compare them is through empirical testing using state-of-the-art quantum simulation software and supercomputers.

APPROACH The team examined whether a quantum algorithm with low implementation costs could provide a quantum speedup over the best-known classical methods. QAOA was applied to the LABS problem, which is classically intractable even for moderately sized instances. To evaluate the quantum algorithm's performance in an ideal, noiseless setting at scale, the researchers used the ALCF's Polaris supercomputer to develop a simulator to carry out large-scale quantum circuit simulations.

RESULTS The team found that the time to solution of QAOA with number of layers $p = 12$ grows as 1.46^N, which is improved to 1.21^N if combined with quantum minimum finding. This scaling is better than that of the best classical heuristic, which has a time to solution that grows as 1.34^N. As an initial step toward practical realization of the speedup in the algorithm, the team demonstrated a small-scale implementation on Quantinuum's System Model H1 and H2 trapped-ion quantum computers. Using algorithm-specific error detection, they reduced the impact of errors on performance by up to 65 percent. The combination of QAOA with quantum minimum finding yields the best empirical scaling of any classical algorithm for the LABS problem. The team's work showed experimental progress in executing QAOA for LABS using the error detection scheme, providing evidence for QAOA's potential to enable quantum speedups.

IMPACT This project provides strong numerical and experimental evidence that QAOA, especially when enhanced with quantum minimum finding, can outperform classical solvers for the LABS problem. Their research highlights the potential for quantum speedups in optimization problems and the need for further improvements in quantum hardware and error correction techniques.

PUBLICATIONS

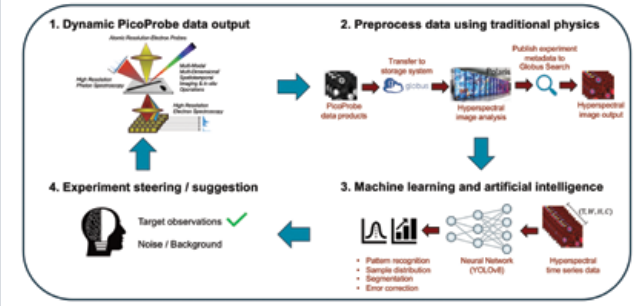
Shaydulin, R., C. Li, S. Chakrabarti, M. DeCross, D. Herman, N. Kumar, J. Larson, D. Lykov, P. Minssen, Y. Sun, Y. Alexeev, et al. "Evidence of Scaling Advantage for the Quantum Approximate Optimization Algorithm on a Classically Intractable Problem." *Science Advances* (May 2024), AAAS. https://doi.org/10.1126/sciadv.adm6761.

Lykov, D., R. Shaydulin, Y. Sun, Y. Alexeev, and M. Pistoia. "Fast Simulation of High-Depth QAOA Circuits," *Proceedings of the SC23 Workshops of The International Conference on High-Performance Computing, Network, Storage, and Analysis* (November 2023), ACM. https://doi.org/10.1145/3624062.3624216

Herman, D., C. Googin, X. Liu, Y. Sun, A. Galda, I. Safro, M. Pistoia, and Y. Alexeev. "Quantum Computing for Finance," *Nature Reviews Physics* (July 2023), Springer Nature. https://doi.org/10.1038/s42254-023-00603-1

Linking the Dynamic PicoProbe Analytical Electron-Optical Beamline/Microscope to Supercomputers

PI Arvind Ramanathan, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



A high-level vision to support computationally mediated science at the Dynamic PicoProbe. *Image: Alexander Brace, Argonne National Laboratory*

To help researchers keep pace with the growing deluge of scientific data, Argonne National Laboratory continues to develop tools and capabilities to connect its powerful supercomputing resources with data-intensive experiments via an Integrated Research Infrastructure. In a recent study, Argonne researchers created and demonstrated an end-to-end software infrastructure that links the lab's Dynamic PicoProbe electron microscope to the ALCF's Polaris supercomputer for rapid data analysis and cataloging.

CHALLENGE As experimental capabilities continue to improve, data generation rates will only continue to grow. The Dynamic PicoProbe, for example, is an advanced electron microscope that is undergoing upgrades that will allow it to generate hundreds of gigabytes of data per day. In the future, state-of-the-art detectors are expected to produce up to 65 gigabytes of data per second (approximately 200 terabytes per hour). Processing this vast amount of data quickly and efficiently requires closer integration between experiments and computational resources. To realize this vision, researchers are developing robust, open-source, modular software components to tightly couple experiments with HPC resources for near-real-time data processing using computationally expensive machine learning/AI models and traditional analysis techniques.

APPROACH To prepare for the Dynamic PicoProbe's increased data streams, Argonne researchers developed a data flow infrastructure that leverages Globus automation services to transfer experimental data to Polaris. Their work focused on two scientific use cases: hyperspectral and spatiotemporal imaging datasets, which involve off-site data transfers; machine learning/AI and traditional data analysis approaches; and automatic metadata extraction and cataloging of experimental results. The team's approach also provides a simple access portal for researchers to view experimental analyses to help guide the next set of experimental measurements and easily share their findings. They leveraged the Globus Search service and the Django Globus Portal Framework to make data Findable, Accessible, Interoperable, and Reusable (FAIR).

RESULTS In a study presented at the SC23 conference, the team detailed how their software infrastructure successfully managed the transfer and analysis of hyperspectral and spatiotemporal data from the Dynamic PicoProbe to the ALCF. In the hyperspectral imaging use case, the system generated detailed plots and metadata, providing insights into the atomic composition of samples. In the spatiotemporal imaging use case, a fine-tuned YOLOv8 neural network model was employed to track and analyze the movement of nanoparticles, demonstrating the system's ability to handle complex data streams and extract scientifically meaningful information.

IMPACT The team's work provides an end-to-end software infrastructure for integrating experimental instruments with HPC and AI resources for rapid data analysis. By automating data processing tasks, their method lays the groundwork for more efficient, experimental workflows that streamline the research process and accelerate discoveries.

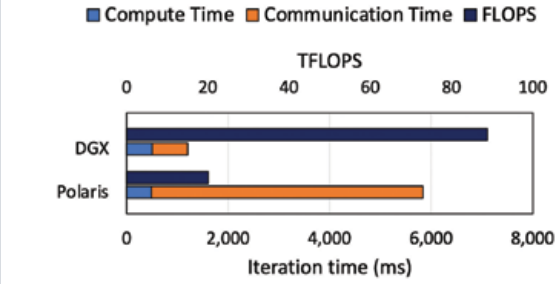
PUBLICATIONS

Bass, B., J. New, A. Berres, M. Adams, N. Clinton, M. Leung, K. Tuxen-Bettman, Brace, A., R. Vescovi, R. Chard, N. D. Saint, A. Ramanathan, N. J. Zaluzec, and I. Foster. "Linking the Dynamic PicoProbe Analytical Electron-Optical Beamline/Microscope to Supercomputers," *Proceedings of the SC23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis* (November 2023), ACM. https://doi.org/10.1145/3624062.3624614.

Computer Science | Data, Learning

Thorough Characterization and Analysis of Large Transformer Model Training At-Scale

PI Venkatram Vishwanath, Argonne National Laboratory
Murali Emani, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Polaris



The FLOPS (floating-point operations per second) metric can not fully reflect the underlying execution profile, especially in a bandwidth-limited system as opposed to DGX systems. *Image: Murali Emani, Argonne National Laboratory*

With the emergence of ChatGPT, DALL-E, and other generative AI tools, large transformer models are at the forefront of the AI revolution. As their capabilities continue to evolve, researchers are leveraging large transformer models for scientific breakthroughs, such as protein structure prediction (AlphaFold) and whole genome analyses (GenSLM). In a recent study, researchers from Argonne and Penn State University carried out an in-depth analysis and characterization of large transformer model training on supercomputers, leveraging parallelization strategies and network optimizations to demonstrate how HPC can significantly accelerate the training of such models.

CHALLENGE Training large transformer models for scientific research presents substantial challenges due to the immense computational and communication demands. As these models grow, the communication overhead between GPUs can present a bottleneck to efficient training. While HPC systems offer the computational power required to address these challenges, quantifying the communication overhead and thereby optimizing communication efficiency across GPUs is crucial to achieving scalable performance.

APPROACH For this study, the researchers used the state-of-the-art Megatron-DeepSpeed training framework to characterize large transformer model training on three GPU-powered supercomputers: ALCF's Polaris system and the NVIDIA DGX-based TG40 and TG80 machines. Their work provides a bottom-up breakdown of training throughput into compute and communication time, and quantitatively analyzes their respective influences on overall end-to-end training scaling. The team's evaluation included an in-depth exploration of data parallelism, scaling up to 512 GPUs with limited bandwidth. They examined three model sharding strategies among six model sizes and evaluated three combinations of model parallelism at high- and low-bandwidth. In addition, the team developed a novel technique to project the estimated speedup of the model training at scale by incorporating the communication overhead.

RESULTS The team's evaluation revealed that communication, on average, consumes 80 percent of the time during training on Polaris, and that the system's lower bandwidth increases communication time by up to 22 percent compared to the DGX machines. Furthermore, their analysis provided precise end-to-end training time estimations with a mean squared error of 2 percent across diverse model sizes and system architectures. By delineating the distinct roles of various communication calls, the team demonstrated how the synergy between data parallelism and model parallelism enables effective training scaling across various network bandwidths.

IMPACT The team's study provides valuable insights into the optimization of large transformer model training on HPC systems. Their work demonstrates how improving communication efficiency and leveraging advanced parallelization strategies can significantly reduce training time, thereby accelerating the development of AI models for science. In addition, they show that the interplay between compute and communication during large-scale training can help inform the efficient, scalable, and performant design of next-generation supercomputers.

PUBLICATIONS

Cheng, S., J.-L. Lin, M. Emani, S. Raskar, S. Foreman, Z. Xie, V. Vishwanath, and M. T. Kandemir. "Thorough Characterization and Analysis of Large Transformer Model Training At-Scale," *Proceedings of the ACM on Measurement and Analysis of Computing Systems* (February 2024), ACM. https://doi.org/10.1145/3639034

Earth Science | Simulation, Learning

Scaling Transformers for Skillful and Reliable Medium-Range Weather Forecasting

PI Rao Kotamarthi, Argonne National Laboratory
AWARD INCITE
SYSTEM Polaris

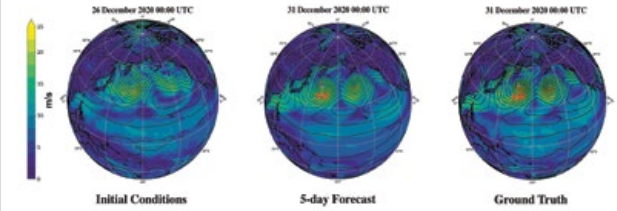


Illustration of an example 5-day forecast of near-surface wind speed (color-fill) and mean sea level pressure (contours). On December 31, 2020, an extratropical cyclone impacted Alaska, setting a new North Pacific low-pressure record. *Image: Rao Kotamarthi, Argonne National Laboratory*

High-resolution climate models that provide accurate weather forecasting are critical tools for mitigating the impacts of climate change and extreme weather events, while also enhancing our understanding of the atmosphere. To address the limitations of existing models, a research team from Argonne and the University of California, Los Angeles leveraged ALCF resources to develop Stormer, a scalable deep learning model that achieves state-of-the-art accuracy and performance for medium-range weather forecasting.

CHALLENGE Atmospheric scientists have traditionally relied on numerical weather prediction (NWP) models for short-term forecasts and coupled Earth system models for climate projections. Despite their widespread use, these models are limited by many factors, including parameterization errors in small-scale physical processes (e.g., cloud physics and radiation) and high computational costs, especially at fine spatial and temporal resolutions. Furthermore, due to the high computational costs, NWP and climate models can only incorporate limited amounts of observational data in simulations and large ensemble forecasts to account for uncertainties. To overcome these challenges, researchers are exploring the use of data-driven approaches based on deep learning for weather forecasting.

APPROACH Starting with a standard vision transformer architecture, the Argonne-UCLA team developed Stormer, a scalable transformer model for weather forecasting. The researchers trained Stormer on the ERA5 dataset from the WeatherBench 2 benchmark, leveraging the ALCF's Polaris supercomputer to complete the training process efficiently. Through extensive ablation studies, they identified and implemented three components that are key to model performance: a weather-specific embedding layer, a randomized dynamics forecasting objective, and a pressure-weighted loss function.

RESULTS The team compared Stormer's performance with two leading deep learning models for weather forecasting: Pangu-Weather and GraphCast. They found that Stormer achieves competitive accuracy for 1–5 day forecasts and surpasses state-of-the-art performance beyond 5 days, with improving performance as model capacity and data size increase. Stormer demonstrated up to 20 percent better accuracy at a 14-day lead time compared to the other models. Notably, Stormer achieves this superior performance with significantly less training data and compute resources, completing training in under 24 hours, compared to the much longer times required by Pangu-Weather and GraphCast. The team's work was recognized with Best Paper at the ICLR 2024 Workshop: Tackling Climate Change with Machine Learning.

IMPACT The Stormer model's ability to maintain accuracy over extended periods marks a substantial improvement in medium-range weather prediction capabilities, providing a powerful new tool for generating more reliable predictions to address issues related to climate change and extreme weather events. In addition, the efficiency of Stormer's training process can inform the development of future weather forecasting models to make high-precision predictions more accessible and less resource-intensive.

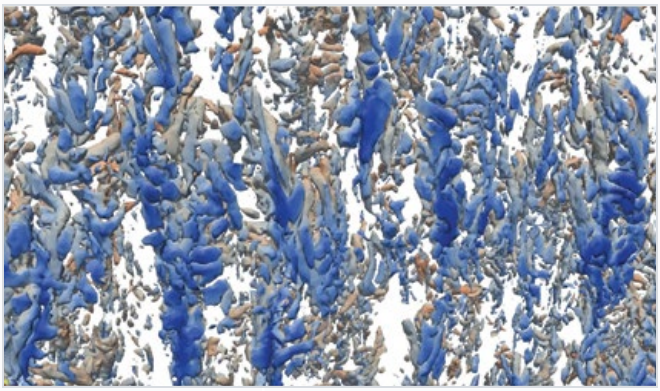
PUBLICATIONS

Nguyen, T., R. Shah, H. Bansal, T. Arcomano, S. Madireddy, R. Maulik, V. Kotamarthi, I. Foster, and A. Grover. "Scaling Transformer Neural Networks for Skillful and Reliable Medium-Range Weather Forecasting," *ICLR 2024 Workshop on Tackling Climate Change with Machine Learning* (May 2024).

Energy Technologies | Simulation

DNS Simulations of Coolant Flow in the High Flux Isotope Reactor

PI Emilian Popov, Oak Ridge National Laboratory
Igor Bolotnov, North Carolina State University
AWARD ALCC
SYSTEM Theta



Turbulent structures in wall vicinity showing the vortex flow colored with velocity magnitude. *Image: Emilian Popov, Oak Ridge National Laboratory*

The High Flux Isotope Reactor (HFIR) at Oak Ridge National Laboratory is a source of thermal and cold neutrons for research projects across numerous scientific disciplines. As part of a DOE initiative to reduce the enrichment of research and test reactors, a project is underway to investigate the conversion of HFIR from a high enriched uranium core to a low enriched uranium core. To support this effort, researchers from Oak Ridge and North Carolina State University are using DOE supercomputers to perform direct numerical simulations (DNS) within HFIR's complex coolant channel geometries, providing data to improve the accuracy of computational fluid dynamics models used for reactor design.

CHALLENGE The HFIR uses involute, or curved, fuel plates to supply intense neutron fluxes, forming cooling channels with high coolant velocities. Due to the complex channel geometry and the difficulty of performing experiments to collect live data, high-fidelity numerical simulations are required to verify and calibrate Reynolds-Averaged Navier-Stokes models. Performing DNS for these geometries requires HPC resources.

APPROACH To overcome these challenges, the team used NERSC and ALCF supercomputing resources to create a detailed DNS simulation to supply high-resolution data and build a flow library for involute coolant channels. The team used an involute coolant channel consistent with the HFIR as the simulation's model to perform the primary analysis. The team followed a DNS approach and used a computational algorithm to numerically integrate Navier–Stokes equations implemented in the parallel hierarchic stabilized transient analysis (PHASTA) code. This application allows for second-order accuracy in both space and time, and is an effective tool for bridging a broad range of length scales in turbulent flows. The team compared their simulations against each other and to other DNS studies performed on parallel flows.

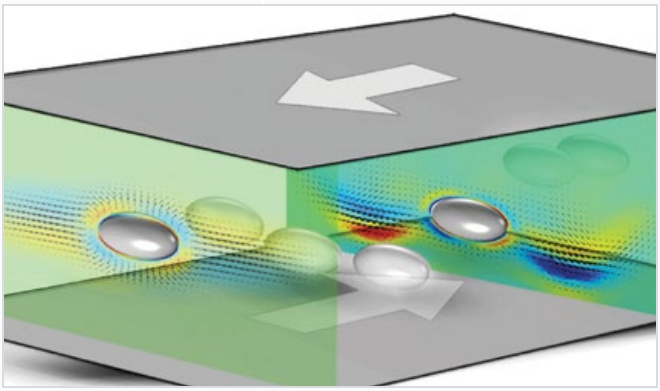
RESULTS The team produced a database representing the turbulent, statistically steady flow in the involute coolant channel. Their results indicate that the bulk involute channel flow does not differ significantly from a flat parallel channel flow and that the curvature of the walls does not significantly alter the mean flow parameters. However, the regions of the involute channel near the side walls exhibited relatively low magnitude twin recirculation structures driven toward the sidewalls from the centerline of the channel, which warrants further study.

IMPACT The project is leveraging DOE supercomputers to perform direct numerical simulations of turbulent flows within HFIR's complex coolant channel geometries. The team's simulations are generating high-fidelity data to enhance the accuracy of computational fluid dynamics models, which will be used to improve the thermal-hydraulic design of HFIR and next-generation energy systems.

PUBLICATIONS
Popov, E. L., N. J. Mecham, and I. A. Bolotnov. "Direct Numerical Simulation of Involute Channel Turbulence," *Journal of Fluids Engineering* (August 2024), ASME. https://doi.org/10.1115/1.4064496

Bifurcation of Equilibrium Positions for Ellipsoidal Particles in Inertial Shear Flows Between Two Walls

PI Zhangli Peng, University of Illinois Chicago
AWARD Director's Discretionary
SYSTEM Theta



Schematic of elliptical particles between two parallel walls moving in opposite directions. The color map represents streamwise vorticity; arrows show the in-plane velocity field. *Image: J. Fluid Mech. (2024), vol. 984, A47*

Inertial migration, which refers to the movement of particles in a fluid flow due to inertia, is an important phenomenon in fluid mechanics. Understanding and controlling this process is key to developing novel separation techniques and devices. To shed light on the mechanisms of this phenomenon, researchers from the University of Illinois Chicago (UIC) carried out simulations on ALCF supercomputers to investigate the inertial dynamics of finite-size particles with various shapes.

CHALLENGE The team's research focused on exploring the behavior of ellipsoidal particles in an inertial shear flow between two walls. Several factors, including complex particle geometries, nonlinear fluid dynamics, and experimental limitations, present challenges to studying such systems. Using supercomputers, researchers have the ability to perform computationally intensive simulations of complex geometries and flow conditions at high Reynolds numbers to explore particle behavior and fluid dynamics in great detail.

APPROACH With access to the ALCF's Theta supercomputer, the UIC team employed smoothed particle hydrodynamics to carry out a systematic numerical investigation of spherical, prolate, and oblate particles in an inertial shear flow between two parallel walls. As part of this work, they explored higher Reynolds numbers than were considered in previous studies. The researchers also performed simulations using the finite element method to validate their results and provide additional insights into the underlying physics.

RESULTS As detailed in a paper published in the *Journal of Fluid Mechanics*, the researchers found multiple parameters, including particle size, shape, initial configuration, and Reynolds number, influence the pitchfork bifurcation of equilibrium positions for rigid shaped particles in a shear flow between two walls. Their observations included the existence of lower and upper thresholds in Reynolds number values that determine the bifurcation behavior of ellipsoidal particles. For the first time, they discovered that ellipsoidal particles can exhibit a reversal in behavior at an increased Reynolds number, while spherical particles become unstable. The researchers found the underlying mechanism of this reversal was the altered streamwise vorticity and symmetry breaking of pressure.

IMPACT The team's research provides new insights into the inertial particle dynamics in complex microfluidic systems. Their findings pave the way for the potential development of novel particle separation methods for various applications, such as industrial processes and biomedical devices.

PUBLICATIONS
Lauricella G., M. M. Naderi, J. Zhou, I. Papautsky, and Z. Peng. "Bifurcation of Equilibrium Positions for Ellipsoidal Particles in Inertial Shear Flows Between Two Walls," *Journal of Fluid Mechanics* (April 2024), Cambridge University Press. https://doi.org/10.1017/jfm.2024.152

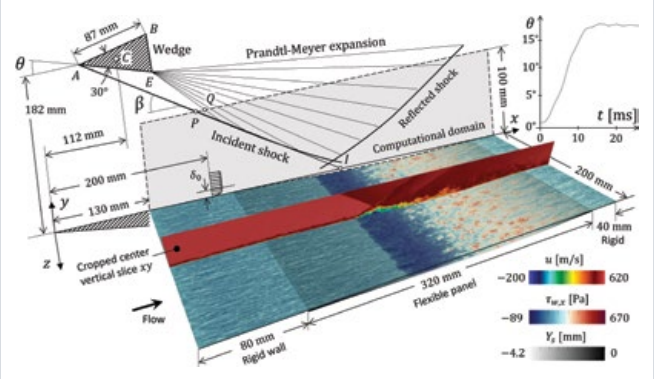
Engineering | Simulation

High-Speed Turbulence with Shocks Over Non-Adiabatic and Flexible Walls

PI Johan Larsson, University of Maryland
Ivan Bermejo-Moreno, University of Southern California

AWARD INCITE

SYSTEM Theta



Simulation results for a wedge deflection of 17.5 degrees at time t=15 ms, showing contour maps of the flexible panel vertical displacement, the wall shear stress, and the fluid flow velocity on a vertical slice at the center of the spanwise domain. *Image: Ivan Bermejo-Moreno, University of Southern California*

To design safe and efficient hypersonic aircraft, engineers must understand how shock waves and turbulence affect the aircraft's performance and structural integrity. Recently, a team from the University of Southern California and the University of Maryland used ALCF supercomputing resources to develop predictive 3D simulations of shock wave and turbulent interactions over flexible walls.

CHALLENGE High-speed airflows create a thin boundary layer along solid surfaces. When a shock hits the boundary layer forcefully enough, it may create high-amplitude, low-frequency oscillations that can damage the aircraft. To address this, we need to understand the mechanics of shock wave and turbulent boundary-layer interactions (STBLI) and the fluid-structure interactions (FSI). Researchers have not extensively studied the fluid-structural coupling of STBLIs. While teams have led efforts to investigate the relationships of STBLIs with flexible panels, it is extremely challenging to characterize this class of interactions experimentally. Numerical simulations are crucial to provide these fundamental insights. Several simulation approaches have been tested, but many suffer from high associated computational costs. Approaches that have lower computational costs suffer from lower accuracy.

APPROACH The team used the ALCF's Theta supercomputer to perform full 3D simulations of the coupling between FSI and STBLI over flexible walls to replicate and complement wind-tunnel experiments. They studied the interactions' characteristic low-frequency motions on flexible panels using wall-modeled large-eddy simulations (WMLES). This method models rather than resolves the inner boundary layer, reducing the computational cost of the simulation and maintaining the physical fidelity of flow features like separation and reattachment. This approach allows for sufficiently long integration times needed to capture the low-frequency motions of interest. The team coupled the WMLES with a finite-element solid mechanics solver to incorporate structural damping. This is the first time researchers have combined these approaches to study such interactions on flexible panels.

RESULTS To validate the high-fidelity simulation methodology, the team used WMLES to replicate experiments at different strengths of the incident shock on the turbulent boundary layer. Based on these results, the team assessed the importance of the 3D effects in those interactions by conducting reduced-span simulations with imposed periodicity in the spanwise direction. The simulations replicated the coupled interactions observed experimentally with better accuracy than prior numerical studies, while also providing additional insights into the wind-tunnel experiments.

IMPACT The ultimate goal of the team's research is to develop improved modeling techniques for the prediction of fluid-thermal-structural interactions through coupled specialized domain-specific solvers. These techniques will reduce the uncertainty factored into designing hypersonic vehicles and propulsion systems, leading to safer and more efficient aircraft designs.

PUBLICATIONS

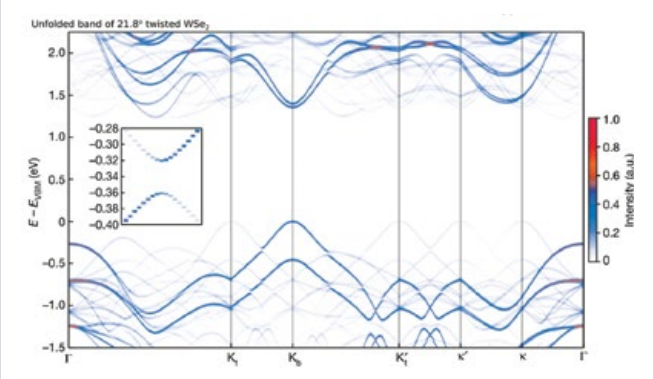
Hoy, J. and I. Bermejo-Moreno. "Fluid-Structural Coupling of an Impinging Shock-Turbulent Boundary Layer Interaction at Mach 3 Over a Flexible Panel," *Flow* (October 2023), Cambridge University Press. https://doi.org/10.1017/flo.2022.28

Hoy, J. and I. Bermejo-Moreno. "Numerical Study of STBLI on Flexible Panels with Wall-Modeled LES," *AIAA Scitech Forum* (2022), AIAA. https://doi.org/10.2514/6.2021-0250

Materials Science | Simulation

Computational Design of Novel Semiconductors for Power and Energy Applications

PI Feliciano Giustino, University of Texas at Austin
AWARD ALCC
SYSTEM Theta



Top: Schematic of the experimental set-up and Gaussian peak fittings for magnified constant-current ∂I/∂V spectroscopies at the 21.8° twisted region. Bottom: DFT-calculated unfolded band structure of 21.8° commensurate moiré along the k-space path. *Image: Viet-Anh Ha, the University of Texas at Austin*

This project aimed to develop and leverage cutting-edge first-principles computational methods to investigate the temperature-dependent optical and transport properties of compound semiconductors, perovskites, and plasmonic ceramics for uses in power devices, solar cells, and light-emitting diodes.

CHALLENGE The emergence of moiré superlattices (MSL) designed in van der Waals bilayers has created opportunities to engineer two-dimensional electronic materials with unique properties. Most superlattices investigated to date are van der Waals bilayers with small twist angles; large-angle moiré structures have been mostly unexplored. The absence of electronic coupling near the Fermi level has limited the effect of quasiperiodicity on other physical properties that are governed primarily by states near E_F in a metallic system.

APPROACH The researchers overcame the absence of electronic coupling near the Fermi level by using tungsten diselenide (WSe_2) twisted bilayers whose flatter dispersion accentuates the effects of quasiperiodicity at accessible doping levels to show they offer a platform to explore moiré physics beyond the MSL. They used DOE supercomputing resources, including ALCF systems, to perform density functional theory calculations implemented in the Quantum ESPRESSO suite.

RESULTS The researchers' results—showing that large-angle twisted bilayers offer distinctly different platforms with a wide range of moiré physics—were published in *Nature*. By using twisted tungsten diselenide bilayers, the team created the incommensurate dodecagon quasicrystals and the commensurate moiré crystals. Valley-resolved scanning tunnelling spectroscopy shows disparate behaviors between moiré crystals and quasicrystals. The formation of mini-gaps near the valence band maximum exemplifies the rich electronic structures identified in the K valley.

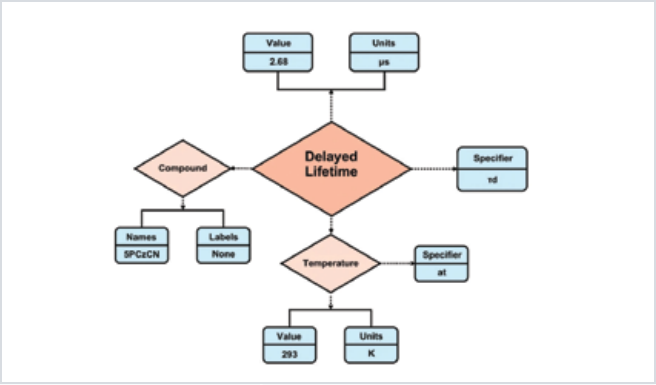
IMPACT As the demand for clean and sustainable energy is increasing, it is imperative to build novel power-conversion and transmission systems, which require the discovery and design of materials with well-balanced electric, transport, thermal, and optical properties that offer reliable stability. The simulations performed for this work have aimed to expedite this process by facilitating and guiding the experimental synthesis and optimization of energy materials.

PUBLICATIONS
Li, Y., F. Zhang, V.-A. Ha, Y.-C. Lin, C. Dong, Q. Gao, Z. Liu, X. Liu, S. H. Ryu, H. Kim, C. Jozwiak, A. Bostwick, K. Watanabe, T. Taniguchi, B. Kousa, X. Li, E. Rotenberg, E. Khalaf, J. A. Robinson, F. Giustino, and C.-K. Shih. "Tuning Commensurability in Twisted Van der Waals Bilayers," *Nature* (January 2024), Springer Nature. https://doi.org/10.1038/s41586-023-06904-w

Materials Science | Simulation, Data, Learning

Data-Driven Materials Discovery for Functional Applications

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



Example of an extracted ChemDataExtractor v2.1 model for TADF-based research. The ontology shows the values, units, chemical compound, specifier, and the temperature at measurement that are associated with a Delayed Lifetime record. *Image: Jacqueline Cole and Dingyun Huang, University of Cambridge*

In the rapidly expanding field of materials design and discovery, the ability to efficiently extract and analyze vast amounts of data from published research is crucial. With help from ALCF computing resources, researchers from the University of Cambridge are developing and demonstrating novel AI-powered tools that can automate the extraction of complex chemical data and the generation of literature reviews.

CHALLENGE Conventional trial-and-error materials synthesis processes can take years of effort, presenting a significant bottleneck to materials discovery. In addition, the growing volume of scientific literature has made it more difficult to sift through thousands of papers to find relevant data. Over the last decade, data-driven design-to-device pipelines, along with advances in HPC and AI capabilities, have significantly advanced efforts to design and discover new functional materials.

APPROACH The Molecular Engineering group at the University of Cambridge is leveraging ALCF resources to develop and employ advanced natural language processing (NLP) tools to automatically process scientific literature for targeted studies. The team's ChemDataExtractor is a chemistry-aware NLP toolkit designed to automatically extract structured chemical data from unstructured text found in scientific publications. They also developed ChemDataWriter, a Python-based toolkit that can auto-generate summaries of scientific papers in a completely unsupervised fashion.

RESULTS The team's research using ALCF resources has resulted in several recent publications. In two separate studies published in *Scientific Data*, they used ChemDataExtractor to automatically generate datasets for photocatalysis and thermally activated delayed fluorescent (TADF) molecules with good precision. Aimed at water-splitting applications, the photocatalysis dataset was comprised of 15,755 records extracted from 47,357 research papers. The TADF dataset, which has applications in organic light-emitting diodes, included 25,482 records from 2,733 scientific articles. In a *Digital Discovery* paper, the team introduced ChemDataWriter—the first open-source toolkit in chemistry capable of automatically composing literature reviews from numerous scientific publications. Additionally, the team published a study in the *Journal of Chemical Information and Modeling* that found pretraining language models on domain-specific materials data for photocatalytic water splitting improves model performance, particularly when extracting detailed chemical attributes.

IMPACT By automating the extraction and analysis of chemical data from scientific literature, the team's AI toolkits and databases can help researchers drastically reduce the time and effort involved in curating databases that are needed to accelerate materials discovery.

PUBLICATIONS

Isazawa, T., and J. M. Cole. "Automated Construction of a Photocatalysis Dataset for Water-Splitting Applications," *Scientific Data* (September 2023), Springer Nature. https://doi.org/10.1038/s41597-023-02511-6

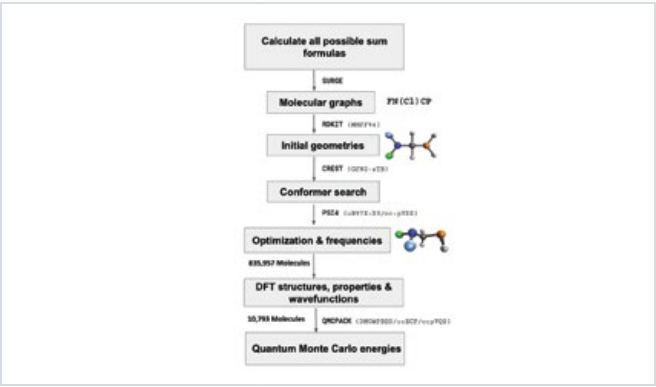
Huang, S., and J. M. Cole. "ChemDataWriter: A Transformer-Based Toolkit for Auto-Generating Books that Summarise Research," *Digital Discovery* (October 2023), Royal Society of Chemistry. https://doi.org/10.1039/D3DD00159H

Huang, D., and J. M. Cole. "A Database of Thermally Activated Delayed Fluorescent Molecules Auto-Generated from Scientific Literature with ChemDataExtractor," *Scientific Data* (January 2024), Springer Nature. https://doi.org/10.1038/s41597-023-02897-3

Isazawa, T., and J. M. Cole. "How Beneficial Is Pretraining on a Narrow Domain-Specific Corpus for Information Extraction about Photocatalytic Water Splitting?," *Journal of Chemical Information and Modeling* (March 2024), ACS Publications. https://doi.org/10.1021/acs.jcim.4c00063

Exascale Simulations of Quantum Materials

PI Paul Kent, Oak Ridge National Laboratory
Anouar Benali, Argonne National Laboratory
AWARD INCITE
SYSTEM Polaris



Workflow used to generate the VQM24 dataset. *Image: Anouar Benali, Argonne National Laboratory*

Navigating the vast chemical compound space (CCS) to identify molecules with desirable properties is critical to the design and discovery of new materials for batteries, catalysts, and other important applications. Machine learning (ML) and AI techniques have emerged as powerful tools for accelerating explorations of CCS. Using ALCF supercomputers, a multi-institutional research team generated a comprehensive quantum mechanical (QM) dataset to enhance the training and testing of ML models for materials discovery.

CHALLENGE Developing effective ML models to sift through CCS requires high-quality QM datasets of molecular properties. Numerous datasets focusing on distinct, chemically relevant subspaces have paved the way for the systematic and quantitative exploration of CCS. However, due to the massive number of possible compounds, existing datasets are often incomplete, introducing biases that can affect the accuracy of the ML models trained and assessed on them. While density functional theory (DFT) has been key to the development of highly accurate and efficient ML models over the past decade, there is still a lack of datasets that exhaustively cover specific regions of chemical space at higher QM levels.

APPROACH To address these challenges, the INCITE team and their collaborators generated VECTOR-QM24 (VQM24), a diverse and comprehensive dataset that aims to cover all possible neutral closed-shell small organic and inorganic molecules and their conformers. The researchers used DFT to optimize 258,242 unique constitutional isomers and 577,705 conformers of varying stoichiometries. To further enhance the accuracy of the VQM24 dataset, the team selected a subset of 10,793 molecules with up to four heavy atoms and calculated their energies using the highly accurate diffusion quantum Monte Carlo method. These computationally intensive calculations were carried out on the ALCF's Polaris supercomputer using the open-source QMCPACK code.

RESULTS The team's research provides one of the largest and most precise quantum Monte Carlo datasets available for small molecules. Covering a total of 835,947 converged molecules, the VQM24 dataset includes internal, atomization, electron-electron repulsion, exchange correlation, dispersion, vibrational frequency, Gibbs free, enthalpy, ZPV, and molecular orbital energies. It also provides thermal properties such as entropy and heat capacities, as well as electronic properties including multipole moments (dipole, quadrupole, octupole, hexadecapole), electrostatic potentials at nuclei (alchemical potential), Mulliken charges, and molecular wavefunctions. Together, these properties make VQM24 a highly accurate and unbiased dataset of molecules that can be used to develop enhanced ML models.

IMPACT The VQM24 dataset is a valuable resource for testing and training ML models aimed at accelerating the discovery of new functional materials. By providing a more complete and accurate representation of the chemical compound space, VQM24 can help researchers improve the predictive power of scalable and generative ML models of real quantum systems.

PUBLICATIONS
Khan, D., A. Benali, S. Y. H. Kim, G. F. von Rudorff, and O. A. von Lilienfeld. "Towards Comprehensive Coverage of Chemical Space: Quantum Mechanical Properties of 836k Constitutional and Conformational Closed Shell Neutral Isomers Consisting of HCNOFSiPSClBr," *arXiv* (May 2024). https://doi.org/10.48550/arXiv.2405.05961

Materials Science | Simulation

First-Principles Electron Dynamics in Complex Systems

PI André Schleife, University of Illinois Urbana-Champaign
AWARD INCITE
SYSTEM Theta

The team's study was featured on the cover of *Nano Letters*. The image depicts a proton irradiating a 2D layer of hot graphene. *Image: Nano Lett. 2024, 24, 17, 5174-5181.*

Understanding the properties of secondary electrons emitted from two-dimensional materials like graphene could form the basis for advances in materials science and chemistry. Recently, a team led by researchers at the University of Illinois Urbana-Champaign used ALCF supercomputers to develop a new approach to understanding fast electronic processes by irradiating 2D materials with ions.

CHALLENGE The intensity and kinetic energy distribution of secondary electrons emitted from ion-irradiated targets is not fully understood. Traditionally, researchers use lasers to irradiate 2D materials and deduce their properties. Using ions instead of lasers could enable more precise insights into how these materials evolve over time. Using ions invokes a complex multi-length and multi-time scale relaxation process that is dominated by interactions within the target. The UIUC team had previously developed codes that could simulate electron excitations on this scale. However, they found that university-level computing centers could not execute their simulations due to the project's high computational costs.

APPROACH The team used the ALCF's Theta supercomputer to develop a computational first-principles description of the secondary electron emission dynamics for proton-irradiated graphene as a prototypical 2D material. They based their calculations on real-time time-dependent density functional theory. This approach has led to successful simulations of femtosecond dynamics under external radiation. The team performed calculations on graphene at absolute zero and at a higher temperature. They simulated irradiating graphene with hydrogen ions, then calculated the release of secondary electrons and the resulting energy spectrum over time. Collaborators at the University of Duisburg–Essen performed irradiation experiments with argon and xenon ions to validate the simulations.

RESULTS The team's simulations provided key insights into the mechanisms behind secondary electron emission. Focusing on the effects of non-vanishing electron and lattice temperatures of the pre-excited graphene, they made key discoveries on how ions can probe thermalization processes. They found that differences in peak intensity and secondary electron yield near the entrance and exit sides suggested multiple contributing emission mechanisms. The team also observed that high temperature graphene released more secondary electrons. By examining the charge distribution, they determined that the material's lattice was responsible for this outcome, rather than the material's electrons.

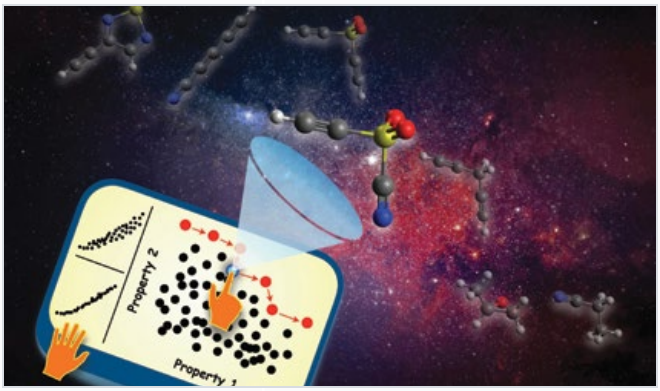
IMPACT This research provides new insights into the electronic properties of 2D materials by irradiating them with ions, allowing for highly precise and localized measurements. Looking forward, ion irradiation has the potential to deliberately introduce defects into materials and manipulate them for targeted applications. These techniques could also lead to advancements in solar-fuel conversion and photo-ionic systems.

PUBLICATIONS

Yao, Y., A. Kononov, A. Metzlaff, A. Wucher, L. Kalkhoff, L. Breuer, M. Schleberger, and A. Schleife. "Nonequilibrium Dynamics of Electron Emission from Cold and Hot Graphene under Proton Irradiation," *Nano Letters* (April 2024), American Chemical Society. https://doi.org/10.1021/acs.nanolett.4c00356

Freedom of Design in Chemical Compound Space

PI Alexandre Tkatchenko, University of Luxembourg
AWARD Director's Discretionary
SYSTEM Theta, Mira



Graphical depiction of the rational molecular design process, which involves a "needle-in-a-haystack" search for molecules with a desired set of properties. *Image: L. Medrano Sandonas et al., Chem. Sci. 14, 10702, (2023), Royal Society of Chemistry*

The chemical compound space (CCS) represents an unfathomably vast space populated by all possible atomic compositions and geometries. Academic and industrial researchers alike have investigated the CCS to understand the fundamental relationships between molecular structures and their physical and chemical properties. Recently, a multi-institutional team used ALCF supercomputing resources to introduce a new "freedom of design" principle aimed at accelerating the discovery of previously unknown molecules with targeted properties.

CHALLENGE Data-driven approaches to studying the CCS have enhanced our qualitative and quantitative understanding of molecular structural and property relationships. While previous studies have led to significant discoveries, a comprehensive understanding of these complex relationships—even in the more manageable sector of CCS spanned by small molecules—is still lacking. Gaining insights into these relationships could improve our ability to discover key molecules that underpin advances in chemistry and material science.

APPROACH With this project, researchers leveraged ALCF supercomputing resources to perform a comprehensive analysis of the CCS corresponding to small organic molecules. The team used Mira and Theta to compute quantum mechanical (QM) properties of millions of molecular systems. They quantitatively analyzed the pairwise correlations between select properties, progressively investigating complex manifolds of the property space and their underlying dependencies on structure and composition. The researchers recognized that the molecular design process often involves the simultaneous optimization of multiple physicochemical properties. To account for this, they used Pareto multi-property optimization to search for molecules with simultaneously large polarizability and electronic gaps, further exploring design approaches in the CCS.

RESULTS The team's quantitative analysis demonstrated that there are very few limitations which prevent markedly distinct molecules from sharing multiple QM properties. They also observed thousands of structurally and/or compositionally distinct molecules that share multiple properties. Together, these findings demonstrated "freedom of design" in CCS—an intrinsic degree of flexibility enabling rational molecule design with very few limitations preventing many molecules from sharing an array of QM properties. The researchers' Pareto fronts also identified paths through CCS consisting of several unexpected molecules connected by structural and/or compositional changes, further reflecting this principle.

IMPACT The "freedom of design" principle has important implications for enabling the rational design of molecules with a desired set of properties. In addition, the team released a database that lays the groundwork for training AI models to accelerate discoveries in chemistry and materials science. Their work provides a foundation for a wide array of innovations, including the inverse design of molecular systems for various applications, such as semiconductors, pharmaceuticals, and batteries.

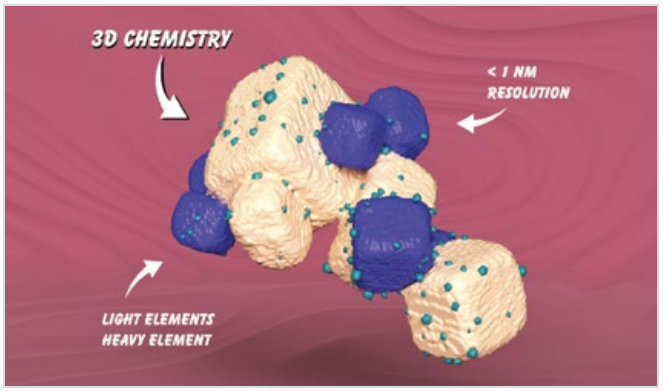
PUBLICATIONS
Sandonas, L. M., J. Hoja, B. G. Ernst, A. Vázquez-Mayagoitia, R. A. DiStasio, Jr., and A. Tkatchenko. "Freedom of Design' in Chemical Compound Space: Towards Rational in Silico Design of Molecules with Targeted Quantum-Mechanical Properties," *Chemical Science* (August 2023), Royal Society of Chemistry. https://doi.org/10.1039/D3SC03598K

Fallani, A., L. Medrano Sandonas, and A. Tkatchenko. "Inverse Mapping of Quantum Properties to Structures for Chemical Space," *Nature Communications* (July 2024), Springer Nature. https://doi.org/10.1038/s41467-024-50401-1

Materials Science | Simulation, Data

Imaging 3D Chemistry at 1 nm Resolution with Fused Multi-Modal Electron Tomography

PI Robert Hovden, University of Michigan
AWARD Director's Discretionary
SYSTEM Theta



A 3D rendering of chemical tomogram using multi-modal electron tomography. *Image: Jason Manassa, University of Michigan*

Electron and x-ray tomography allow researchers to perform 3D characterization of materials at the nano- and mesoscale, generating data that is critical to the development of nanomaterials for a wide range of applications, including solar cells and semiconductor devices. With support from ALCF computing resources, a University of Michigan-led research team developed an entirely new approach to seeing chemistry in 3D by combining the measured elastic and inelastic signals into a single volumetric reconstruction. Coined multi-modal electron tomography, the work demonstrated unprecedented resolution below 1 nanometer and dose reduction for chemical imaging.

CHALLENGE Measuring the 3D distribution of chemistry in nanoscale matter is a longstanding challenge. The inelastic scattering events required for 3D chemical imaging are too rare, requiring high-beam exposure that destroys the specimen before an experiment is completed. Even larger electron doses are required to achieve high resolution. Thus, chemical mapping in 3D has been unachievable except at lower resolution with the most radiation-hard materials.

APPROACH With access to DOE supercomputing resources, the researchers conducted comprehensive simulations for real-time electron tomography and developing reconstruction methods for through-focal tomography. As shown in a recent paper, the team achieved high-resolution 3D chemical imaging in an Au-Fe_3O_4 metamaterial within an organic ligand matrix, Co_3O_4-Mn_3O_4 core-shell nanocrystals, and ZnS-$Cu_{0.64}S_{0.36}$ nanomaterial using fused multi-modal electron tomography. Multi-modal data fusion enables high-resolution chemical tomography with electron doses reduced by as much as 99 percent by linking information encoded within both elastic and inelastic signals.

RESULTS Achieving high-precision 3D chemical imaging near or below 1-nanometer resolution using multimodal electron tomography, the team published a paper in *Nature Communications* that demonstrated that sub-nanometer 3D resolution of chemistry is measurable for a broad range of geometrically and compositionally complex materials.

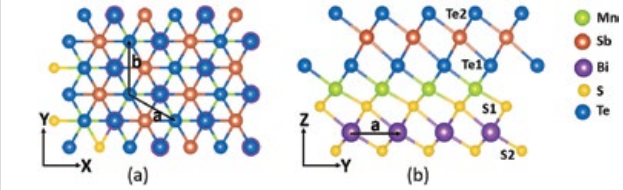
IMPACT The team's approach will help advance materials characterization research by enabling real-time analysis of 3D specimens while an experiment progresses. By integrating their framework with an open-source 3D visualization and tomography software package, the team's techniques will be accessible to a wide range of researchers and enable new material characterizations across academia and industry. Their research demonstrates that sub-nanometer 3D resolution of chemistry can be measured for a wide array of complex materials for which high-resolution mapping was previously unachievable.

PUBLICATIONS

Schwartz, J., Z. W. Di, Y. Jiang, J. Manassa, J. Pietryga, Y. Qian, M. G. Cho, J. L. Rowell, H. Zheng, R. D. Robinson, J. Gu, A. Kirilin, S. Rozeveld, P. Ercius, J. A. Fessler, T. Xu, M. Scott, and R. Hovden. "Imaging 3D Chemistry at 1 nm Resolution with Fused Multi-Modal Electron Tomography," *Nature Communications* (April 2024), Springer Nature. https://doi.org/10.1038/s41467-024-47558-0

Materials Informatics Study of Two-Dimensional Magnetic Materials

PI Trevor Rhone, Rensselaer Polytechnic Institute
AWARD Director's Discretionary
SYSTEM Theta



The crystal structure of monolayer $MnSbBiS_2Te_2$ from the top view (a) and side view (b). The black arrows show the lattice vectors. *Image: Romakanta Bhattarai, Rensselaer Polytechnic Institute*

2D magnetic Janus materials are a special class of magnetic materials that lack symmetry when flipped. Their unique crystal structure and strong spin-orbit interactions make them significant to fields like nanotechnology, quantum computing, and spintronics. Recently, a research team from Rensselaer Polytechnic Institute used ALCF supercomputers to investigate how strain can be used to tune the magnetic and topological properties of a 2D magnetic Janus material.

CHALLENGE In a previous study, the team used a combination of density functional theory (DFT) calculations and machine learning to predict the properties of several monolayer van der Waal materials similar to $MnBi_2Te_4$ (MBT). MBT is the first identified antiferromagnetic topological insulator, and has been a major focus in 2D materials research. This prior study highlighted several Janus and non-Janus MBT-type materials that could exhibit topological order. Before that, only a handful of these materials had been studied systematically. At the outset of this current project, no one had reported on the behavior of MBT-type materials that exhibit itinerant ferromagnetism in the monolayer. The team realized that a systematic investigation of novel 2D Janus phases in the MBT family was necessary to have a deep understanding MBT-type materials' behaviors.

APPROACH Using the ALCF's Theta supercomputer, the team performed first-principles DFT calculations to study the effects of strain on $MnSbBiS_2Te_2$. They applied a biaxial strain to the $MnSbBiS_2Te_2$ monolayer to investigate the changes in its electronic, magnetic, and topological properties within the range of -8 percent to +8 percent. The team also examined the chemical stability of this phase by calculating the formation energy, the phonon spectra, and ab initio molecular dynamics simulations.

RESULTS When applying the biaxial strain, the researchers found that a band gap appears with a compressive strain between -4 percent and -7 percent. They observed a topological phase transition when the strain reached -5 percent, where the material becomes a Chern insulator exhibiting a quantum anomalous hall effect. They also found that applying strain allowed them to control the magnetic easy axis of $MnSbBiS_2Te_2$. The biaxial strain and spin-orbit coupling are responsible for the topological phase transition in $MnSbBiS_2Te_2$. The team calculated the Curie temperature of $MnSbBiS_2Te_2$ using the Heisenberg model and found it to be 24 kelvins, which is comparable to that of other MBT-type materials.

IMPACT Identifying the key features of this special class of 2D magnetic topological materials has applications in an array of fields, including topological quantum computing, magnetic sensors, photoelectronic nanodevices, and spintronics. Further investigations into the effects of strain on $MnSbBiS_2Te_2$ will enable researchers to explore the electronic, magnetic, and topological properties of this class of materials in greater detail.

PUBLICATIONS

Bhattarai, R., P. Minch, Y. Liang, S. Zhang, and T. D. Rhone. "Strain-Induced Topological Phase Transition in Ferromagnetic Janus Monolayer $MnSbBiS_2Te_2$," *Physical Chemistry Chemical Physics* (March 2024), Royal Society of Chemistry. https://doi.org/10.1039/d3cp05578g

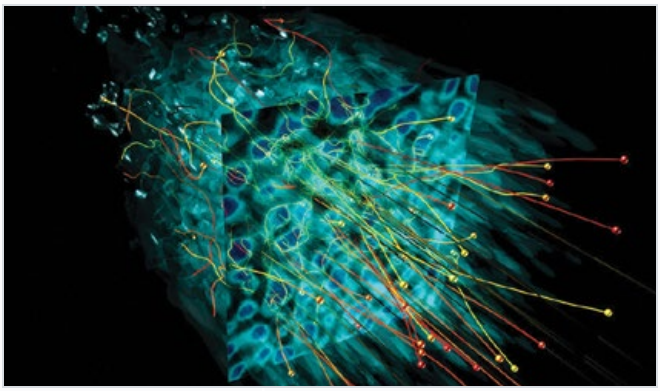
Bhattarai, R., P. Minch, and T. D. Rhone. "Investigating Magnetic van der Waals Materials Using Data-Driven Approaches," *Journal of Materials Chemistry C* (March 2023), Royal Society of Chemistry. https://doi.org/10.1039/D3TC00001J

Minch, P., R. Bhattarai, and T.D. Rhone. "Data-driven Study of Magnetic Anisotropy in Transition Metal Dichalcogenide Monolayers," *Solid State Communications* (June 2023), Elsevier. https://doi.org/10.1016/j.ssc.2023.115248

Physics | Simulation

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

PI Frederico Fiúza, Instituto Superior Técnico and SLAC National Accelerator Laboratory
AWARD ALCC
SYSTEM Theta



3D kinetic simulation of the injection of electrons in collisionless shocks. Electrons (yellow-red spheres colored by energy) are accelerated by scattering off the turbulence magnetic fields (blue-green) produced at the shock transition. *Image: Frederico Fiúza, Instituto Superior Técnico*

Collisionless shocks are powerful cosmic events that play a crucial role in high-energy astrophysical processes, driving particle acceleration and magnetic field amplification. While the mechanism behind these shock waves remains mysterious, recent developments in high-energy-density (HED) laser-plasma experiments, such as DOE's National Ignition Facility (NIF), are opening the door to study magnetized collisionless shocks in conditions relevant to astrophysical environments. To help advance such experiments, researchers are using ALCF supercomputers to carry out large-scale simulations of particle acceleration in laboratory-produced collisionless magnetized shocks.

CHALLENGE Diffusive shock acceleration (DSA) is considered the dominant particle acceleration mechanism at play in astrophysical shock waves. Understanding how electrons are preaccelerated to sufficiently high energies to engage in DSA has been a longstanding challenge in astrophysics. In high-Mach-number shocks, traditional mechanisms do not account for the efficient injection of electrons into the acceleration process. Leveraging the power of DOE supercomputers, researchers can simulate these complex plasma interactions at kinetic scales, providing insights that are otherwise inaccessible through observations or simpler models.

APPROACH With this ALCC project, researchers used the particle-in-cell code OSIRIS to model large-scale magnetized shocks in 2D and 3D on the ALCF's Theta supercomputer. For a study on electron injection in DSA, the team carried out simulations of NIF experimental conditions to better understand the shock structure and to identify the electron acceleration mechanism. They varied parameters, such as ion-to-electron mass ratios and spatial resolution, to ensure their simulations accurately captured the physical processes at play.

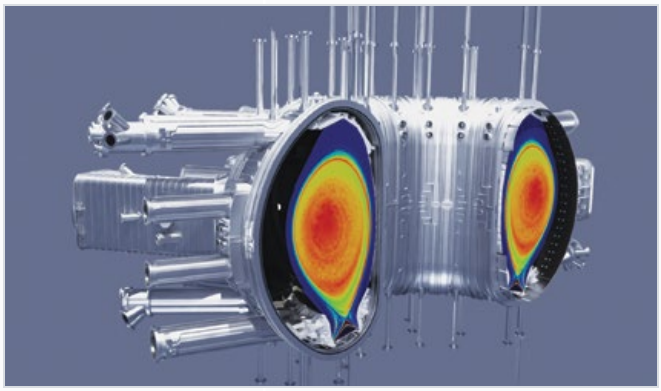
RESULTS The team's fully kinetic plasma simulations demonstrated how high-Mach-number shocks can effectively inject electrons by scattering in kinetic-scale magnetic turbulence produced near the shock transition via the ion Weibel, or current filamentation, instability. In a paper published in *The Astrophysical Journal Letters*, the researchers describe this process as a modified DSA mechanism that accounts for the flow velocity profile at the shock transition, in contrast to conventional DSA in which the shock front is considered an infinitely sharp transition. The electron energization rate, diffusion coefficient, and acceleration time obtained in their injection model are consistent with particle-in-cell simulations and with the results of recent laboratory experiments where nonthermal electron acceleration was observed. The team's injection model represents a natural extension of DSA and could account for electron injection in high-Mach-number astrophysical shocks, such as those associated with young supernova remnants and accretion shocks in galaxy clusters.

IMPACT The team's simulations provide new insights into how kinetic-scale magnetic turbulence can accelerate electrons in collisionless shocks. In addition, their research will help support the design and interpretation of future NIF experiments aimed at studying particle acceleration in magnetized shocks.

PUBLICATIONS
Grassi, A., H. G. Rinderknecht, G. F. Swadling, D. P. Higginson, H.-S. Park, A. Spitkovsky, and F. Fiúza. "Electron Injection via Modified Diffusive Shock Acceleration in High-Mach-Number Collisionless Shocks." *The Astrophysical Journal Letters* (November 2023), IOP Publishing. https://doi.org/10.3847/2041-8213/ad0cf9

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

PI CS Chang, Princeton Plasma Physics Laboratory
AWARD INCITE, Aurora ESP
SYSTEM Aurora, Polaris



Tungsten density profile in the German tokamak ASDEX-U. Like ITER, ASDEX-U uses a tungsten wall, which allows sputtering of tungsten particles into the core plasma. *Image: ALCF Visualization and Data Analytics Team; CS Chang, Princeton Plasma Physics Laboratory*

This project aims to significantly advance our understanding of fundamental edge plasma physics in fusion reactors, answering questions critical to the successful operation of ITER and to the design of fusion power plants.

CHALLENGE The goal of this project is to perform two-pronged, interrelated fundamental edge physics studies of critical importance to the successful operation of ITER and to the design of fusion power plants. The first prong is the mitigation of high stationary heat-flux densities that will damage material walls while maintaining the high edge plasma pedestal within a safe operational window. The second prong is avoiding explosive transient power flow to material walls caused by edge localized mode crash.

APPROACH Achieving their goals necessitates that the researchers employ DOE supercomputing resources. The team uses the electromagnetic edge gyrokinetic particle-in-cell code XGC, which enables the inclusion of two important but computationally expensive components: (a) the addition of tungsten impurity particles that are sputtered from ITER's material wall as a third species along with deuterium and tritium fuel particles, and (b) the capability for plasma detachment from the divertor plates. Tungsten impurity particles, beyond their deleterious effect of radiating away plasma energy in the core, are known to significantly impact the edge physics, and the detached plasma is known to significantly reduce the divertor heat load.

RESULTS A paper published in *Physics of Plasmas* demonstrated the application of a bundling technique to model the diverse charge states of tungsten impurity species in total-f gyrokinetic simulations. XGC was used to simulate a JET H-mode-like plasma across an entire plasma volume, spanning from the magnetic axis to the divertor. Tungsten impurities were found to affect the deuterium fluxes of particles and heat.

IMPACT As fusion power represents a paradigm-shifting breakthrough, success of the ITER project is a high-priority challenge, which will ultimately lead to more economical fusion power plants. To accomplish this goal, simulations based on first-principles must be deployed to solve issues with power exhaust, including mitigating stationary heat-flux densities and avoiding unacceptably high transient power flow to material walls.

PUBLICATIONS

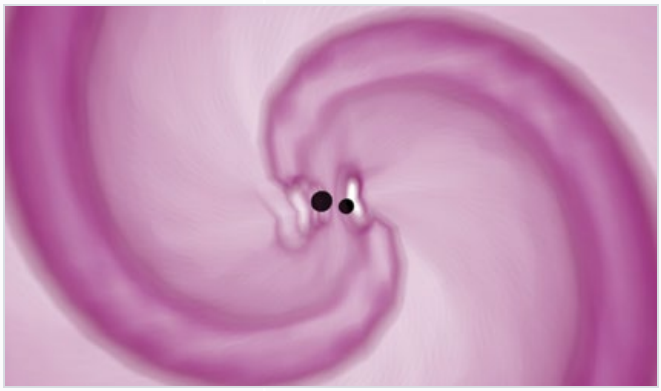
Chang, C. S., S. Ku, R. Hager, J. Choi, D. Pugmire, S. Klasky, A. Loarte, and R. A. Pitts, "Role of Turbulent Separatrix Tangle in the Improvement of the Integrated Pedestal and Heat Exhaust Issue for Stationary-Operation Tokamak Fusion Reactors," *Nuclear Fusion* (April 2024), IOP Publishing. https://doi.org/10.1088/1741-4326/ad3b1e

Dominski, J., C. S. Chang, R. Hager, S. Ku, E. S. Yoon, and V. Pareil, "Neoclassical Transport of Tungsten Ion Bundles in Total-f Neoclassical Gyrokinetic Simulations of a Whole-Volume JET-Like Plasma," *Physics of Plasmas* (March 2024), AIP Publishing. https://doi.org/10.1063/5.0144509

Physics | Data, Learning

Physics-Inspired AI Ensemble for the Detection of Higher Order Wave Mode Signals of Spinning Binary Black Hole Mergers

PI Eliu Huerta, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Theta, Polaris



Scientific visualization of the gravitational wave emission of a binary black hole merger. *Image: ALCF Visualization and Data Analytics Team; Eliu Huerta, Argonne National Laboratory; and Roland Haas, University of Illinois Urbana-Champaign*

When two black holes orbit each other, they generate ripples in spacetime called gravitational waves. These waves carry away energy and angular momentum from the pair, causing them to spiral closer together. Eventually, this spiral ends in a powerful collision that forms a single, larger black hole. During this collision, the newly formed black hole can receive a "kick" from the gravitational waves, pushing it away. Scientists use AI technology to detect these gravitational waves, helping us learn more about black holes and the universe. Recently, an Argonne-led team developed a new class of AI models to detect higher order wave modes, similar in nature to piano overtones, emitted by spinning binary black hole mergers.

CHALLENGE Creating a production-scale framework for gravitational wave discovery poses several computational challenges. AI models describing the physics of higher order wave modes of spinning binary black hole mergers require a massive, multi-terabyte sized training dataset with tens of millions of modeled waveforms. They also require distributed training and inference algorithms that optimally utilize hundreds of GPUs on supercomputing platforms to reduce time to solution.

APPROACH To address these challenges, the researchers developed physics-inspired AI models, combining hybrid dilated convolutional neural networks to model short and long-range temporal features of gravitational waves, and graph neural networks to model the spatial correlations between the sky location of gravitational wave sources and the geographical location of the LIGO and Virgo gravitational wave observatories. The team used ALCF supercomputers to train several AI models, using distributed training over 256 A100 GPUs on Polaris. The models demonstrated strong scaling using up to 512 A100 GPUs.

RESULTS Using an ensemble of four AI models, the team distributed AI inference using 128 nodes on Theta and 128 A100 GPUs on Polaris, processing a decade of synthetic data in 3.5 hours, reporting only one misclassification. These AI models were then fine-tuned, using real gravitational wave data available at the Gravitational Wave Open Science Center, and subsequently used to process the entire month of February 2020, which is part of the O3b LIGO/Virgo observation run. This AI search found six gravitational waves, concurrently identified in Advanced LIGO and Advanced Virgo data, and reported zero false positives. This analysis was completed in one hour using a single A100 GPU.

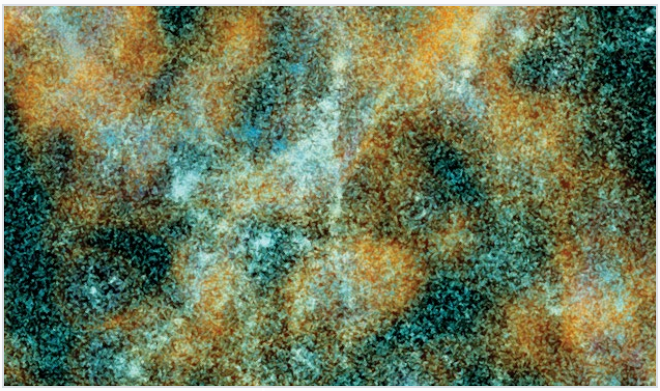
IMPACT This project represents the first time AI achieved this level of accuracy and sensitivity in identifying gravitational waves over decade-long datasets. This approach sets the stage to develop intuitive, physics-inspired AI that can detect the complete symphony of gravitational wave signals for black holes that describe astrophysically motivated, yet complex, gravitational waves, such as compact sources that spin and precess.

PUBLICATIONS
Tian, M., E. A. Huerta, H. Zheng, and P. Kumar. "Physics-Inspired Spatiotemporal-Graph AI Ensemble Detection of Higher Order Wave Mode Signals of Spinning Binary Black Hole Mergers," *Machine Learning: Science and Technology* (June 2024), IOP Publishing. https://doi.org/10.1088/2632-2153/ad4c37

Physics | Simulation, Data

Preparing HACC for Exascale Cosmology

PI Salman Habib, Argonne National Laboratory
Esteban Rangel, Argonne National Laboratory
AWARD Exascale Computing Project, Aurora ESP
SYSTEM Aurora



Researchers are using Aurora to perform extreme-scale cosmological simulations that will advance our understanding of the universe. *Image: ALCF Visualization and Data Analytics Team; HACC Collaboration*

HACC (Hardware/Hybrid Accelerated Cosmology Code) is a cosmological N-body and hydrodynamics simulation code designed to run at extreme scales. HACC computes the complicated emergence of structure in the universe across cosmological history, the core of the code's functionality consisting of gravitational calculations along with the more recent addition of gas dynamics and astrophysical sub-grid models. The solvers are integrated with a large set of sophisticated analysis methods encapsulated within HACC's CosmoTools library. With this project, Argonne researchers aimed to port HACC to SYCL to enable deployment on exascale supercomputers, including the ALCF's Aurora system.

CHALLENGE The increased computing power brought about by today's exascale systems has allowed HACC to support additional baryonic physics through a newly developed smoothed particle hydrodynamics formalism called conservative reproducing kernel. Versions of HACC being developed for exascale systems incorporate basic hydrodynamics to enable more detailed studies of structure formation on the scales of galaxy clusters and individual galaxies. These versions also include sub-grid models that integrate phenomena like star formation, supernova, and active galactic nuclei feedback.

APPROACH The research team ported HACC from a CUDA codebase to SYCL. Because "shuffle" operations used by HACC are not performance-portable on the Intel Data Center GPU Max 1550, a straightforward workaround using shared local memory was proposed. In bringing HACC to exascale, the developers have aimed to evaluate Aurora's early hardware and software development kit on a set of more than 60 complex kernels primarily written in CUDA or otherwise under active development, minimize divergence between CUDA and SYCL versions of the codebase, identify configurations and implementation optimizations specific to Intel GPUs, and identify more generally applicable implementation optimizations.

RESULTS Scaling on the pre-production Aurora system was demonstrated on up to 1792 nodes, and HACC simulations have been performed on Aurora in runs using as many as 1920 nodes. Visualizations of results generated on Aurora illustrate the large-scale structure of the universe. Single-GPU performance on Aurora using the Intel Data Center GPU Max 1550 exceeds that of compared systems: figure-of-merit assessments measuring particle-steps per second used 33 million particles per GPU and a 50 percent performance improvement over the AMD MI250X and NVIDIA A100.

IMPACT Modern cosmology provides a unique window to fundamental physics, and has led to remarkable discoveries culminating in a highly successful model for the dynamics of the universe. Simulations and predictions enabled by the HACC code deployed at exascale will help deepen our understanding of the structure of the universe and its underlying physics. Furthermore, new generations of cosmological instruments will depend on exascale systems to interpret the measurements. Exascale cosmological simulations developed through HACC will allow researchers to analyze observational data from state-of-the-art telescopes and test different theories of cosmological evolution simultaneously.

PUBLICATIONS

Rangel, E. M., S. J. Pennycook, A. Pope, N. Frontiere, Z. Ma, and V. Madananth. "A Performance-Portable SYCL Implementation of CRK-HACC for Exascale." *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis* (November 2023), ACM. https://doi.org/10.1145/3624062.3624187.

Physics | Simulation

Simulating the Cosmos for the Roman and Rubin Telescopes

PI Katrin Heitmann, Argonne National Laboratory
AWARD Director's Discretionary
SYSTEM Theta

This pair of simulations images shows the same region of sky as the Vera C. Rubin Observatory (left) and Nancy Grace Roman Space Telescope (right) will see it. *Image: J. Chiang, SLAC; C. Hirata, Ohio State University; and NASA's Goddard Space Flight Center*

The Vera C. Rubin Observatory and Nancy Grace Roman Space Telescope are state-of-the-art telescopes set to revolutionize our understanding of the universe when they begin operations in 2025 and 2027, respectively. To prepare for the vast amounts of observational data they will generate, researchers from DOE, NASA, and U.S. universities joined forces to use the ALCF's Theta supercomputer to produce nearly 4 million simulated images that depict the cosmos as the telescopes will see it.

CHALLENGE NASA's space-based Roman telescope and the ground-based Rubin telescope, jointly funded by the National Science Foundation (NSF) and DOE, will provide unprecedented views of the cosmos to help advance research on dark energy, dark matter, and the evolution of the universe. Because the two surveys will produce highly complementary datasets, joint data processing and analysis has been identified as a critical step for achieving the most robust and powerful cosmological and astrophysical results. This approach, however, requires powerful computing resources and advanced simulation techniques to develop methods that take full advantage of the overlapping survey data.

APPROACH Researchers from DOE (Argonne, SLAC), NASA, and academia partnered to use the ALCF's Theta supercomputer to produce a set of overlapping joint synthetic Roman-Rubin time-domain surveys. Carried out as part of the broader NASA-led OpenUniverse project, the team's simulations account for the telescopes' unique instrument performances, making them the most accurate predictions to date of what the telescopes will observe.

RESULTS The team leveraged Theta's processing power to generate approximately 4 million simulated images for the Rubin and Roman telescopes in about nine days—a task that would have taken around 300 years on a personal computer. The simulations cover the same patch of the sky, spanning 70 square degrees (roughly equivalent to the sky area covered by 350 full Moons). The team has released an initial 10-terabyte subset of the simulation data for the community to explore, with the remaining 390 terabytes to follow once all the data has been processed. The simulated images give researchers the opportunity to exercise their data processing pipelines, better understand their analysis codes, and accurately interpret the results so they are ready to use the real observational data as soon as it starts coming in.

IMPACT The team's simulations will help inform future research with the Roman and Rubin telescopes to advance our understanding of dark energy, dark matter, and the evolution of the universe. By overlapping the simulation data for each survey, scientists can learn how to use the best aspects of each telescope—Rubin's broader view and Roman's sharper, deeper vision. The combination will provide more precise and accurate insights than could be achieved from either observatory alone.

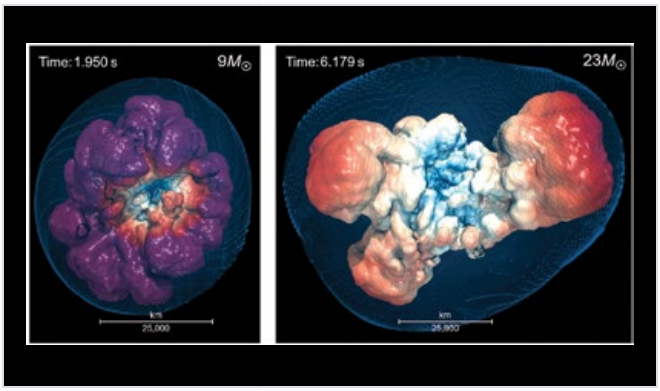
PUBLICATIONS

OpenUniverse Team. "OpenUniverse 2024 Simulated Roman & Rubin Images: Preview" (2024), ISRA. https://doi.org/10.26131/IRSA569

Physics | Simulation

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

PI Adam Burrows, Princeton University
AWARD INCITE
SYSTEM Polaris, Theta



This is a representative depiction of a generic feature of core-collapse explosions by the turbulence-aided neutrino-driven mechanism. *Image: ALCF Visualization and Data Analytics Team; Princeton University*

Supernovae shape the universe and life as we know it, but the physical mechanisms that cause a star to explode remain a mystery. Using ALCF's supercomputing resources, a team from Princeton University has been generating one of the largest collections of 3D supernova simulations to shed light on the physics behind these cosmic events.

CHALLENGE Modeling the physics of supernovae has posed a persistent challenge to astrophysicists for decades. Models must not only be shown to explode, but must reach the asymptotic state of the blast to determine many of the observables. Many relevant simulations have been developed. However, the required sophisticated and expensive 3D simulations typically have not run long enough after bounce to capture asymptotic kick speeds. Even the few existing longer-term studies do not explore the systematics with the broad range of progenitor masses to determine the relationships between the kick speed and progenitor mass or the initial core structure.

APPROACH The team carried out 20 state-of-the-art 3D long-term core-collapse simulations generated using the code FORNAX. They performed these simulations on multiple supercomputers, including the ALCF's Theta and Polaris systems. The researchers focused their study on the kicks during the simultaneous accretion and explosion phase, with attention towards the crucial first few seconds post-bounce. They then complemented their kick study with a study of the associated induced spins. The ALCF Catalyst team provided support to transition the code to Polaris (NVIDIA GPUs) and is working with the researchers to port the FORNAX GPU version to Aurora for future exascale simulations.

RESULTS For the first time, using a large and uniform collection of 3D supernova models ranging from 9- to 60-solar-mass stars, the researchers asymptoted the kicks or came within 20 percent of doing so. They obtained an integrated and wide-angle perspective of the overall dependence of the recoil kicks and induced spins upon progenitor mass and their Chandrasekhar-like core structures, the latter indexed approximately by compactness. The team found that the mass and compactness of the progenitors directly correlated to the size of the neutron star's kicks. These two classes can be correlated to the gravitational mass of the residual neutron star, which suggests the survival of binary neutron star systems may be due to their lower observed kick speeds. Their new 3D model suite provides a greatly expanded perspective and appears to explain some observed pulsar properties by default.

IMPACT The team's simulations represent the largest set of long-term 3D state-of-the-art core collapse simulations ever created. These simulations lay the groundwork for more comprehensive research that will address other aspects and outcomes of core collapse and their dependence upon progenitors. They also provide a qualitative picture, paving the way to develop a quantitative explanation of the survival of binary neutron star systems.

PUBLICATIONS

Burrows, A., T. Wang, D. Vartanyan, M. S. B. Coleman. "A Theory for Neutron Star and Black Hole Kicks and Induced Spins," *The Astrophysical Journal* (February 2024), American Astronomical Society. https://doi.org/10.3847/1538-4357/ad2353.

ALCF Projects

INCITE 2024

BIOLOGICAL SCIENCES

COMbining Deep-Learning with Physics-Based affinIty estimatiOn 3 (COMPBIO3)

PI Peter Coveney, University College London
HOURS ALCF: 330,000 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: 60,000 Node-Hours
OLCF: 240,000 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: 320,000 Node-Hours

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

PI Mohammed AlQuraishi, Columbia University
HOURS ALCF: 200,000 Node-Hours

COMPUTER SCIENCE

Democratizing AI by Training Deployable Open-Source Language Models

PI Abhinav Bhatele, University of Maryland
HOURS ALCF: 150,000 Node-Hours
OLCF: 600,000 Node-Hours

CHEMISTRY

Exascale Catalytic Chemistry

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

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

PI Anastassia Alexandrova, University of California, Los Angeles
HOURS ALCF: 250,000 Node-Hours

Heteropolymer Design Harnessing New and Emerging Computing Technologies

PI Vikram Mulligan, Flatiron Institute
HOURS ALCF: 375,000 Node-Hours

EARTH SCIENCE

Energy Exascale Earth System Model

PI Peter Caldwell, Lawrence Livermore National Laboratory
HOURS ALCF: 500,000 Node-Hours
OLCF: 1,000,000 Node-Hours

ENGINEERING

Flight-Scale Simulations of a Transport Aircraft in High-Lift Conditions

PI Eric Nielsen, NASA Langley Research Center
HOURS ALCF: 350,000 Node-Hours
OLCF: 1,000,000 Node-Hours

High Reynolds Number Hypersonic Transition Control via Porous Walls

PI Carlo Scalo, Purdue University
HOURS ALCF: 300,000 Node-Hours

Interface-Resolved Simulations of Scalar Transport in Turbulent Bubbly Flows

PI Parisa Mirbod, University of Illinois Chicago
HOURS ALCF: 50,000 Node-Hours

Online Machine Learning for Large-Scale Turbulent Simulations

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

MATERIALS SCIENCE

Carbon at Extremes: Discovery Science with Exascale Computers

PI Ivan Oleynik, University of South Florida
HOURS ALCF: 300,000 Node-Hours
OLCF: 1,000,000 Node-Hours

Exascale Simulations of Quantum Materials

PI Paul Kent, Oak Ridge National Laboratory
HOURS ALCF: 590,000 Node-Hours
OLCF: 600,000 Node-Hours

Heterogeneous Reaction Dynamics for Energy Storage and Hydrogen Production

PI Boris Kozinsky, Harvard University
HOURS ALCF: 150,000 Node-Hours
OLCF: 300,000 Node-Hours

PHYSICS

3D Imaging of Strong Interaction Nambu-Goldstone Bosons

PI Yong Zhao, Argonne National Laboratory
HOURS ALCF: 150,000 Node-Hours

Ab-initio Nuclear Structure and Nuclear Reactions

PI Gaute Hagen, Oak Ridge National Laboratory
HOURS ALCF: 500,000 Node-Hours
OLCF: 1,000,000 Node-Hours

Advances in Quark and Lepton Flavor Physics with Lattice QCD

PI Andreas Kronfeld, Fermilab
HOURS ALCF: 1,200,000 Node-Hours
OLCF: 1,000,000 Node-Hours

Exascale Cosmology: Lighting Up the Dark Universe

PI Katrin Heitmann, Argonne National Laboratory
HOURS ALCF: 375,000 Node-Hours
OLCF: 800,000 Node-Hours

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

PI CS Chang, Princeton Plasma Physics Laboratory
HOURS ALCF: 550,000 Node-Hours
OLCF: 1,000,000 Node-Hours

Exascale Models of Astrophysical Thermonuclear Explosions

PI Michael Zingale, Stony Brook University
HOURS ALCF: 75,000 Node-Hours
OLCF: 400,000 Node-Hours

Novel Calculation of Nucleon Generalized Form Factors in Lattice QCD at the Physical Point

PI Martha Constantinou, Temple University
HOURS ALCF: 315,000 Node-Hours

Radiation-Dominated Black Hole Accretion

PI James Stone, Institute for Advanced Study
HOURS ALCF: 425,000 Node-Hours
OLCF: 800,000 Node-Hours

Resolving Cosmic Ray Transport by Pushing the Frontier of MHD Turbulence

PI Drummond Fielding, Cornell University
HOURS ALCF: 200,000 Node-Hours
OLCF: 700,000 Node-Hours

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

PI Adam Burrows, Princeton University
HOURS ALCF: 1,000,000 Node-Hours

QCD Under Extreme Conditions

PI Zoltan Fodor, Pennsylvania State University
HOURS ALCF: 400,000 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

Autonomy for DOE Simulations

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

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

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

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
NERSC: 500,000 Node-Hours

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

PI Taehun Lee, City College of New York
HOURS ALCF: 115,000 Node-Hours

ENGINEERING

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

MATERIALS SCIENCE

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

Hadronic Contributions to the Muon g-2 from Lattice QCD

PI Thomas Blum, University of Connecticut
HOURS AALCF: 1,000,000 Node-Hours
OLCF: 3,846,000 Node-Hours

Simulating Large-Scale Long-Lived Neutron Star Remnants from Binary Neutron Star Mergers

PI Ore Gottlieb, Flatiron Institute
HOURS ALCF: 400,000 Node-Hours

ALCC 2023–2024

BIOLOGICAL SCIENCES

Probabilistic Comparative Modeling of Colorectal Cancer Screening Strategies

PI Jonathan Ozik, Argonne National Laboratory
HOURS ALCF: 160,000 Node-Hours
NERSC: 100,000

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

PI Ravi Madduri, Argonne National Laboratory
HOURS ALCF: 210,000 Node-Hours

CHEMISTRY

Microscopic Insight Into Transport Properties of Li-Battery Electrolytes

PI Wei Jiang, Argonne National Laboratory
HOURS ALCF: 710,000 Node-Hours

Relativistic Quantum Dynamics in the Non-Equilibrium Regime

PI Eugene DePrince, Florida State University
HOURS ALCF: 700,000 Node-Hours

ENERGY TECHNOLOGIES

Accelerating Deployment of Next-Generation Nuclear Power Using High-Fidelity CFD

PI Dillon Shaver, Argonne National Laboratory
HOURS ALCF: 500,000 Node-Hours

High Energy Density Physics of Inertial Confinement Fusion Ablator Materials

PI Ivan Oleynik, University of South Florida
HOURS ALCF: 500,000 Node-Hours
OLCF: 1,500,000 Node-Hours

Large Eddy Simulation on Flow and Heat Transfer Behavior in Involute Plate Research Reactor Supporting the Needs of the Materials Management and Minimization (M3) Reactor Conversion Program

PI Yiqi Yu, Argonne National Laboratory
HOURS ALCF: 510,000 Node-Hours

ENGINEERING

Two-Phase Flow Interface Capturing Simulations

PI Igor Bolotnov, North Carolina State University
HOURS ALCF: 200,000 Node-Hours
NERSC: 300,000 Node-Hours

MATERIALS SCIENCE

Computational Design of Novel Semiconductors for Power and Energy Applications

PI Feliciano Giustino, The University of Texas at Austin
HOURS ALCF: 100,000 Node-Hours

Large-Scale Simulations of Materials for Quantum Information Science

PI Giulia Galli, University of Chicago
HOURS ALCF: 600,000 Node-Hours
NERSC: 400,000 Node-Hours

Quantum Accurate Large-Scale Atomistic Simulations of Advanced Fusion Reactor Materials

PI Aidan Thompson, Sandia National Laboratories
HOURS ALCF: 850,000 Node-Hours
OLCF: 500,000 Node-Hours
NERSC: 250,000 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

Hadronic Contributions to the Muon g-2 from Lattice QCD

PI Thomas Blum, University of Connecticut
HOURS ALCF: 5,000 Node-Hours
OLCF: 3,283,000 Node-Hours

Using GPU to Reconstruct LHC Collisions Recorded with the CMS Detector

PI Dirk Hufnagel, Fermilab
HOURS ALCF: 70,000 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

Extreme-Scale Unstructured Adaptive CFD

PI Kenneth Jansen, University of Colorado Boulder

High-Fidelity Simulation of Fusion Reactor Boundary Plasmas

PI CS 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 Theresa Windus, Iowa State University and Ames 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

Accelerating Drug Discovery with Deep Learning

PI Archit Vasan, Argonne National Laboratory

APACE: AlphaFold2 and Advanced Computing as a Service for Accelerated Discovery in Biophysics

PI Eliu Huerta, Argonne National Laboratory

Isochronic Development of Cortical Synapses in Primates and Mice

PI Gregg Wildenberg, Argonne National Laboratory and University of Chicago

Probabilistic Comparative Modeling of Colorectal Cancer Screening Strategies

PI Jonathan Ozik, Argonne National Laboratory

Protein Generation via Genome-scale Language Models with Bio-physical Scoring

PI Arvind Ramanathan, Argonne National Laboratory; Venkatram Vishwanath, Argonne National Laboratory

COMPUTER SCIENCE

Demonstrating Cross-Facility Data Processing at Scale with Laue Microdiffraction

PI Michael Prince, Argonne National Laboratory

Efficient Algorithms for Monte Carlo Particle Transport on AI Accelerator Hardware

PI John Tramm, Argonne National Laboratory

Evidence of Scaling Advantage for the Quantum Approximate Optimization Algorithm on a Classically Intractable Problem

PI Yuri Alexeev, Argonne National Laboratory

Linking the Dynamic PicoProbe Analytical Electron-Optical Beam Line / Microscope to Supercomputers

PI Arvind Ramanathan, Argonne National Laboratory

Thorough Characterization and Analysis of Large Transformer Model Training At-Scale

PI Venkatram Vishwanath, Argonne National Laboratory

ENGINEERING

Bifurcation of Equilibrium Positions for Ellipsoidal Particles in Inertial Shear Flows Between Two Walls

PI Zhangli Peng, University of Illinois Chicago

MATERIALS SCIENCE

Data-Driven Materials Discovery for Functional Applications

PI Jacqueline Cole, University of Cambridge

Freedom of Design in Chemical Compound Space: Towards Rational In Silico Design of Molecules with Targeted Quantum-Mechanical Properties

PI Alexandre Tkatchenko, University of Luxembourg

Imaging 3D Chemistry at 1 nm Resolution with Fused Multi-Modal Electron Tomography

PI Robert Hovden, University of Michigan

Materials Informatics Study of Two-Dimensional Magnetic Materials

PI Trevor Rhone, Rensselaer Polytechnic Institute

Physics-Inspired Spatiotemporal-Graph AI Ensemble for the Detection of Higher Order Wave Mode Signals of Spinning Binary Black Hole Mergers

PI Eliu Huerta, Argonne National Laboratory

Simulating the Cosmos for the Roman and Rubin Telescopes

PI Katrin Heitmann, Argonne National Laboratory

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-24/3)

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 (Director of Operations), Susan Coghlan (ALCF-X Project Director), Kalyan Kumaran (Director of Technology), Jini Ramprakash (Deputy Division 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.