



TPC 25

Developing best practices for
utilizing AI for scientific discovery
and engineering at scale.

Monday-Thursday, July 28-31
DoubleTree by Hilton San Jose

WELCOME TO TPC!

We're delighted that you'll be joining the Trillion Parameter Consortium's annual all-hands meeting in San Jose. Since its founding in 2023, TPC has united researchers who push the boundaries of AI and science. This four-day conference convenes 300 participants from over 100 universities, national laboratories, companies, and institutes and 14 countries to pursue that mission on three fronts:

1. Building Community: Sharing Knowledge and Tools

The plenary program brings together two dozen international leaders to survey recent breakthroughs in AI-driven science and to frame the challenges — and opportunities — posed by today's frontier models and tomorrow's super-scale systems.

2. Building Collaborations

Eighteen current or prospective working groups will meet across 30 parallel breakout sessions. Each of these 90-minute breakout sessions is anchored by 4–6 lightning talks that highlight emerging results. Lively discussions will follow, aimed at matching complementary expertise and launching new joint projects.

3. Growing Capacity

The opening day and a half feature three hands-on tutorials and a hackathon focused on agentic systems for science. Participants will leave with practical skills and a network of peers to draw on for future quarterly hackathons and collaborative coding sprints.

We encourage you to explore every facet of the program — especially the breakouts — where serendipitous conversations often spark the next big idea. Thank you for bringing your expertise to TPC25; we look forward to the discoveries that will follow.



Rick Stevens
TPC Executive Committee, Argonne National
Laboratory | The University of Chicago



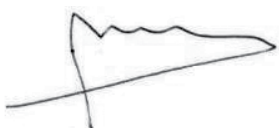
Charlie Catlett
TPC25 Executive Director, Argonne National
Laboratory | The University of Chicago



Satoshi Matsuoka
TPC Executive Committee, RIKEN R-CCS



Tom Tabor
TPC25 Executive Producer, Tabor Communications



Mateo Valero
TPC Executive Committee, Barcelona
Supercomputing Center



AGENDA-AT-A-GLANCE

MONDAY, JULY 28			
9:00	HACKATHON / TUTORIAL OPENING PLENARY		
10:30	BREAK		
11:00	HACKATHON SESSION 2	TUTORIAL SESSIONS 2	
13:00	LUNCH IN GATEWAY BALLROOM		
14:00	HACKATHON SESSION 3	TUTORIAL SESSIONS 3	
15:30	BREAK		
16:00	HACKATHON SESSION 4	TUTORIAL SESSIONS 4	
17:30	BREAK		
18:00-19:30	HAMMERSPACE HACKATHON/TUTORIAL NETWORKING GATHERING		
TUESDAY, JULY 29			
9:00	HACKATHON SESSION 5	TUTORIAL SESSIONS 5	EXHIBITION
10:30	BREAK		
10:45	HACKATHON SESSION 6	TUTORIAL SESSIONS 6	
13:00	LUNCH		
14:00	OPENING PLENARY		
15:30	BREAK		
16:00	PLENARY 2		
17:30			
18:00-19:30	GOOGLE WELCOME RECEPTION		
WEDNESDAY, JULY 30			
9:00	PLENARY 3		EXHIBITION
10:30	BREAK		
11:00	PLENARY 4		
12:30	LUNCH & PANEL DISCUSSION		
14:00	PARALLEL BREAKOUTS A		
15:30	BREAK		
16:00	PARALLEL BREAKOUTS B		
THURSDAY, JULY 31			
8:30	PARALLEL BREAKOUTS C		
10:30	BREAK		
11:00	PARALLEL BREAKOUTS D		
12:30	LUNCH & PANEL DISCUSSION		
14:00	PARALLEL BREAKOUTS E		
15:30	BREAK		
16:00	PLENARY 5: CLOSING SESSION		

AGENDA

MONDAY, JULY 28		
9:00	HACKATHON/TUTORIAL OPENING PLENARY: INTRODUCTION TO AI FOR SCIENCE Moderator: Neeraj Kumar, <i>Pacific Northwest National Laboratory</i> Advancing Science and Medicine with AI Physician-scientists Vivek Natarajan, <i>Google DeepMind</i> Session 1: Plenary session with all Tutorial and Hackathon participants: Foundations in AI for Science	
10:30	COFFEE BREAK	
11:00	HACKATHON SESSION 2 <i>Room: Cascade</i> Building Agentic Systems for Science Session 2: Intro to Agentic Systems and Use Cases	TUTORIAL SESSIONS 2 <div> AI for Science: Foundations and Frontiers Session 2: Case Studies and Emerging Frontiers in AI for Science <i>Room: Siskiyou</i> </div> <div> Evaluation of AI Model Scientific Reasoning Skills Session 2: Use Cases and Basic Evaluation Techniques <i>Room: Donner</i> </div>
12:30	LUNCH	
14:00	HACKATHON SESSION 3 <i>Room: Cascade</i> Building Agentic Systems for Science Session 3: Team Formation and Project Kickoff	TUTORIAL SESSIONS 3 <div> AI for Science: Foundations and Frontiers Session 3: Parallelization Strategies for Large-scale Pre-training <i>Room: Siskiyou</i> </div> <div> Evaluation of AI Model Scientific Reasoning Skills Session 3: Advanced Evaluation Technique <i>Room: Donner</i> </div>
15:30	COFFEE BREAK	
16:00	HACKATHON SESSION 4 <i>Room: Cascade</i> Building Agentic Systems for Science Session 4: Hands-on Hacking with Expert Mentorship	TUTORIAL SESSIONS 4 <div> AI for Science: Foundations and Frontiers Session 4: Fine-tuning Techniques: From Theory to Practice <i>Room: Siskiyou</i> </div> <div> Evaluation of AI Model Scientific Reasoning Skills Session 4: Hands On <i>Room: Donner</i> </div>
17:30	BREAK	
18:00-19:30	HAMMERSPACE HACKATHON/TUTORIAL NETWORKING GATHERING	

TUESDAY, JULY 29		
9:00	HACKATHON SESSIONS 5 <i>Room: Cascade</i> Building Agentic Systems for Science Session 5: Midpoint Sync, Debugging, Breakouts	TUTORIAL SESSIONS 5 <div> AI for Science: Foundations and Frontiers Session 5: Profiling AI Workloads with PARAVR <i>Room: Siskiyou</i> </div> <div> Using AI to Accelerate Day-to-Day Scientific Productivity Session 1: LLM Refresher + Deep Research & Idea Generation <i>Room: Donner</i> </div>
10:30	COFFEE BREAK	
10:45	HACKATHON SESSION 6 <i>Room: Cascade</i> Building Agentic Systems for Science Session 6: Project Showcases, Wrap-up Discussion	TUTORIAL SESSIONS 6 <div> AI for Science: Foundations and Frontiers Session 6: Building RAG-based Workflows OR AI Agents <i>Room: Siskiyou</i> </div> <div> Using AI to Accelerate Day-to-Day Scientific Productivity Session 2: Coding Faster & Better (Usually) + AI-enabled Science Applications <i>Room: Donner</i> </div>



TUESDAY, JULY 29, continued	
12:30	LUNCH IN GATEWAY BALLROOM
14:00	<p>Opening Plenary Session</p> <p>SCALING UP TO GW DATA CENTERS AND AI FACTORIES</p> <p>Reinventing Discovery: Accelerating Science in the Age of Artificial Super-Intelligence Rick Stevens, <i>Argonne National Laboratory</i></p> <p>HPC and Science: The Need for Hybrid Thierry Pellegrino, <i>AWS</i></p> <p>Modeling and Simulating Complex Behavior in Dynamic Cyber-Physical-Social Systems Flora Salim, <i>University of New South Wales</i></p>
15:30	COFFEE BREAK
16:00	<p>Plenary Session 2</p> <p>TPC AND AI FOR SCIENCE TWO YEARS LATER: NEW DIRECTIONS IN CONVERGENCE OF AI AND HPC</p> <p>"Some" Challenges for Using LLMs/ML in Science Moderator: Satoshi Matsuoka, <i>RIKEN R-CCS</i></p> <p>Enabling Scientific Discovery With Generative Quantum AI Steve Clark, <i>Quantinuum</i></p> <p>An Overview of Recent Studies of the Use of AI for Technical Computing Workloads Earl Joseph, <i>Hyperion Research</i></p> <p>Secure AI Infrastructure for Scientific Computing and General-purpose Applications at RIKEN Jens Domke, <i>RIKEN R-CCS</i></p>
17:30	BREAK
18:00-19:30	GOOGLE WELCOME RECEPTION

WEDNESDAY, JULY 30	
9:00	<p>Plenary Session 3</p> <p>AI AND THE FUTURE OF SCIENTIFIC DISCOVERY</p> <p>Scaling Reasoning, Scaling Science: Engineering an AI-Native Scientific Discovery Platform Moderator: Ian Foster, <i>Argonne National Laboratory</i></p> <p>Agents, Autonomy, and Agency: A Brave New World Preeth Chengappa, <i>Microsoft</i></p> <p>The Automation of Biological Discovery with Language Model Agents Siddharth Narayanan, <i>FutureHouse</i></p> <p>Active Inference AI Systems for Scientific Discovery Karthik Duraisamy, <i>University of Michigan</i></p>
10:30	COFFEE BREAK
11:00	<p>Plenary Session 4</p> <p>MULTIMODAL DATA, EVALUATION, AND NON-LLM MODEL ARCHITECTURES</p> <p>Responsible AI Moderator: Ricardo Baeza-Yates, <i>Barcelona Supercomputing Center</i></p> <p>ORNL's AI Initiative: Advancing Secure, Assured, and Efficient AI for Scientific Discovery Prasanna Balaprakash, <i>Oak Ridge National Laboratory</i></p> <p>OLMoTrace: Tracing LM Output Back to its Multi-trillion-token Training Data in Real Time Jiacheng Liu, <i>Allen Institute for AI</i></p> <p>Fairness of Geospatial Foundation Models Kyoung-Sook Kim, <i>National Institute of Advanced Industrial Science and Technology (AIST)</i></p>
12:30	<p>Lunch & Panel Discussion</p> <p>INDUSTRY, ACADEMIA, AND GOVERNMENT COLLABORATION: ACCELERATING TRUSTWORTHY AI FOR SCIENCE Moderator: Karthik Duraisamy, <i>University of Michigan</i></p> <p>Hal Finkel, <i>U.S. Department of Energy</i></p> <p>Raj Hazra, <i>Quantinuum</i></p> <p>Pradeep Dubey, <i>Intel</i></p> <p>Molly Presley, <i>Hammerspace</i></p>

WEDNESDAY, JULY 30, continued						
14:00	PARALLEL BREAKOUTS A <div> <div>Workflows Open Slot</div> <div>Initiatives BOF: Building Foundation Models for the Electric Grid (GridFM)</div> <div>Life Sciences AI for Cancer</div> <div>Evaluation Model Skills, Reasoning, and Trust Evaluation (EVAL) <i>Intro and Benchmarks</i></div> <div>Scale & Services BOF: Deployment of Inference-for-Science Services at HPC Centers <i>Session 1</i></div> <div>Applications AI Models for Software Engineering and Development</div> </div>					
15:30	COFFEE BREAK					
16:00	PARALLEL BREAKOUTS B <div> <div>Workflows Data, Workflows, Agents, and Reasoning Frameworks (DWARF) <i>Keynote and Systems Software for Agents</i></div> <div>Initiatives BOF: Leveraging ICICLE for TPC Applications Across the Computing Continuum</div> <div>Life Sciences Agentic AI and Foundation Models</div> <div>Evaluation Model Skills, Reasoning, and Trust Evaluation (EVAL) <i>UQ and Safety</i></div> <div>Scale & Services BOF: Deployment of Inference-for-Science Services at HPC Centers <i>Session 2</i></div> <div>Applications BOF: Foundation Models for Fusion Energy</div> </div>					

THURSDAY, JULY 31						
8:30	PARALLEL BREAKOUTS C <div> <div>Workflows Data, Workflows, Agents, and Reasoning Frameworks (DWARF) <i>Scalable Scientific Data/Scientific Data for AI</i></div> <div>Initiatives BOF: AI in Decision Sciences</div> <div>Life Sciences AI for Biology</div> <div>Evaluation Model Skills, Reasoning, and Trust Evaluation (EVAL) <i>Automatic Benchmark Generation</i></div> <div>Scale & Services Model Architecture and Performance Evaluation (MAPE) <i>Session 1</i></div> <div>Applications AI for Scientific Discovery in Materials Science (AI4MS)</div> </div>					
10:30	COFFEE BREAK					
11:00	PARALLEL BREAKOUTS D <div> <div>Workflows Data, Workflows, Agents, and Reasoning Frameworks (DWARF) <i>Scalable Processing Pipelines</i></div> <div>Initiatives BOF: Public AI: Policy, Community, and the Future of National Labs</div> <div>Life Sciences HPC-AI Society Meeting From the TPC25 Conference</div> <div>Evaluation Model Skills, Reasoning, and Trust Evaluation (EVAL) <i>Advance Evaluation</i></div> <div>Scale & Services Model Architecture and Performance Evaluation (MAPE) <i>Session 2</i></div> <div>Applications Education and Outreach</div> </div>					
12:30	Lunch & Panel Discussion BUILDING AGENTIC SYSTEMS FOR SCIENCE: REPORTS FROM THE FIELD Moderator: Addison Snell, <i>Intersect360 Research</i> Elahe Vedadi, <i>Google DeepMind</i> Preeth Chengappa, <i>Microsoft Discovery</i> Kexin Huang, <i>Stanford University (Boimni project)</i> Arvind Ramanathan, <i>Argonne National Laboratory (Scientia project)</i> Siddharth Narayanan, <i>FutureHouse (Agentic life sciences project)</i>					
14:00	PARALLEL BREAKOUTS E <div> <div>Workflows BOF: LLMs for Living Docs</div> <div>Initiatives BOF: Energy Efficient HPC for AI <i>Workloads</i></div> <div>Life Sciences BOF: Federated Learning at Scale</div> <div>Evaluation BOF: LLMs and Scientific Reasoning</div> <div>Scale & Services Model Architecture and Performance Evaluation (MAPE) <i>Session 3</i></div> <div>Applications Earth and Environment (AI for Digital Earth)</div> </div>					
15:30	COFFEE BREAK					
16:00	Plenary Session 5 SCIENCE UPDATES FROM KEY TPC LEADERS Recent Progress on Japanese LLMs Rio Yokota, <i>Institute for Science Tokyo</i> EAIRA: Establishing a Methodology for Evaluating AI Models as Scientific Research Assistants Franck Cappello, <i>Argonne National Laboratory</i> CLOSING PANEL: THE FUTURE OF SCIENCE AND SOCIETY ENTERING THE ERA OF ARTIFICIAL SUPER INTELLIGENCE Moderator: Charlie Catlett, <i>Trillion Parameter Consortium</i> Ian Foster, <i>Argonne National Laboratory</i> Karthik Duraisamy, <i>University of Michigan</i> Satoshi Matsuoka, <i>RIKEN R-CCS</i> Thierry Pellegrino, <i>AWS</i>					



Plenary Sessions

Opening Plenary

SCALING UP TO GW DATA CENTERS AND AI FACTORIES

Reinventing Discovery: Accelerating Science in the Age of Artificial Super-Intelligence

Moderator: Rick Stevens, *Argonne National Laboratory*

Frontier AI models have crossed a threshold. They no longer merely assist scientists, but now co-design not only which questions to pursue, but how to pursue them. This keynote examines how we can accelerate scientific discovery using these advanced models. Drawing an analogy to Amdahl's Law, we'll see how extraordinary speed-ups in hypothesis generation, simulation, and data interpretation collide with bottlenecks in chemistry, fabrication, and field observation, forcing a strategic rebalancing of the entire research pipeline. We'll explore embedding human values in autonomous goal setting, preserving trust and reproducibility amid synthetic data, and redesigning the workforce to align automated cognition with irreplaceable human judgment. Last, we'll introduce high-level considerations and concrete actions to collectively explore how we can navigate this rapidly changing landscape.

HPC and the Science: The Need for Hybrid

Thierry Pellegrino, *AWS*

There is a strategic necessity for hybrid approaches in modern HPC. This talk explores three critical dimensions transforming scientific computing today: the integration of AI with traditional HPC workflows, seamless extension of on-premises infrastructure to the cloud, and the convergence of quantum and classical computing. Through concrete examples from weather forecasting, computational fluid dynamics, and molecular modeling, we'll demonstrate how AWS's hybrid solutions deliver measurable advantages, dramatically reducing computation time while providing access to next-generation hardware without capital expenditure constraints. As computational demands grow exponentially across scientific disciplines, hybrid architectures emerge as the practical foundation for accelerating scientific discovery, while enabling organizations to fully leverage the scalability, flexibility, and innovation velocity that cloud computing provides.

Modeling and Simulating Complex Behavior in Dynamic Cyber-Physical-Social Systems

Flora Salim, *University of New South Wales*

Understanding and anticipating complex dynamic behavior is fundamental to both computational social science and the scientific modeling of socio-technical systems. Behaviors of humans and systems in the wild could unfold dynamically — often shaped by diverse contexts and evolving intentions. Yet data capturing real-world behaviors — such as mobility routines, energy use patterns, and decision-making in urban and digital environments — are inherently noisy, context-dependent, and often only partially observed. This talk synthesizes recent progress in understanding behavior at scale through data-driven modeling and simulation, highlighting the convergence of data-efficient learning, generative models, and agentic AI for complex systems analysis.

Drawing from years of work in spatiotemporal data mining and multimodal machine learning, we examine how AI systems are now capable of learning from sparse signals while generalizing across heterogeneous settings. These advances reveal how latent routines, dynamics, and behavioral patterns can be learned without explicit ground-truth supervision. We will also demonstrate the use of LLMs for synthetic data generation. These approaches reflect a shift toward data-efficient, transferable, and context-sensitive models that are aimed at generalization beyond narrow domains.

We also discuss the rise of agentic AI for enabling automated tooling and simulation. We will present our new cyber-physical-social simulation generation framework, enabling automated scenario generation, behavior testing, and what-if analysis. This framework opens new possibilities for integrating empirical data with simulated environments.

Plenary 2

TPC AND AI FOR SCIENCE TWO YEARS LATER: NEW DIRECTIONS IN CONVERGENCE OF AI AND HPC

"Some" Challenges For Using LLMs/ML In Science

Moderator: Satoshi Matsuoka, *RIKEN R-CCS*

Large-scale language and learning models are poised to transform computational science, yet integrating them with exascale simulations, streaming experiments, and deep domain priors remains an open research frontier. This talk draws on RIKEN's AI4Science effort and recent Fugaku/Frontier deployments to outline a technical agenda for science-grade LLMs: scalable tokenization and sparse attention for multi-terabyte, multi-modal inputs; adaptive patching and sequence-tiling that push ViTs into the billion-token regime; cross-channel and physics-constrained operators that fuse neural surrogates with PDE solvers; and distributed orchestration of ensembles and agent workflows across CPU-GPU-QC hybrids. Performance results — up to 1.8 EF/s and 6.9x speedups — are presented for climate down-scaling, high-resolution 3-D imaging, and materials discovery, followed by open challenges in provenance, uncertainty quantification, and human-in-the-loop steering for petascale-and-beyond ML stacks.

Enabling Scientific Discovery With Generative Quantum AI

Steve Clark, *Quantinuum*

By leveraging the highest fidelity quantum systems in tandem with HPC, we can now harness previously unobtainable information to expand the scope of AI datasets, enabling fine-tuned, bespoke AI techniques with new capabilities to discover solutions for complex problems in areas like drug discovery and materials science — a bold initiative we call Generative Quantum AI. This talk will shed light on new territory for HPC to revolutionize scientific discovery of commercial relevance by unlocking value at the intersection of quantum computing and AI.

An Overview of Recent Studies of the Use of AI for Technical Computing Workloads

Earl Joseph, *Hyperion Research*

Hyperion Research has recently conducted a number of studies on the use of AI at technical computing (or HPC) centers around the world. This presentation will cover the highlights and key findings from these studies, including: AI use cases and the level of AI usage at HPC sites; growth rate of applying AI to technical workloads; budgets, cloud usage, frameworks, use of LLMs, hardware preferences, etc.; and barriers to applying AI more quickly and attributes that sites would like to see improved.

Secure AI Infrastructure for Scientific Computing and General-purpose Applications at RIKEN

Jens Domke, *RIKEN R-CSS*

This talk presents RIKEN's approach to building a secure, open-source AI infrastructure tailored for scientific and industrial use. We showcase how frontier open models, agentic AI, and RAG pipelines can be deployed locally, integrated tightly with HPC systems and simulation apps, while preserving privacy and performance. We will also share our experiences testing RAGFlow and our plans to set up a model serving stack for applications like the Spring-8 light source. Lastly, we highlight our work to eliminate inefficient cross-process communication by enabling large models to run directly in the memory space of our HPC simulations written in C, C++, or Fortran.

Plenary 3

AI AND THE FUTURE OF SCIENTIFIC DISCOVERY

Scaling Reasoning, Scaling Science: Engineering an AI-native Scientific Discovery Platform

Moderator: Ian Foster, *Argonne National Laboratory | University of Chicago*

Trillion-parameter reasoning engines promise great cognitive power, yet scientific breakthroughs still hinge on integrating models with experiments, data, and people. Let's examine an AI-native Scientific Discovery Platform that couples frontier language models to simulation codes, knowledge graphs, and autonomous laboratories, orchestrated by policy-aware scheduling and ultra-low-latency "thought-action" fabrics. The platform will autonomously generate hypotheses, execute in-silico or in-vitro experiments, assess uncertainty, and iteratively refine understanding, all while preserving provenance. I'll outline needed advances in data logistics, model-simulation co-execution, and trustable evaluation, and argue that careful co-design of infrastructure, benchmarks, and culture is vital to realize this potential.

Agents, Autonomy, and Agency: A Brave New World

Preeth Chengappa, *Microsoft*

With Agents set to proliferate across every facet of our lives, we discuss the state-of-the-art and the state of the future. We also explore complex questions of ownership, agency, control, decision-making and outcomes in an agentic world.

The Automation of Biological Discovery with Language Model Agents

Siddharth Narayanan, *FutureHouse*

The intellectual bottlenecks of science are growing, as evidenced by the increasing complexity of fields, the volume of scientific

papers published, and the number of humans involved. To manage this complexity, it is likely that major breakthroughs will increasingly rely on automation of the stages of the scientific method. One approach has been scientific agents — AI models equipped with tools and data to manipulate and observe the world. Such systems are increasingly automating tasks such as literature research, hypothesis generation, and data analysis. They can scale in dimensions beyond what has been previously possible, like checking every claim of a paper against all previous literature for disagreement. In this talk, we'll examine work at FutureHouse to apply scientific agents across stages of biological research and discovery, including challenges faced in defining well-posed problems, scaling compute, and evaluating AI scientist performance.

Active Inference AI Systems for Scientific Discovery

Karthik Duraisamy, *University of Michigan*

Current AI systems, while already useful in accelerating some aspects of scientific research, remain fundamentally limited by their operational architectures, brittle reasoning mechanisms, and separation from reality. Progress in AI-driven science now depends on closing three fundamental gaps — the abstraction gap, the reasoning gap, and the reality gap — rather than on model size/data/test time compute. Scientific reasoning demands internal representations that support simulation of actions and response, causal structures that distinguish correlation from mechanism, and continuous calibration. A vision is proposed for "active inference" AI systems: a multi-layered stack where discovery arises from the interplay between internal models that enable counterfactual reasoning and external validation that grounds hypotheses in reality. It is also argued that human judgment is indispensable, not as a temporary scaffold but as a permanent architectural component.

Plenary 4

MULTIMODAL DATA, EVALUATION, AND NON-LLM MODEL ARCHITECTURES

Responsible AI

Moderator: Ricardo Baeza-Yates, *Barcelona Supercomputing Center*

To set the stage, we will cover irresponsible AI: (1) discrimination (e.g., facial recognition, justice); (2) pseudoscience (e.g., biometric based predictions); (3) limitations (e.g., human incompetence, minimal adversarial AI); (4) indiscriminate use of computing resources (e.g., large language models); and (5) the impact of generative AI (disinformation, mental health and copyright issues). These examples do have a personal bias, but set the context for the second part, where we address three challenges: (1) principles and governance; (2) regulation; and (3) our cognitive biases. We will finish by discussing responsible AI initiatives and the near future.

ORNL's AI Initiative: Advancing Secure, Assured, and Efficient AI for Scientific Discovery

Prasanna Balaprakash, *Oak Ridge National Laboratory*

Oak Ridge National Laboratory's Artificial Intelligence Initiative is driving the advancement of AI methods to accelerate discovery and innovation in science, energy, and national security. With access to world-class computational resources, including the Frontier exascale system, the initiative prioritizes the development of AI foundation models and adaptive AI systems tailored to non-language modalities such as time series, spatial-temporal, multimodal sensor data, and physics-based simulations.



These efforts integrate physics-informed learning, uncertainty quantification, and causal reasoning to enable robust, explainable AI applications in complex scientific environments. The initiative supports a diverse portfolio and industry collaborations, spanning strategic domains such as nuclear energy, materials discovery, and national security. It also plays a key role in workforce development through the AI Academy, which engages over 200 researchers across directorates. Through targeted investments and cross-cutting coordination with ORNL's experimental facilities, the AI Initiative is enabling next-generation AI systems that go beyond language and LLMs to transform and accelerate scientific discovery.

OLMoTrace: Tracing LM Output Back to its Multi-trillion-token Training Data in Real Time

Jiacheng Liu, *Allen Institute for AI*

As LLMs gain adoption in higher-stakes scenarios, it is critical to understand why they generate certain responses. To address this challenge, we developed OLMoTrace — a system that traces outputs of LLMs back into their multi-trillion-token training data in real time. Given an LLM response to a user prompt, OLMoTrace finds long and unique spans in this response that appear verbatim in the training data, and shows users these spans and their enclosing training documents.

The purpose is to reveal where LLMs may have learned to generate certain word sequences. By combining algorithmic innovations and low-level system optimizations, our production system can return tracing results (i.e., spans and documents) in 4.5 seconds for typical LLM responses (~450 tokens), making it a real-time experience for users. We found OLMoTrace to be useful for fact-checking LLM outputs, understanding LLM hallucinations, tracing the “creative expressions” generated by LLMs, tracing some of their math capabilities, debugging erratic model outputs, etc. OLMoTrace is available in the public Ai2 model playground so that anyone can use this tool to trace the outputs of OLMo 2 and Tulu 3 models.

To our knowledge, OLMoTrace is the first system to scale up model behavior tracing beyond the trillion-token ballpark. Complementary to the body of work in mechanistic interpretability that traces LLM outputs into model weights and circuits, OLMoTrace traces directly into the training data, serving as an important piece in our model understanding toolbox.

Fairness of Geospatial Foundation Models

Kyoung-Sook Kim, *Deputy Director of Intelligent Platforms Research Institute, Advanced Industrial Science and Technology (AIST)*

GeoAI foundation models hold great promise for applications in diverse fields such as disaster prevention, environmental monitoring, and urban planning. However, their widespread adoption is hindered by significant challenges, including issues related to data quality, model transferability across regions, and the presence of geographical bias. A critical issue that warrants attention is the measurement of geo-bias in these models. Addressing this is essential for ensuring fairness, robustness, and generalizability. In this context, it is important to explore the current limitations and future directions for developing effective geo-bias metrics, while also discussing the broader challenges and opportunities for achieving sustainable and equitable GeoAI deployment.

WEDNESDAY LUNCH & PANEL

Industry, Academia, and Government Collaboration: Accelerating Trustworthy AI for Science

Moderator: Karthik Duraisamy, *University of Michigan*

This cross-sector panel convenes leaders from industry, academia, national laboratories, and government to chart a collaborative roadmap for trustworthy AI for science. We will dissect the technical and governance hurdles that impede reliable scientific AI and debate the institutional mechanisms — algorithmic improvements, shared testbeds, open data standards, federated model hubs, and national infrastructure — needed to overcome them. Panelists will assess how near-term breakthroughs in computing, advanced accelerators, and HPC-AI convergence could redefine algorithmic efficiency and scientific discovery over the next 3–5 years, while confronting inequities in compute access across sectors.

Panelists: Hal Finkel, *U.S. Department of Energy*
Raj Hazra, *Quantinuum*
Pradeep Dubey, *Intel Labs*
Molly Presley, *Hammerspace*

THURSDAY LUNCH & PANEL

Building Agentic Systems for Science: Reports From the Field

Moderator: Addison Snell, *Intersect360 Research*

Agentic AI systems — collections of autonomous, goal-seeking entities that plan, act, and learn across open-ended environments — have moved from research prototypes to production pipelines, yet the field still lacks a shared formal definition. This panel convenes builders of agentic platforms from national labs, academia, and industry to dissect what differentiates an ‘agent’ from a sophisticated subroutine, compare architectural families (LLM-centric tool use, multi-agent swarms, hybrid symbolic-neural controllers), and discuss design and implementation trade-offs among scalability, maintainability, security, and standards-driven interoperability. Through concrete case studies the speakers will expose design heuristics, failure modes, and metrics that matter, offering the TPC25 community a roadmap for considering frameworks and hardening agentic workflows for scientific discovery.

Panelists: Elahe Vedadi, *Google DeepMind*
Preeth Chengappa, *Microsoft Discovery*
Kexin Huang, *Stanford University (Boimni project)*
Arvind Ramanathan, *Argonne National Laboratory (Scientia project)*
Siddharth Narayanan, *FutureHouse (Agentic life sciences project)*

Plenary 5

SCIENCE UPDATES FROM KEY TPC LEADERS

Recent Progress on Japanese LLMs

Rio Yokota, *Institute for Science Tokyo*

LLMs are not mechanical tools that provide the same benefit for everyone. Rather, they are intellectual tools that interact with human culture and creativity. The influence is mutual, as not only are the models affected by the data we train on, but our culture and the data we generate will be influenced by LLMs. Therefore, it is paramount to develop sovereign AI models that adhere to

our cultural norms and acquire the technology to develop such models independently. This talk will showcase some of the major efforts to train Japanese LLMs, focusing on training data, training methodologies, and challenges.

EAIRA: Establishing a Methodology for Evaluating AI Models as Scientific Research Assistants

Franck Cappello, *Argonne National Laboratory*

Recent advancements have positioned Large Language Models (LLMs) as transformative tools for scientific research, capable of addressing complex tasks that require reasoning, problem-solving, and decision-making. Their exceptional capabilities suggest their potential as scientific research assistants, but also highlight the need for holistic, rigorous, and domain-specific evaluation to assess effectiveness in real-world scientific applications.

First, this talk motivates and describes the current effort at Argonne National Laboratory to develop a multifaceted methodology for evaluating AI models as scientific Research Assistants (EAIRA). This methodology incorporates four primary classes of evaluations:

1) Multiple Choice Questions to assess factual recall; 2) Open Response to evaluate advanced reasoning and problem-solving skills; 3) Lab-Style Experiments involving detailed analysis of capabilities as research assistants in controlled environments; and 4) Field-Style Experiments to capture researcher-LLM interactions at scale in a wide range of scientific domains and applications.

For each of these four classes of evaluation, we develop testing methods (e.g., benchmarks) and tools for manual and automatic QA generation and validation, as well as for collecting and analyzing researcher-LLM interactions.

We will present a selection of tools and generated benchmarks, as well as the early analysis of the largest Field-Style Experiments to date (the 1,000 Scientists AI JAM). These complementary methods enable a comprehensive analysis of LLM strengths and weaknesses with respect to their scientific knowledge, reasoning abilities, and adaptability. Although developed within a subset of scientific domains, the methodology is designed to be generalizable to a wide range of scientific domains.

Closing Panel: The Future Of Science And Society Entering The Era Of Artificial Super Intelligence

Moderator: Charlie Catlett, *Trillion Parameter Consortium*

Today's frontier models, with ASI-class systems on the horizon, can deliver the greatest scientific returns for humanity. This panel of community leaders will summarize a series of discussions in order to stimulate the community to identify and examine areas of science that can be meaningfully accelerated, ultimately identifying a few grand challenges whose solutions multiply benefits across society. Examples might include decarbonizing the global energy system, solutions to inequality and scarcity, overcoming threats to health that cause enormous human suffering, and removing the threat of nuclear weapons. Such a dialog is needed to distill a prioritized research portfolio in terms of challenges that commercial and public sector (academia, national laboratories) AI and science communities can pursue immediately in order to turn AI's accelerating capabilities into tangible, equitable gains.

Panelists: Ian Foster, *Argonne National Laboratory*
Karthik Duraisamy, *University of Michigan*
Satoshi Matsuoka, *RIKEN R-CCS*
Thierry Pellegrino, *AWS*



Scaling Intelligence
Breaking Barriers
Building Futures

Hackathon

HACKATHON / TUTORIAL OPENING PLENARY TALK

Advancing Science and Medicine with AI Physician-scientists

Vivek Natarajan, *Google DeepMind*

This talk highlights general purpose AI systems designed at Google to democratize medical expertise and accelerate scientific discovery. We will first take a look at the AI co-scientist, built to accelerate scientific breakthroughs by assisting scientists in generating novel hypotheses and aiding experimental design. This system has yielded validated results in areas like genetic discovery, drug repurposing, target discovery, and understanding antimicrobial resistance. Secondly, we will examine the AI co-physician, AMIE, developed to make medical expertise universally accessible through capabilities such as advanced diagnostic dialogue. In simulations, AMIE outperformed primary care physicians on multiple clinical evaluation axes and showed promise as an assistive tool, with ongoing real-world validations. Together, these AI initiatives demonstrate the potential to transform scientific research and care delivery.

BUILDING AGENTIC SYSTEMS FOR SCIENCE

Monday and Tuesday Morning in the Cascade room

This is a hands-on tutorial and hackathon for academic scientists with limited experience in agentic AI systems. Over 1.5 days, participants will learn to build and extend AI agents tailored to scientific challenges, particularly in biology and chemistry.

With guidance from mentors and access to NVIDIA and Cerebras compute resources, teams will collaborate on projects such as molecular tool development, protein engineering, and reasoning agents. This open-format event emphasizes collaborative learning and practical implementation, building on foundational AI concepts from a shared tutorial session. It is ideal for researchers eager to explore agentic systems and apply them to their own scientific work.

Learning Takeaways

Participants will gain a working knowledge of agentic system architecture, learn how to apply agentic methods to domain-specific scientific problems, and develop prototype tools or agents. They'll collaborate with peers and mentors, access advanced compute resources, and leave with hands-on experience that empowers further exploration of AI in science.

Session One

Shared foundations in AI for science

Session Two

Intro to agentic systems and use cases

Session Three

Team formation and project kickoff

Session Four

Hands-on hacking with expert mentorship

Session Five

Midpoint sync, debugging, optional breakouts

Session Six

Final demos, project showcases, wrap-up discussion

The logo for USAI (United States Artificial Intelligence) is displayed. It features a stylized grid of dots above the text "USAI" in a large, bold, sans-serif font. Below "USAI" is the phrase "Built by intel" in a smaller, lowercase sans-serif font. The entire logo is set against a dark blue background with a subtle pattern of glowing circuitry or data points.

Trusted AI from a trusted partner.

With five decades of US-based manufacturing and ready-to-deploy AI solutions, Intel is a proven partner to help the public sector implement AI securely and at scale.

That's the power of Intel Inside.®

Visit us at TPC25.

Intel.com/USAI

Tutorials

HACKATHON / TUTORIAL OPENING PLENARY TALK

Advancing Science and Medicine with AI Physician-scientists

Vivek Natarajan, *Google DeepMind*

This talk highlights general purpose AI systems designed at Google to democratize medical expertise and accelerate scientific discovery. We will first take a look at the AI co-scientist, built to accelerate scientific breakthroughs by assisting scientists in generating novel hypotheses and aiding experimental design. This system has yielded validated results in areas like genetic discovery, drug repurposing, target discovery, and understanding antimicrobial resistance. Secondly, we will examine the AI co-physician, AMIE, developed to make medical expertise universally accessible through capabilities such as advanced diagnostic dialogue. In simulations, AMIE outperformed primary care physicians on multiple clinical evaluation axes and showed promise as an assistive tool, with ongoing real-world validations. Together, these AI initiatives demonstrate the potential to transform scientific research and care delivery.

Tutorial Plenary: Introduction to AI for Science

AI for Science
(case studies)

Model Skills Evaluation

Lunch

AI for Science
(frameworks/tools)

Model Skills Evaluation

AI for Science
(frameworks/tools)

Model Skills Evaluation

Tuesday

AI for Science
(advanced topics)

AI for Scientific Productivity

AI for Science
(advanced topics)

AI for Scientific Productivity

AI FOR SCIENCE: FOUNDATIONS AND FRONTIERS

Monday and Tuesday Morning in the Siskiyou room

This is a hands-on tutorial designed to equip researchers with practical skills and conceptual grounding in the application of large-scale AI models to scientific challenges. The program covers key components of the AI model lifecycle — from distributed strategies for pre-training generative models to fine-tuning techniques for domain-specific tasks using models like LLAMA-70B and Stable Diffusion.

Participants will also learn to analyze and optimize performance through workload profiling with PARAVR, and to build intelligent scientific workflows using Retrieval-Augmented Generation (RAG) and agent-based approaches. The tutorial concludes with real-world case studies across disciplines — biology, climate, physics, chemistry — highlighting lessons learned from deployment and emerging trends such as simulation models and neural-symbolic systems.

Learning Takeaways

Participants will develop a practical understanding of large-scale AI model development, including:

- Parallelized pre-training strategies and fine-tuning techniques for domain-specific tasks
- How to analyze and optimize AI workloads using profiling tools, and gain
- Hands-on experience building Retrieval-Augmented Generation (RAG) pipelines
- Agent-based workflows

Attendees will come away with exposure to real-world scientific applications and current research frontiers in AI for science. This tutorial will be conducted by Neeraj Kumar (PNNL), Samantika Sury (HPE), Laura Morselli (CINECA), and Prasanna Balaprakash (ORNL).

Session One

Plenary session with all Tutorial and Hackathon participants: Foundations in AI for Science

Session Two

Case Studies and Emerging Frontiers in AI for Science

Session Three

Parallelization Strategies for Large-Scale Pre-Training

Session Four

Fine-Tuning Techniques: From Theory to Practice

Session Five

Profiling AI Workloads with PARAVR

Session Six

Building RAG-based Workflows OR AI Agents



EVALUATION OF AI MODEL SCIENTIFIC REASONING SKILLS

Monday in the Donner room

This is a hands-on tutorial designed to equip researchers with practical skills and conceptual grounding in the application of LLMs to scientific challenges. Large Language Models (LLMs) are becoming capable of solving complex problems while presenting the opportunity to leverage them for scientific applications. However, even the most sophisticated models can struggle with simple reasoning tasks and make mistakes.

This tutorial focuses on best practices for evaluating LLMs for science applications. It guides participants through methods and techniques for testing LLMs at basic and intermediate levels. It starts with the fundamentals of LLM design, development, application, and evaluation while focusing on scientific application. Participants will also learn various complementary methods to rigorously evaluate LLM responses in benchmarks and end-to-end scenario settings. The tutorial features a hands-on session where participants use LLMs to solve provided problems.

Learning Takeaways

Participants will learn the principles and approaches for the use of LLMs as scientific assistants and how these can be evaluated with respect to scientific knowledge and reasoning skills, such as:

- Use cases of LLMs for scientific applications
- Importance of prompting and performance
- Basic of LLM evaluation
- Evaluation of LLMs for science and engineering
- Advanced evaluation techniques of LLMs for Science and Engineering
- Hands-on

This tutorial will be conducted by leaders from the TPC EVAL working group: Franck Cappello, Neil Getty, and Sandeep Madireddy from Argonne National Laboratory, along with Javier Aula-Blasco from Barcelona Supercomputing Center.

Session One

Plenary session with all Tutorial and Hackathon participants: Foundations in LLMs for Science

Session Two

Use cases and basic evaluation techniques

Session Three

Advanced evaluation techniques

Session Four

Hands on

USING AI TO ACCELERATE DAY-TO-DAY SCIENTIFIC PRODUCTIVITY

Tuesday Morning in the Donner room

This is a hands-on tutorial designed to equip researchers, students, engineers, and scientific leaders with practical skills for the use of AI models and systems to accelerate planning, inquiry, coding, and other common tasks.

This tutorial will demonstrate how computational scientists can effectively harness AI-powered tools across the research lifecycle to increase their productivity. Attendees will learn to generate novel research ideas and hypotheses using agents for Deep Research and Idea Generation. We then cover structuring comprehensive research plans with AI assistance.

For implementation, attendees will see how to efficiently port, develop, and optimize code using tools like Google's Gemini Code Assist and CLI, alongside advanced optimizers such as AlphaEvolve. While this tutorial will use Google technologies for the examples (and attendees will be given accounts to access them), the core principles and strategies are designed to be portable, enabling scientists to effectively use any comparable AI tool in their own scientific endeavors.

Session One

LLM Refresher + Deep Research & Idea Generation

Session Two

Coding Faster & Better (Usually) + AI-enabled Science Applications

Learning Takeaways

Participants will learn how AI-powered tools can help in every phase of the computational research/application development process:

- Prompt engineering: how to get the best results, and why some methods work better than others
- Leveraging context windows to improve accuracy and usefulness of answers
- Providing sources/content to further improve accuracy and usefulness of responses

This tutorial will be conducted by Jay Boisseau, Advanced Computing Strategist at Google.

Breakout Group Sessions

TPC25 breakout groups are designed to identify, form, and pursue collaborations that will accelerate the development of new AI capabilities and services for scientific discovery. Some sessions are organized by TPC working groups, others are prospective working groups or birds-of-a-feather gatherings. Each session comprises a small set of lightning talks followed by group discussion.

The six-way parallel breakout schedule is loosely organized around six themes: Workflows, Initiatives, Life Sciences, Evaluation, Scale and Services, and Applications.

WORKFLOWS

DATA WORKFLOWS, AGENTS, AND REASONING FRAMEWORKS (DWARF)

KEYNOTE AND SYSTEMS SOFTWARE FOR AGENTS

Wednesday, July 30, 16:00

SCALABLE SCIENTIFIC DATA/SCIENTIFIC DATA FOR AI

Thursday, July 31, 8:30

SCALABLE PROCESSING PIPELINES

Thursday, July 31, 11:00

Organizers: Ian Foster, *Argonne National Laboratory and University of Chicago*
Neeraj Kumar, *Pacific Northwest National Laboratory*
Robert Underwood, *Argonne National Laboratory*
Ravi Madduri, *Argonne National Laboratory*

This multi-session track explores emerging systems and strategies for building intelligent, scalable platforms to accelerate scientific discovery. Talks and discussions will cover the design of agent-based architectures, integration of scientific workflows with large language models, scalable data pipelines, and novel reasoning frameworks. Emphasis will be placed on domain-specific applications spanning biology, climate, and materials, highlighting new approaches to real-time discovery, autonomous labs, and LLM-driven scientific tools. Participants will engage in dialogue on the future of scientific AI infrastructure and the coordination required to realize a distributed, agent-enabled discovery ecosystem.

See insert for associated lightning talks.

BOF: LLMS FOR LIVING DOCS

Thursday, July 31, 14:00

Organizer: Daniel Ratner, *SLAC*

Large-scale scientific facilities generate extensive “living documents”: logbooks, wikis, hardware/software standards, etc. Large language models are promising tools for both extracting information from and maintaining these documents, with example applications including improving search, educating new employees, and generating shift summaries. However, the wealth of highly specialized terminology as well as multi-modal information complicates the direct application of existing LLMs. This session will discuss opportunities, challenges, and existing solutions, and search for ways to collaborate across a range of DOE facilities.

See insert for associated lightning talks.



INITIATIVES

BOF: BUILDING FOUNDATION MODELS FOR THE ELECTRIC GRID (GRIDFM)

Wednesday, July 30, 14:00

Organizers: Kibaek Kim, *Argonne National Laboratory*
Hendrik Hamann, *Stony Brook University and Brookhaven National Laboratory*
Hongwei Jin, *Argonne National Laboratory*

GridFM (Grid Foundation Model) is a fast-growing, community-driven initiative uniting researchers from national labs, academia, and industry to develop foundation models tailored to the electric grid. Unlike typical foundation models, GridFM focuses on graph-based models pretrained on multi-modal grid data to capture the system's complexity and dynamics. This session will introduce GridFM to the TPC community, present the motivation for domain-specific foundation models, and feature invited talks highlighting early technical progress and real-world applications. An open forum will follow to discuss shared challenges in data, modeling, software, and infrastructure, and explore connections with HPC, scientific ML, and energy systems modeling. Researchers in AI/ML, scalable algorithms, and complex systems are invited to join and help shape the future of GridFM..

See insert for associated lightning talks.

BOF: LEVERAGING ICICLE FOR TPC APPLICATIONS ACROSS THE COMPUTING CONTINUUM

Wednesday, July 30, 16:00

Organizers: Raghu Machiraju, *The Ohio State University*
DK Panda, *The Ohio State University*
Zhao Zhang, *Rutgers University*

This session will explore the use of AI models across the computing continuum, highlighting diverse application domains and key technical challenges. Participants will discuss current obstacles in data collection, model training, and deployment workflows, particularly in distributed environments. The session will also focus on how the NSF-funded ICICLE AI Institute components can be rapidly adapted to meet TPC community requirements, including domain-specific data curation and labeling, scalable and distributed model training, and robust evaluation for bias and alignment with scientific goals. In addition, we will examine the role of agentic workflows in scientific computing, identifying reusable patterns that can be instantiated within ICICLE and deployed effectively on HPC systems.

BOF: AI IN DECISION SCIENCES

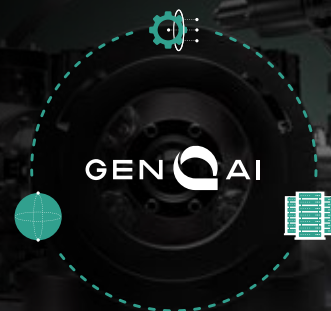
Thursday, July 31, 8:30

Organizers: Francis Alexander, *Argonne National Laboratory*
Peter Nugent, *Lawrence Berkeley National Laboratory*
Suzanne Pierce, *U. Texas (Austin) / TACC*

This Birds of a Feather session explores how AI models can transform high-consequence decision-making under uncertainty. Building on the consortium's collaborative development of foundation models for science and engineering, we'll examine applications in disaster response, supply chain resilience, pandemic management, and other areas. The discussion will connect TPC's work on scalable architectures, scientific data curation, and exascale optimization to breakthrough capabilities in time-critical decision support. Participants



Enhancing Your
LLMs Today with
**Quantum +
HPC + AI**



GenQAI is a breakthrough framework that harnesses unique quantum-generated data to tackle complex problems intractable for classical computing.

Today, we operate **Quantinuum H2**, the world's most powerful quantum computer, as demonstrated by a record-breaking **Quantum Volume of 2^{23} (8,388,608)** and the **first ever universally fully fault-tolerant gate set** with repeated error correction. Launching in 2025, **Quantinuum Helios** will succeed H2 as the world's most powerful quantum computer. Helios will be the **first commercial system to offer 50 logical qubits** and the **highest two-qubit gate fidelity at 99.95%**.

Learn how we're unlocking AI's full potential with quantum computing



INITIATIVES, continued

from industry, academia, and national laboratories will explore how TPC's multi-institutional approach can accelerate next-generation decision support systems while maintaining the rigor essential for high-stakes applications.

See insert for associated lightning talks.

BOF: PUBLIC AI: POLICY, COMMUNITY, AND THE FUTURE OF NATIONAL LABS

Thursday, July 31, 11:00

Organizers: Joshua Tan, *Metagov*
Nick Vincent, *Simon Fraser University*
Avani Wildani, *Cloudflare*

Governments around the world are beginning to invest in public AI — AI built and maintained as public infrastructure. New initiatives like AuroraGPT (U.S.), SEA-LION (Singapore), and the EU's AI Gigafactories signal growing recognition that public institutions have a critical role to play in the next generation of AI systems. Meanwhile, national labs are being asked to take on more: building sovereign capabilities, supporting regulated sectors, and producing high-trust, mission-aligned models. This breakout introduces the emerging public AI movement and key proposals such as "Airbus for AI" and "CERN for AI." It aims to provide national lab leaders with a strategic view of the policy landscape — beyond narrow regulation — and to catalyze coordination across jurisdictions. Blending technical and institutional design challenges, this session may seed a working group focused on aligning national labs with the new public AI infrastructure agenda.

BOF: ENERGY EFFICIENT HPC FOR AI WORKLOADS

Thursday, July 31, 14:00

Organizers: Natalie Bates, *Lawrence Berkeley National Laboratory*
Siddhartha Jana, *Intel*

As AI workloads and HPC systems scale to new extremes, energy efficiency has become a defining challenge for sustainable computing. This BOF, led by the Energy Efficiency HPC Working Group (EE HPC WG), brings together experts from computing centers, industry, and research institutions to explore recent advances in sustainable infrastructure, including liquid cooling innovations, facility energy reuse, alternative power sources, and software-driven energy optimization. Participants will discuss trends in AI/HPC operations, insights from systems like Fugaku, and strategies for integrating sustainability into procurement and operational policies. The session invites open dialogue on actionable approaches for reducing energy and environmental impacts across the global HPC and AI ecosystem.

See insert for associated lightning talks.

LIFE SCIENCES

AI FOR CANCER

Wednesday, July 30, 14:00

AGENTIC AI AND FOUNDATION MODELS

Wednesday, July 30, 16:00

AI FOR BIOLOGY

Thursday, July 31, 8:30

Organizers: Arvind Ramanathan, *Argonne National Laboratory*
Tom Brettin, *Argonne National Laboratory*
Miguel Vazquez, *Barcelona Supercomputing Center*
Silvia Crivelli, *Lawrence Berkeley National Laboratory*
Heidi Hanson, *Oak Ridge National Laboratory*

This track will focus on the development of foundation models / large language models and agentic systems for biology. Given the shared interests and the broader implications for how AI models and agentic systems can potentially alter the scope of biological research, the goal of this session is to catalyze discussions and build collaborations in areas such as: (1)



LIFE SCIENCES, continued

how to build shared datasets for creating a rich repertoire of downstream evaluation tasks for foundation models; (2) discuss and develop shared strategies for model sharing and scoping across diverse biological applications; (3) evaluate approaches towards incorporating robust strategies to reflect implicitly on the bias and trust/safety into the context of biological data; and (4) how to develop agentic systems for a variety of tasks ranging from discovery to laboratory operations. The track will include multiple topic areas in each 90-minute breakout, including Agentic systems, AI for Cancer, and emerging opportunities at the intersection of HPC, AI, biomedicine and precision population health. Agentic AI systems in particular offer a novel paradigm for simulating complex disease trajectories and personalizing interventions at scale. Balancing predictive accuracy with explainability poses a core challenge. Advancing this field requires coordinated innovation in AI algorithms, data infrastructure, and HPC workflows for biological and bio-medical research and healthcare.

See insert for associated lightning talks.

HPC-AI SOCIETY MEETING FROM THE TPC25 CONFERENCE

Thursday, July 31, 11:00

Organizer: Doug Norton, *The HPC-AI Society*

Hear from TPC Executive Director Charlie Catlett about TPC's work in advancing AI for science and engineering, and discuss how the two non-profit, vendor neutral organizations, TPC and The HPC-AI Society, are holistically collaborating to advance common goals through their respective communities.

BOF: FEDERATED LEARNING AT SCALE

Thursday, July 31, 14:00

Organizers: Ravi Madduri, *Argonne National Laboratory*
Jason Haga, *Advanced Industrial Science and Technology (AIST)*
Feiyi Wang / Sahil Tyagi, *Oak Ridge National Laboratory*
Miguel Vazquez, *Barcelona Super Computing Center*

In this BOF, we will explore the application of Federated Learning in building foundational AI models for science, as well as the technical policy challenges in training foundational models across the geographical boundaries. We will specifically focus on identifying computing, networking, and other challenges in training these models across different scientific domains. We will focus on identifying key bottlenecks in compute, networking, and coordination, and what it would take to overcome them. We'll also look at how these challenges play out across different scientific domains, and what's needed to make federated model training a practical reality for the broader research community.

See insert for associated lightning talks.

EVALUATION

MODEL SKILLS, REASONING, AND TRUST EVALUATION (EVAL)

INTRO AND BENCHMARKS

Wednesday, July 30, 14:00

UQ AND SAFETY

Wednesday, July 30, 16:00

AUTOMATIC BENCHMARK GENERATION

Thursday, July 31, 8:30

ADVANCE EVALUATION

Thursday, July 31, 11:00

EVALUATION, continued

Organizers: Franck Cappello, *Argonne National Laboratory*
Sandeep Madireddy, *Argonne National Laboratory*
Javier Aula-Blasco, *Barcelona Supercomputing Center*

One of the main thrusts behind the rapid evolution of LLMs is the availability of benchmarks that assess the skills and trustworthiness of LLMs. Not only do they enable a rigorous evaluation of LLMs skills and trustworthiness from accepted metrics, but they also generate competition between LLM developers. While several frameworks/benchmarks have emerged as de facto standards for the evaluation of general-purpose LLMs (Eleuther AI Harness and HELM for skills, DecodingTrust for trustworthiness), only very few of them specifically are related to science. In this segment, we will discuss the challenges of developing methods to evaluate the skills, trustworthiness, and safety of large Foundation Models for science. This track will include multiple sessions focused on different facets of model evaluation.

See insert for associated lightning talks.

BOF: TRUSTWORTHINESS IN SCIENTIFIC MACHINE LEARNING: FROM INFERENCE-TIME-COMPUTE TO REASONING

Thursday, July 31, 14:00

Organizers: Diane Oyen, *Los Alamos National Laboratory*
Bradley Love, *Los Alamos National Laboratory*
Ayan Biswas, *Los Alamos National Laboratory*

For SciML models to be trustworthy and broadly deployable, we must balance accuracy, complexity, and computational cost. Like LLMs, SciML faces “data walls” where scale alone yields diminishing returns — prompting growing interest in models capable of genuine reasoning. But what counts as reasoning in SciML? Unlike LLMs, which reward-learn response patterns, scientific reasoning demands adherence to physical laws, logical transparency, and uncertainty quantification. This includes symbolic derivation, solver integration, and interpretable chains of thought. Yet, embedding constraints can reduce flexibility, while data-driven models risk losing interpretability. Navigating this space requires community-developed evaluation frameworks — and clarity on what “reasoning” means across disciplines — to distinguish brute-force prediction from scientific understanding and advance robust, insightful AI for science.

See insert for associated lightning talks.

SCALE AND SERVICES

BOF: DEPLOYMENT OF INFERENCE-FOR-SCIENCE SERVICES AT HPC CENTERS

SESSION 1

Wednesday, July 30, 14:00

SESSION 2

Wednesday, July 30, 16:00

Organizers: Weicheng Huang, *National Center for High Performance Computing*
Venkat Vishwanath, *Argonne Leadership Computing Facility*
Aleksi Kallio, *IT Center for Science (CSC)*
Dan Stanzione, *Texas Advanced Computing Center*

As foundation models and domain-specific AI systems gain traction in science, HPC centers are actively developing infrastructure to support scalable, high-performance inference services. This session convenes leaders from labs, vendors, and the open-source community to share experiences, challenges, and emerging best practices in deploying inference for science. Topics include integration of inference with simulations and workflows, support for diverse architectures, sustainability of open-weight models, supporting proprietary models, and software stacks tuned for reliability and reproducibility. Training and workforce development will also be addressed, reflecting the growing institutional investment in upskilling staff and users. The session will gather use cases, propose next steps — including a community webpage and Slack/Discord channel — and catalyze collaboration across the TPC community.

See insert for associated lightning talks.



SCALE AND SERVICES, continued

MODEL ARCHITECTURE AND PERFORMANCE EVALUATION (MAPE)

SESSION 1

Thursday, July 31, 8:30

SESSION 2

Thursday, July 31, 11:00

SESSION 3

Thursday, July 31, 14:00

Organizers: Rio Yokota, *Institute of Science Tokyo*
Murali Emani, *Argonne National Laboratory*

Architectures for AI models are evolving rapidly, with frequent innovations in transformer variants, mixture-of-experts extensions, and state-space models. Frameworks like Megatron-LM, DeepSpeed, and their forks support different architectures, parallelism strategies, and system optimizations. Identifying the optimal architecture and framework for training trillion-parameter models on scientific data is vital to unlocking the next generation of AI for science. Equally crucial is efficient inference, which enables the practical use of pre-trained models in downstream scientific applications. This multi-session track will bring together researchers and practitioners to discuss cutting-edge strategies for large-scale training, inference, and test-time scaling, alongside robust methods.

See insert for associated lightning talks.

APPLICATIONS

AI MODELS FOR SOFTWARE ENGINEERING AND DEVELOPMENT

Wednesday, July 30, 14:00

Organizers: Anshu Dubey, *Argonne National Laboratory*
Valerie Taylor, *Argonne National Laboratory*
Pete Beckman, *Northwestern University*

Generative AI has rapidly evolved, now demonstrating strong performance in scientific code generation, refactoring, and formal verification. This session highlights recent breakthroughs, including agentic systems that plan and refine scientific workflows, execution-guided code generation, domain-specific LLMs for HPC and HEP applications, and tools for energy-aware refactoring and formal specification synthesis. Talks will explore datasets, multi-agent pipelines, and inference-time feedback loops that dramatically improve correctness and usability. We will discuss how these methods are accelerating software development across science domains, marking a shift from early experimentation to practical, performant systems that redefine how we build and verify scientific code.

See insert for associated lightning talks.

BOF: FOUNDATION MODELS FOR FUSION ENERGY

Wednesday, July 30, 16:00

Organizers: William Tang, *Princeton Particle Physics Laboratory*
Shantenu Ja, *Rutgers University and Princeton Particle Physics Laboratory*

The aim of Fusion Foundation Models (FFMs) is to transform fusion-energy research and accelerate commercialization this decade. Embedding fundamental physics directly within associated architectures enables overcoming scarce experimental data and extrapolating while adhering to validation, verification, and uncertainty quantification principles. These models will deliver predictive accuracy, operational intelligence, and design optimization across the fusion lifecycle. Priorities include delivering physics-informed, multimodal architectures, enabling human-AI-enabled co-discovery, and orchestrating specialized AI agents that ensure scalability, robustness, real-time response, and strict ethical and safety compliance. This BoF surveys the current landscape of FFM development and deployment, probing applications for real-time control and

APPLICATIONS, continued

prediction in complex scientific settings. Participants will map actionable, collaborative opportunities, positioning fusion as a compelling proving ground and inviting active engagement in the science community.

See insert for associated lightning talks.

AI FOR SCIENTIFIC DISCOVERY IN MATERIALS SCIENCE (AI4MS)

Thursday, July 31, 8:30

Organizers: Eliu Huerta, *Argonne National Laboratory and University of Chicago*
Samuel Blau, *Lawrence Berkeley National Laboratory*

This session explores the transformative role of generative AI and autonomous agents in accelerating discovery in materials science. As this field increasingly adopts data-centric approaches, the session will focus on four key themes: (1) the generation and curation of high-quality, interoperable datasets spanning simulations, synthesis protocols, and experimental measurements; (2) the development of foundation models and domain-specific AI systems, including multimodal LLMs tailored to scientific data; (3) the integration of AI agents with robotic platforms and autonomous labs, enabling self-driving experimentation, real-time decision-making, and natural language control of complex workflows; and (4) the application of these technologies in real-world discovery.

See insert for associated lightning talks.

AI EDUCATION AND OUTREACH

Thursday, July 31, 11:00

Organizers: Valerie Taylor, *Argonne National Laboratory*
Jason Haga, *Advanced Industrial Science and Technology (AIST)*
Javier Aula-Blasco, *Barcelona Supercomputing Center*
Claudio Domenico Arlandini, *CINECA*

Over the past year, the accelerating integration of AI into scientific research has intensified demand for a skilled, AI-fluent scientific workforce. Institutions worldwide have launched new training programs, curricula, and upskilling initiatives to prepare both early-career researchers and established staff to effectively harness AI for discovery. These efforts are producing valuable insights into what pedagogical strategies succeed — and where gaps remain. As scientific domains rapidly adopt LLMs and foundation models, building a diverse, globally connected, and continuously learning workforce is both a strategic imperative and a social responsibility. This session will explore how the Trillion Parameter Consortium can support inclusive, cross-disciplinary workforce development through governance, shared curricula, and collaborative training infrastructures.

See insert for associated lightning talks.

EARTH AND ENVIRONMENT (AI FOR DIGITAL EARTH)

Thursday, July 31, 14:00

Organizer: Po-Lun Ma, *Pacific Northwest National Laboratory*

Foundation models are reshaping how we simulate, understand, and interact with the Earth system. This session explores how large-scale AI models accelerate Earth system prediction, transform data assimilation and coupling, and enable new forms of discovery and decision support. We invite contributions on digital Earth development, analysis, and prediction, and LLM-based tools for querying, debugging, or orchestrating digital twins. Case studies, open-source tools, and critical perspectives on trustworthiness, reproducibility, and interpretability are welcome. We will discuss the challenges and opportunities at the intersection of AI and Earth science, and their role in shaping next-generation digital Earth infrastructure and Earth system sciences.

See insert for associated lightning talks.



Plenary Speakers and Panel Moderators

Ricardo Baeza-Yates

Director, AI Institute,
Barcelona Supercomputing Center

Ricardo Baeza-Yates is the inaugural director of the AI Institute at the Barcelona Supercomputing Center since May 2025. Earlier roles include Director of Research at the Inst. for Experiential AI of Northeastern University (2021-2025) and VP of Research of Yahoo Labs (2006-2016). He has a Ph.D. in CS from the Univ. of Waterloo and is co-author of the Modern Information Retrieval textbook, that won the ASIST 2012 Book of the Year award. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow.



Prasanna Balaprakash

Director of AI Programs,
Oak Ridge National Laboratory

Prasanna Balaprakash is the Director of AI Programs and a Distinguished R&D Scientist in the Computing and Computational Sciences Directorate at Oak Ridge National Laboratory (ORNL), where he co-leads the AI Initiative — an LDRD portfolio focused on secure, trustworthy, and efficient AI for scientific discovery, experimental facilities, and national security. He serves as the AI lead for several U.S. Department of Energy-funded projects and received the DOE Early Career Award in 2018.



Franck Capello

R&D Lead, Senior Computer Scientist,
Argonne National Laboratory

Frank Cappello is an R&D lead and Senior Computer Scientist at Argonne National Laboratory. He leads a research team exploring resilience for HPC and large-scale distributed systems, lossy compression of scientific data and LLMs for science. He is an IEEE Fellow, the recipient of the 2024 IEEE CS Charles Babbage Award, the 2024 Europar Achievement Award, the 2022 HPDC Achievement Award, two R&D100 awards (2019 and 2021), the 2018 IEEE TCPP Outstanding Service Award, and the 2021 IEEE Transactions of Computer Award for Editorial Service and Excellence.



Charlie Catlett

Executive Director, *Trillion Parameter Consortium* |
Senior Computer Scientist, *Argonne National Laboratory* | *The University of Chicago*

Charlie Catlett is a Senior Computer Scientist at the U.S. Department of Energy's Argonne National Laboratory, and a Visiting Scientist at the University of Chicago's Mansueto Institute for Urban Innovation. His research focuses on building cyberinfrastructure to embed edge-AI in urban, environmental, and emergency sensing and response settings. He was founding chair of Grid Forum / Global Grid Forum from 1999-2005 and director of NSF's TeraGrid initiative from 2004-2007. Charlie was part of the team that established the National Center for Supercomputing Applications (NCSA) in 1985, leading efforts there including the deployment and operation of the NSFNET backbone network,



an early component of the Internet, and serving as Chief Technology Officer prior to joining Argonne and UChicago in 2000. He was one of GovTech magazine's "25 Doers, Dreamers & Drivers" of 2016 and in 2019 received the Argonne Board of Governors Distinguished Performer award. Charlie is a Computer Engineering graduate of the University of Illinois at Urbana-Champaign.

Preeth Chengappa

Head of Industry, Semiconductors & Physics,
Microsoft

Preeth Chengappa leads GenAI-focused industry engagements across Physics & Semiconductors at Microsoft for the newly launched Discovery platform, which brings enterprise-grade agentic capabilities to science and engineering. Preeth has deep industry expertise in the semiconductor industry, working across commercial and government entities, and represents Microsoft on the US CHIPS Act — EU and other nations' equivalents for semiconductors, secure cloud, and GenAI. In 2011, Preeth co-founded SiCAD, a startup that built the first cloud-based chip design platform leveraging industry standard tools. He has held engineering, product, and business development leadership roles in large enterprises and startups. Preeth graduated from NITK, India, with a degree in Chemical Engineering.



Steve Clark

Head of Artificial Intelligence, *Quantinuum*

Stephen Clark is Head of Artificial Intelligence at Quantinuum. Prior to this, he spent 14 years as a Member of Faculty at the Departments of Computer Science of the Universities of Oxford and Cambridge, including four years as a Tutorial Fellow at Keble College, Oxford. From 2016 to 2020, Steve was a Research Scientist at DeepMind in London, acting as a Team Lead for the research area of Grounded Language Learning. Steve holds an undergraduate degree in Philosophy from the University of Cambridge (Gonville and Caius College). He obtained an M.Sc. in Cognitive Science at the University of Manchester (Department of Computer Science) and a Ph.D. in Computer Science and Artificial Intelligence at the University of Sussex (School of Cognitive and Computing Sciences). Steve is currently an Honorary Professor at Queen Mary University of London.



Jens Domke

Team Principal of the Supercomputing Performance Research Team, *RIKEN R-CCS*

Jens Domke is the Team Principal of the Supercomputing Performance Research Team at RIKEN R-CCS. He received his doctoral degree in 2017 from TU Dresden for his work on HPC routing algorithms and interconnects. Jens started his career in HPC in 2008, after he and a team of five students from TU Dresden and Indiana University won the Student Cluster Competition at SC08. Since then, Jens has published dozens of peer-reviewed articles, contributed routing algorithms to InfiniBand's subnet manager, and built the first large-scale HyperX prototype at TokyoTech. Jens' research interests include system co-design, performance evaluation,



extrapolation and modeling, interconnect networks, AI, and optimization of parallel applications and architectures.

Pradeep Dubey

Intel Senior Fellow and Parallel Computing Lab Director, *Intel Labs*

Pradeep K. Dubey is an Intel Senior Fellow and director of the Parallel Computing Lab, a part of the Intel Labs organization at Intel Corporation. Since 2003, he has led a team of top researchers focused on state-of-the-art research in parallel computing. Pradeep and his team are responsible for defining computer architectures that can efficiently handle emerging machine learning/artificial intelligence, traditional HPC applications for data-centric computing environments, and deriving product differentiation opportunities for Intel's CPU and GPU processing platforms. Pradeep holds 36 patents and has published more than 100 peer-reviewed technical papers. Throughout his career, he has made significant contributions to the design, architecture and application performance of various microprocessors, including the IBM Power PC, the Intel386™, Intel486™, Intel® Pentium®, and Intel Xeon® processors. Pradeep earned a bachelor's degree in electronics and communication engineering from Birla Institute of Technology, India; a master's degree in electrical engineering from the University of Massachusetts at Amherst; and a Ph.D. in electrical engineering from Purdue University. He was named a Fellow of the Institute of Electrical and Electronics Engineers in 2001 for his contributions to computer architecture supporting multimedia processing.



Karthik Duraisamy

Professor of Aerospace Engineering, *University of Michigan*

Karthik Duraisamy is a Professor of Aerospace Engineering at the University of Michigan (U-M) where he also directs the Michigan Institute for Computational Discovery and Engineering (MICDE). He holds a PhD in Aerospace Engineering and a Masters in Applied Mathematics from the University of Maryland. His research interests span a broad spectrum of computational science and AI, including data-driven and reduced order modeling, statistical inference, numerical methods, and Generative AI for science. Karthik is the PI of the U-M/Los Alamos Center on Advanced Computational Sciences. He is also the founder and chief scientist of the Silicon Valley-based startup Geminus.AI, which is focused on physics-informed AI to accelerate autonomous industrial operations.



Hal Finkel

Director, Computational Science Research and Partnerships (CSRP) Division, *DOE Office of Science, Advanced Scientific Computing Research Program*

Hal Finkel is a program manager for computer science research in the US Department of Energy Office of Science's Advanced Scientific Computing Research (ASCR) program. Prior to joining ASCR, Hal was the Lead for Compiler Technology and Programming Languages at Argonne's Leadership Computing Facility. As part of DOE's Exascale Computing Project (ECP), Hal was a PathForward technical lead and PI/Co-PI of several multi-institution activities. Hal also helped develop the Hardware/Hybrid Accelerated Cosmology Code (HACC), a two-time IEEE/ACM Gordon Bell Prize finalist. He graduated from Yale University in 2011



with a Ph.D. in theoretical physics focusing on numerical simulation of early-universe cosmology.

Ian Foster

Data Science and Learning Division Director, *Argonne National Laboratory*

Dr. Ian Foster is Senior Scientist and Distinguished Fellow, and also director of the Data Science and Learning Division, at Argonne National Laboratory, and the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago. Ian received a BSc degree from the University of Canterbury, New Zealand, and a PhD from Imperial College, United Kingdom, both in computer science. His research deals with distributed, parallel, and data-intensive computing technologies, and innovative applications of those technologies to scientific problems in such domains as materials science, climate change, and biomedicine. Foster is a fellow of the AAAS, ACM, BCS, and IEEE, and an Office of Science Distinguished Scientists Fellow.



Raj Hazra

CEO, *Quantinuum*

Raj Hazra has more than three decades of experience in supercomputing, quantum, and technical roles across the globe. Prior to joining Quantinuum, he served as the General Manager, Compute and Networking Business Unit at Micron Technologies, and spent 25 years at Intel Corporation, leading the Enterprise and Government Group, Technical Computing Group, Supercomputer Architecture and Planning, and Systems Technology Research. Before joining Intel, Raj was with the Lockheed Corporation based at NASA's Langley Research Center. He prides himself on building high-performing teams with a growth mindset and a culture of truth and transparency. Raj has a Ph.D. and Master's degree in Computer Science from the College of William and Mary in Virginia, U.S., as well as a Bachelor's degree in Computer Science from Jadavpur University in Kolkata, India, and holds 16 patents.



Kexin Huang

PhD Student, *Stanford University*

Kexin Huang is a fourth-year PhD student in Computer Science at Stanford University, advised by Prof. Jure Leskovec. His research focuses on leveraging AI to drive novel, deployable, and interpretable biomedical discoveries, while also tackling fundamental AI challenges such as multi-modal modeling, uncertainty quantification, and agentic reasoning. Kexin's work has been published in Nature Medicine, Nature Biotechnology, Nature Chemical Biology, Nature Biomedical Engineering, Nature, and machine learning conferences including NeurIPS, ICML, and ICLR. He has received numerous best paper awards at NeurIPS/ICML workshops, ISMB, and ASHG, with cover articles in Nature Biotechnology and Cell Patterns. Kexin's research has been featured in major media outlets such as Forbes, WIRED, and MIT Technology Review. He has also contributed to machine learning research at leading companies and institutions, including Genentech, GSK, Pfizer, IQVIA, Flatiron Health, Dana-Farber Cancer Institute, and Rockefeller University.



Earl Joseph

CEO, *Hyperion Research*

Earl Joseph, Chief Executive Officer of Hyperion Research, drives research and consulting efforts associated with the United States, Europe and Asia-Pacific markets for technical computing. He advises Hyperion Research clients on the competitive, managerial, technological, integration and implementation issues for HPC and AI, and heads up Hyperion Research's high-end HPC user forum activities. Earl's areas of expertise include technical computers (from entry-level servers to high-end capability supercomputers), software, AI technologies, storage, and networking solutions. He has worked for four technical computing companies in multiple marketing and R&D roles. Earl holds a Ph.D. from the University of Minnesota, where his research focus was the strategic management of high technology firms, and an undergraduate degree in business and technology from the University of Minnesota.



Kyoung-Sook Kim

Deputy Director of Intelligent Platforms Research Institute, *Advanced Industrial Science and Technology (AIST)*

Kyoung-Sook Kim is a deputy director at the Intelligent Platforms Research Institute at the National Institute of Advanced Industrial Science and Technology (AIST) in Japan. Her research interests include spatiotemporal data platforms, big data analysis, data integration, etc. She served as a researcher at the National Institute of Information and Communications Technology in Japan from November 2007 to March 2014 and is currently working on a testbed for data and AI quality management. Kyoung-Sook also contributes to international standardization for AI data ecosystem interoperability, serving as the project leader of ISO/IEC 5259-2 and an expert in several working groups, including ISO/IEC JTC 1/SC 42/WG 2 (AI Data), IEC SyC Smart Cities/WG 3 (Reference Architecture), and ISO/TC 204/WG 3 (ITS Database Technology). Additionally, she co-chairs the Moving Features SWG, GeoAI DWG, and Urban Digital Twins DWG within the Open Geospatial Consortium (OGC).



Jiacheng Liu

PhD Student Researcher,
Allen Institute for AI | University of Washington

Jiacheng Liu is a researcher at the Allen Institute for AI (AI²) and a PhD student at University of Washington. His research area spans LLM pre-training, post-training, text generation, commonsense reasoning, and mathematical reasoning. Most recently, he focuses on developing efficient search engines for understanding the massive training data of LLMs and tracing LLM behaviors. Jiacheng's work has been generously supported by the Meta AI Mentorship Program and Qualcomm Innovation Fellowship.



Satoshi Matsuoka

Director, *RIKEN R-CCS*

Professor Satoshi Matsuoka has been the director of Riken Center for Computational Science (R-CCS), the Tier-1 national HPC center for Japan, since April 2018, developing and hosting Japan's flagship



'Fugaku' supercomputer. Fugaku was the fastest supercomputer in the world in 2020 and 2021, supporting cutting-edge HPC research, including investigating Post-Moore era computing and especially the future FugakuNEXT supercomputer. Satoshi led the TSUBAME series of supercomputers that received much international acclaim at the Tokyo Institute of Technology, where he holds a professor position pursuing research in HPC, scalable big data, and AI. His longtime contribution was commended with the Medal of Honor with Purple ribbon by his Majesty Emperor Naruhito of Japan in 2022. Satoshi is a Fellow in ACM, ISC, IPSJ, and the JSSST and has won numerous awards, including ACM Gordon Bell Prizes, the IEEE-CS Sidney Fernbach Award, and the IEEE-CS Computer Society Seymour Cray Computer Engineering Award.

Siddharth Narayanan

Technical Staff, *FutureHouse*

Siddharth Narayanan is a physicist who is deeply passionate about making science more efficient by building systems that scale human capability. During his PhD, he worked on jet substructure and dark matter searches at the Large Hadron Collider. He has since also conducted ML research at Fidelity Investments and Flagship Pioneering, focusing on representation learning for transcriptomics, protein design, and scientific reasoning using language models.



Vivek Natarajan

Research Scientist, *Google DeepMind*

Vivek Natarajan is a Research Scientist at Google DeepMind, leading research at the intersection of AI, science, and medicine. He is the lead researcher behind Med-PaLM (Nature, 2023) and Med-PaLM 2 (Nature Medicine, 2025), the first AI systems to obtain passing and expert-level scores on US Medical License exam questions, respectively. Vivek also co-leads Project AMIE, a research program aiming to build and democratize medical superintelligence. Over the past year, AMIE has shown promising potential in controlled settings, including primary care, specialty care, and complex diagnostic challenges, as both a standalone (Nature, 2025) and assistive (Nature 2025) tool for clinicians. Finally, Vivek recently co-led the development of the AI co-scientist — a virtual AI collaborator designed to augment scientists, help uncover new original knowledge, and accelerate the clock speed of scientific discoveries. Prior to Google, Vivek worked on multimodal assistant systems at Facebook AI Research. He is also part of the faculty for executive education at Harvard T.H. Chan School of Public Health in a part-time capacity.



Thierry Pellegrino

Global Head of Advanced Computing,
Amazon Web Services

Thierry Pellegrino is the Global head of Advanced Computing at AWS, a role in which he oversees HPC, domain-specific ML, IOT, and Quantum for the company. In his last industry role, Thierry was CEO of Penguin Computing, and prior to that he spent 23 years with Dell, where he was the Head of the HPC and AI business. Thierry has held multiple leadership roles over his career, ranging from engineering to strategy, M&A, and business leadership, and has had the privilege to sit on the board of GRC's and Penn State's ICDS.



Molly Presley

SVP Global Marketing, *Hammerspace*

Molly Presley is the SVP of Global Marketing for Hammerspace and the host of the very popular Data Unchained podcast. She brings a wealth of experience from leading product and marketing organizations, user communities, and customer advisory boards for global technology innovators including DDN, Qumulo, and Quantum. Molly is the founder of the Active Archive Alliance, co-author of three books focused on putting data to use in research, analytics, and AI environment, and was a previous board member of the Storage Networking Industry Association (SNIA).



Arvind Ramanathan

Computational Science Leader,
Argonne National Laboratory

Arvind Ramanathan is a computational biologist in the Data Science and Learning Division at Argonne National Laboratory and a senior scientist at the University of Chicago Consortium for Advanced Science and Engineering (CASE). His research interests are at the intersection of data science, high performance computing and biological/biomedical sciences.



Flora Salim

Professor, *University of New South Wales*

Flora Salim a full Professor in the School of Computer Science and Engineering at the University of New South Wales (UNSW) Sydney, where she also serves as the Deputy Director (Engagement) of the UNSW AI Institute. Her work focuses on multimodal machine learning and foundation models for time-series and spatio-temporal data, multimodal sensors, and wearables, and on applications of AI and LLMs for smart and sustainable cities, and for mobility, transport, energy, and grid systems. She is a member of the Australian Academy of Sciences' National Committee for Information and Computing Sciences and an elect member of the Australian Research Council (ARC) College of Experts. She has received multiple nationally and internationally competitive fellowships, such as Humboldt Fellowship, Bayer Fellowship, and many accolades and awards such as the Women in AI Award Australia and New Zealand (2022) and IBM Smarter Planet Industry Innovation Award. She is a Vice Chair of the IEEE Task Force on AI for Time-Series and Spatio-Temporal Data and serves as an editorial board member of many journals including ACM TIST, ACM TSAS, IMWUT, IEEE Pervasive Computing, and Nature Scientific Data.



Addison Snell

Co-Founder & Chief Executive Officer,
Intersect360 Research

Addison Snell is a veteran of the HPC industry and the co-founder and CEO of Intersect360 Research, now in its 15th year delivering forecasts and insights for high-performance markets. Intersect360 Research is a premier source of market information, analysis, and consulting for



HPC and hyperscale industries worldwide. Addison is a frequent keynote speaker and panel moderator at industry events, has testified before the U.S.-China Economic & Security Review Commission Congressional Subcommittee, and was named one of 2010's "People to Watch" by *HPCwire*. Prior to Intersect360 Research, Addison was an HPC industry analyst for IDC. He originally gained industry recognition as a marketing leader and spokesperson for SGI's supercomputing products and strategy. Addison holds a master's degree from the Kellogg School of Management at Northwestern University and a bachelor's degree in Mathematics from the University of Pennsylvania.

Rick Stevens

Associate Laboratory Director - CELS, *Argonne National Laboratory* | Professor of Computer Science, *The University of Chicago*

Rick Stevens is a Professor of Computer Science at the University of Chicago and the Associate Laboratory Director of the Computing, Environment and Life Sciences (CELS) Directorate and Argonne Distinguished Fellow at Argonne National Laboratory. His research spans the computational and computer sciences from high-performance computing architecture to the development of advanced tools and methods. Recently, he has focused on developing AI methods for a variety of scientific and biomedical problems, and also has significant responsibility in delivering on the U.S. national initiative for Exascale computing and developing the DOE's Frontiers in Artificial Intelligence for Science, Security, and Technology (FASST) initiative.



Elahe Vedadi

Research Scientist, *Google DeepMind*

Elahe Vedadi is a Research Scientist at Google DeepMind, where she specializes in applying large language models (LLMs) to major challenges in biomedicine and scientific discovery. Elahe is a core contributor to several of Google's pioneering AI initiatives, including the Co-Scientist effort, MedGemini, and AMIE, Google's AI-powered diagnostic conversational system. Her path to Google includes targeted research internships at Google Research and Seagate Technology during her doctoral studies. Elahe received her PhD in Electrical Engineering from the University of Illinois Chicago in 2023, and her Bachelor of Science in Electrical Engineering from Sharif University of Technology in 2018.



Rio Yokota

Professor, *Institute of Science Tokyo*

Rio Yokota is a Professor at the Supercomputing Research Center, Institute of Science Tokyo. He also leads the AI for Science Foundation Model Research Team at RIKEN CCS. His research interests lie at the intersection of HPC and ML. He has been optimizing algorithms on GPUs since 2007, and was part of a team that received the Gordon Bell prize in 2009 using the first GPU supercomputer. He has been leading distributed training efforts on Japanese supercomputers such as ABCI, TSUBAME, and Fugaku.



Sponsors

FOUNDATION SPONSOR



Amazon Web Services

Since 2006, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud. AWS has been continually expanding its services to support virtually any workload, and it now has more than 240 fully featured services for compute, storage, databases, networking, analytics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, media, and application development, deployment, and management from 114 Availability Zones within 36 geographic regions. Millions of customers — including the fastest-growing startups, largest enterprises, and leading government agencies — trust AWS to power their infrastructure, become more agile, and lower costs.

Learn more at www.aws.amazon.com.

PLATINUM SPONSORS



Intel

Intel, the world leader in silicon innovation, develops technologies, products, and initiatives to continually advance customer missions. Providing advanced manageability, security, and sustainable performance, Intel business-optimized technologies address the challenges and opportunities within the public sector both today and tomorrow.

Learn more at www.intel.com/content/www/us/en/government/public-sector-solutions-overview.html.



Quantinuum

As the world's largest integrated quantum company, Quantinuum is leading the development of the most powerful quantum computers and the most advanced quantum software solutions. With our full-stack technology and world-class scientists, we are rapidly scaling quantum computing to solve tomorrow's biggest challenges. We are making the impossible possible.

Learn more at www.quantinuum.com.

CORPORATE SPONSORS



Articul8 AI

Articul8 AI is transforming enterprise data and expertise into high-performance engines of growth, value, and impact. Our full-stack GenAI platform is purpose-built for mission-critical environments, combining domain-specific models, autonomous reasoning, and multimodal orchestration to solve complex enterprise challenges with ROI in hours to weeks. General-purpose GenAI isn't enough — we enable expert-level applications that encode domain knowledge, automate decisions, and deliver precise, explainable outcomes. Designed for regulated industries, our platform meets the highest standards for security, performance, and traceability. Trusted by NVIDIA, Intel, EPRI, and others, Articul8 brings secure, scalable GenAI to the frontlines of enterprise operations at HPC speed.

Learn more at www.articul8.ai.



Google Cloud

Google Cloud is the new way to the cloud, providing AI, infrastructure, developer, data, security, and collaboration tools built for today and tomorrow. Google Cloud offers a powerful, fully integrated and optimized AI stack with its own planet-scale infrastructure, custom-built chips, generative AI models, and development platform, as well as AI-powered applications, to help organizations transform. Customers in more than 200 countries and territories turn to Google Cloud as their trusted technology partner.

Learn more at cloud.google.com.



Hammerspace

Hammerspace is a high-performance data platform built to simplify AI infrastructure at scale. It makes all your data immediately accessible, anywhere — across on-prem and cloud environments — without copying or migrating data. Hammerspace integrates with your existing storage, networking, and applications to create a unified, high-speed data backbone for AI, accelerating every stage of the AI pipeline while eliminating data silos.

Learn more at www.hammerspace.com.



Hewlett Packard

Hewlett Packard Enterprise

HPE is a leader in essential enterprise technology, bringing together the power of AI, cloud, and networking to help organizations achieve more. As pioneers of possibility, our innovation and expertise advance the way people live and work. We empower our customers across industries to optimize operational performance, transform data into foresight, and maximize their impact. Unlock your boldest ambitions with HPE.

Learn more at www.hpe.com.



InspireSemi

InspireSemi provides revolutionary high-performance, energy-efficient accelerated computing solutions for HPC, AI, graph analytics, and other compute-intensive workloads. The Thunderbird “supercomputer-cluster-on-a-chip” is a disruptive, next-generation datacenter accelerator designed to address multiple underserved and diversified industries, including computer-aided engineering, energy, climate modeling, cybersecurity, financial services, and life sciences & drug discovery. Based on the open standard RISC-V instruction set architecture, InspireSemi’s solutions set new standards of performance, energy efficiency, and ease of programming.

Learn more at www.inspiresemi.com.



Parallel Works

Parallel Works is the creator of ACTIVATE, a hybrid multi-cloud computing control plane that simplifies the provisioning, management, and scaling of complex compute environments. Designed for HPC, AI, and enterprise workflows, ACTIVATE enables seamless access to on-premises and cloud resources through a unified, intuitive interface. With built-in cost controls, budgeting tools, and containerized user environments, ACTIVATE accelerates collaboration, improves infrastructure efficiency, and supports secure, large-scale computing for research, analytics, and simulation. Organizations use ACTIVATE to modernize operations and unlock the full potential of their hybrid infrastructure.

Learn more at parallelworks.com.



SambaNova

Welcome to SambaNova: Revolutionizing AI Capacity. At SambaNova, we’re empowering developers, enterprises, governments, and data centers to unlock their full AI potential. Our full-stack infrastructure, from chips to models, enables lightning-fast performance, low power consumption, and high-efficiency computing.

Learn more at www.sambanova.ai.

INSTITUTIONAL SPONSORS



Argonne National Lab

Argonne National Laboratory (ANL) is a multidisciplinary science and engineering research center where leading scientists and engineers work together to answer the biggest questions facing humanity — from how to obtain reliable and affordable energy, to how to protect ourselves from emerging threats. The laboratory works in concert with universities, industry, and other national laboratories on questions and experiments too large for any one institution to do by itself. Through collaborations here and around the world, we strive to discover new ways to develop energy innovations through science, create novel materials molecule-by-molecule, and gain a deeper understanding of our planet and the cosmos.

Learn more at www.anl.gov.



The Institute of Science Tokyo

The Institute of Science Tokyo was established on October 1, 2024, following the merger between Tokyo Medical and Dental University (TMDU) and Tokyo Institute of Technology (Tokyo Tech), with the mission of “advancing science and human wellbeing to create value for and with society.” Tokyo Tech is known in the HPC field for installing the first GPU supercomputer, TSUBAME 1.2, back in 2008. The academic structure of the institute consists of six schools, two graduate schools, two faculties, and an institute for liberal arts, with 13,000 students (6,000 undergraduate, 7,000 graduate) and 1,900 faculty members.

Learn more at www.isct.ac.jp/en.





Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory (Berkeley Lab) is committed to groundbreaking research focused on discovery science and solutions for abundant and reliable energy supplies. The lab's expertise spans materials, chemistry, physics, biology, earth and environmental science, mathematics, and computing. Researchers from around the world rely on the lab's world-class scientific facilities for their own pioneering research. Founded in 1931 on the belief that the biggest problems are best addressed by teams, Berkeley Lab and its scientists have been recognized with 16 Nobel Prizes. Berkeley Lab is a multiprogram national laboratory managed by the University of California for the U.S. Department of Energy's Office of Science. DOE's Office of Science is the single largest supporter of basic research in the physical sciences in the United States, and is working to address some of the most pressing challenges of our time.

Learn more at www.lbl.gov.



Oak Ridge National Lab

Oak Ridge National Laboratory (ORNL) delivers the scientific discoveries and technical breakthroughs required to realize solutions in energy and national security and provide economic benefit to the nation. ORNL addresses national needs through impactful research and world-leading research centers. A wide range of partnerships with other US Department of Energy (DOE) laboratories and programs, universities, and industry pairs ORNL's strengths with others for outstanding contributions to science.

Learn more at www.ornl.gov.



Riken

RIKEN, a National Research and Development Agency, is Japan's largest comprehensive research institution renowned for high-quality research in a diverse range of scientific disciplines. Founded in 1917, initially as a private research foundation, RIKEN has grown rapidly in size and scope, today encompassing a network of world-class research centers and institutes across Japan. RIKEN promotes the science of computing, by computing, and for computing — implementing the newest research that integrates simulation, big data analysis, and AI through HPC to solve scientific and social issues and to bring about revolutionary development in our society.

Learn more at www.riken.jp/en.



University of Illinois Chicago

The University of Chicago is a world-renowned private research university dedicated to pushing the boundaries of knowledge and fostering innovative solutions to global challenges. Its faculty and students pursue transformative scholarship across the physical, biological, and social sciences, the humanities, and professional schools. Home to pioneering work in economics, physics, medicine, and computer science — including foundational contributions to artificial intelligence — the university partners closely with Argonne National Laboratory and Fermilab to accelerate discovery at scale. With an ethos of rigorous inquiry, interdisciplinary collaboration, and public impact, the University of Chicago is proud to support the Trillion Parameter Consortium's vision and research-driven mission.

Learn more at www.uic.edu.

NON-PROFIT SPONSOR



The HPC-AI Society

The HPC-AI Society is a vendor-neutral, non-profit organization whose sole purpose is to educate and promote the common business and technology interests of the HPC-AI community, with the goal of accelerating HPC, AI, data science, and quantum computing through community collaboration. The Society organizes and conducts open forum meetings and manages noncompetitive research activities that address the use, availability, standardization, and evaluation of existing technology, while introducing emerging technology to the HPC-AI community. Membership is open to all HPC and AI professionals serving all industries, government, and academia.

Learn more at www.hpc-ai-society.org.

TPC25 Breakout Group Lightning Talks

This schedule is up to date as of July 29. Minor changes may occur.

WORKFLOWS TRACK (Main Plenary room)

Data Workflows, Agents, and Reasoning Frameworks (DWARF)

Keynote and Systems Software for Agents (Wed, July 30, 16:00)

Building a Scientific Reasoning Platform: Realizing a Discovery Cloud (Session Keynote) — *Ian Foster (University of Chicago and Argonne National Laboratory)*

A Case Study of the System Software/Middleware Needs for Agents — *Robert Underwood (Argonne National Laboratory)*

Enabling Autonomous Labs: The NSDF-ORNL Partnership for Real-Time Scientific Discovery — *Michela Taufer (University of Tennessee at Knoxville)*

Academy: Empowering Scientific Workflows with Federated Agents — *Kyle Chard (University of Chicago)*

Scalable Scientific Data/Scientific Data for AI (Thu, July 31, 8:30)

LangChain-Parsl: Connect Large Language Model Agents to High Performance Computing Resources — *Heng Ma (Argonne National Laboratory)*

Building AI Scientific Assistants for Accelerating Understanding of Complex Biological Systems — *Arvind Ramanathan (Argonne National Laboratory)*

A Grassroots Network and Community Roadmap for Interconnected Autonomous Science Laboratories for Accelerated Discovery — *Rafael Ferreira da Silva (Oak Ridge National Laboratory)*

Integrating Data and AI to Advance Earth System Predictability — *Po-lun Ma (Pacific Northwest National Laboratory)*

LLM Agent-based Code Translation for Low Resource Programming Languages — *Le Chen (Argonne National Laboratory)*

ChatVis: Automating Scientific Visualization with a Large Language Model — *Tanwi Mallick (Argonne National Laboratory)*

Towards Enhancing Reliability in Agentic Scientific Workflows — *Amal Gueroudji (Argonne National Laboratory)*

From Models to Missions: The Path of Agentic AI — *Nelli Babayan (Microsoft)*

Agents and the Model Context Protocol — *Rodolfo Tonoli (Articul8 AI)*

Scalable Processing Pipelines (Thu, July 31, 11:00)

Towards Scalable Memory Runtimes for LLM Agents with DataStates — *Avinash Maurya (Argonne National Laboratory)*

Imperfect Recognition: A Study of OCR Limitations in the Context of Scientific Documents — *Chinmay Sahasrabudhe (Sandia National Laboratories)*

Why Use Model Context Protocol (MCP) in Scientific Application Domains? — *Elliot Jacopin (RIKEN Center for Biosystems Dynamics Research)*

Empowering Scientific and Supercomputing Users and Their Workloads Using Context Engineering and Domain Specific Models (DSMs) — *Rodolfo Tonoli (Articul8 AI)*

BOF: LLMs for Living Docs (Thu, July 31, 14:00)

AI Systems for Technical Logbooks — *Aaron Reed (SLAC National Accelerator Laboratory)*

Enhancing APS Logbook Search with Retrieval-Augmented Generation Models — *Rajat Sainju (Argonne National Laboratory)*

LCLS-Elog-Copilot: An AI Agent to Navigate LCLS Experiment Metadata — *Cong Wang (SLAC National Accelerator Laboratory)*

LLMs for Dark Matter: Knowledge Management Across Multi-Modal Silosed Resources — *Maria Elena Monzani (SLAC and Stanford University)*

LLMs for Living Docs — *Kuktae Kim (SLAC National Accelerator Laboratory)*

INITIATIVES TRACK (Cascade room)

BOF: Building Foundation Models for the Electric Grid (GridFM) (Wed, July 30, 14:00)

Designing a Unified Data Structure for Multi-task Training of GridFM — *Zhirui Liang (Argonne National Laboratory)*

Federated Learning Framework for Collaborative Training of Electric Grid Foundation Models — *Yijiang Li (Argonne National Laboratory)*

Foundation Models for the Electric Grid — *Hendrik Hamann (Stony Brook University and Brookhaven National Laboratory)*

GridFM Models for Distribution System State Estimation — *Stefano Fenu (Argonne National Laboratory)*

GridFM: A Foundation Model for Power Grid Intelligence via Heterogeneous GNNs — *Hongwei Jin (Argonne National Laboratory)*

The First Set of Domain-specific GenAI Models for Electric and Power Systems to Accelerate the Open Power AI Consortium (OPAI) — *Rodolfo Tonoli (Articul8 AI)*



BOF: AI in Decision Sciences (Thu, July 31, 8:30)

AI in Epidemiology — *Peter Nugent (Lawrence Berkeley National Lab)*

Satisficing at Scale: How AI Connects Community Knowledge with Scientific Models for Actionable Solutions — *Suzanne A Pierce (Texas Advanced Computing Center, The University of Texas at Austin)*

The Wall Confronting Large Language Models — *Peter Coveney (Argonne National Laboratory and University College London)*

Urgent AI for Decision Science — *Manish Parashar (Scientific Computing and Imaging Institute, University of Utah)*

BOF: Energy Efficient HPC for AI Workloads (Thu, July 31, 14:00)

AI & HPC Facility Trends — *Jason Hick (Los Alamos National Laboratory)*

Balancing Power and Performance for AI and ML on Heterogeneous HPC Systems — *Brice Videau (Argonne National Laboratory)*

LIFE SCIENCES TRACK (Sierra room)

AI for Cancer (Wed, July 30, 14:00)

Scalable AI for Pediatric Cancer: Unlocking Precision Medicine from Discovery to Treatment — *Ninad Oak (St. Jude Children's Research Hospital)*

Global Federated Learning Enabled by the Planet9 Ecosystem — *Xun Zhu (St. Jude Children's Research Hospital)*

Accelerating Peptide Binder Design for Cancer Using Generative AI and Multiscale Simulations — *Matt Sinclair (Argonne National Laboratory)*

Antibody Design Using Preference Optimization and Structural Inference — *Archit Vasan (Argonne National Laboratory)*

Visualizing Collaborative Intelligence — *Bharat Kale (Argonne National Laboratory)*

Agentic AI and Foundation Models (Wed, July 30, 16:00)

Biomni: A General-purpose Biomedical AI Agent — *Keixin Huang (Stanford University)*

Cracking Shells: Streamlining MCP Server Management for Scientific Software — *Jacopin Elliott (RIKEN Center for Biosystems Dynamics Research)*

Biological Reasoning System (BioR5): A Three-Layer AI Architecture — *Peng Ding (Argonne National Laboratory)*

AI for Biology (Thu, July 31, 8:30)

Leveraging AI-driven Protein Structure Prediction to Decode Phosphorylation Effects on CSF1R Kinase Domain Dimerization — *Moeen Meigooni (Argonne National Laboratory)*

Workflow for Fine-tuning Genome-scale Language Models for Generative Enzyme Design — *Xinran Lian (Argonne National Laboratory)*

BOF: Federated Learning at Scale (Thu, July 31, 14:00)

Differentially Private Federated Averaging with James-Stein Estimator — *Xinran Zhao (Arizona State University)*

Federated Learning at Scale: Privacy-preserving Collaboration on Frontier Supercomputer — *Olivera Kotevska (Oak Ridge National Laboratory)*

Scalable Federated Learning Across DOE HPC Clusters — *Yijiang Li (Argonne National Laboratory)*

Scaling Secure Collaboration: Real-world Federated Learning with FLARE — *Chester Chen (NVIDIA)*

SR-APPFL: Scalable and Resilient Advanced Privacy-preserving Federated Learning — *Xiaoyi Lu (University of California, Merced)*

SyftBox: A Networked Protocol for Federated Training with Minimal Coordination — *Irina Bejan (OpenMined)*

Training Scalability of the APPFL Framework — *Zilinghan Li (Argonne National Laboratory)*

EVALUATION TRACK (Siskiyoo room)

Model Skills, Reasoning, and Trust Evaluation (EVAL)

Intro and Benchmarks (Wed, July 30, 14:00)

Guarding the Future: Advancing Risk Assessment, Safety Alignment, and Guardrail Systems for AI Agents (Session Keynote) — *Bo Li (University of Illinois Urbana-Champaign and Virtue AI)*

EAIIRA: Establishing a Methodology for Evaluating AI Models as Scientific Research Assistants — *Franck Capello (Argonne National Laboratory)*

SciCode: A Research Coding Benchmark Curated by Scientists — *Eliu Huerta (Argonne National Laboratory)*

LLM Evaluation on Biological Science — *Shinjae Yoo (Brookhaven National Laboratory)*

Exploring the Capabilities of the Frontier Large Language Models for Nuclear Energy Research — *Prasanna Balaprakash (Oak Ridge National Laboratory)*

Astrophysics Benchmarking of LLMs — *Nesar Ramachandra (Argonne National Laboratory)*

UQ and Safety (Wed, July 30, 16:00)

Safety Evaluation of Test-time Compute-constrained Reasoning Language Models — *Adarsha Balaji (Argonne National Laboratory)*

UProp: Investigating the Uncertainty Propagation of LLMs in Multi-step Agentic Decision-Making — *Kaidi Xu / Jinhao Duan (Drexel University)*

Robustness and Safety-constrained Generation — *Ferdinando Fioretto (University of Virginia)*

Double-blind Evaluation via GPU Enclaves: A Path to Trustworthy Model Assessment — *Irina Bejan (OpenMined)*

Evaluating Probability Consistency and Trust in Large Language Models — *Bradley Love (Los Alamos National Laboratory)*

Automatic Benchmark Generation (Thu, July 31, 8:30)

Automated Multiple Choice Question Answering Benchmark Generation and Model Evaluation — *Ozan Gokdemir (Argonne National Laboratory)*

LLM Judges — *Neil Getty (Argonne National Laboratory)*

DoReMi: Difficulty-oriented Reasoning Effort Modeling of Science Problems for Language Models — *Cong Xu (HPE)*

Evaluation of Multimodal Understanding in Foundation Models for Geo-Spatial Data — *Tanwi Mallick (Argonne National Laboratory)*

Advance Evaluation (Thu, July 31, 11:00)

End-to-end Evaluation: Lab Style and Field Style Experiments (And 1000 Scientists AI JAM) — *Franck Capello (Argonne National Laboratory)*

Prioritizing Skills and Capabilities for Science Assistant Evaluation — *Patrick Emami (National Renewable Energy Lab)*

Semibench: A Microtask Benchmark for Semiconductor Manufacturing AI Agents — *Angel Yanguas-Gil (Argonne National Laboratory)*

Evaluating Agentic AI for Science: Insights from ChemGraph — *Thang Duc Pham (Argonne National Laboratory)*

BOF: Trustworthiness in Scientific Machine Learning: From Inference-time-compute to Reasoning (Thu, July 31, 14:00)

Challenges in Reasoning in Scientific Machine Learning with PDEs — *Siddharth Mansingh (Los Alamos National Laboratory)*

Evaluating Probability Consistency and Trust in Large Language Models — *Bradley Love (Los Alamos National Laboratory)*

Diagnostic Metrics and Reasoning for Trustworthy PDE Surrogates — *James Amarel (Los Alamos National Laboratory)*

SCALE AND SERVICES TRACK (Donner room)

BOF: Deployment of Inference-for-Science Services at HPC Centers

Session 1 (Wed, July 30, 14:00) and **Session 2** (Wed, July 30, 16:00)

Persistent Inference Services for Science: Progress at TACC to Date — *Dan Stanzione (Texas Advanced Computing Center / The University of Texas at Austin)*

Exploring GPT-based AI services at Pitt and PSC to Support Scientific Research — *Barr von Oehsen (Pittsburgh Supercomputing Center)*

Leveraging a First-party Inference Service for HPC User Support — *Mitja Sainio (CSC – IT Center for Science)*

Serving AI Models on NCSA's HPC Systems — *Volodymyr Kindratenko (National Center for Supercomputing Applications)*

The AI RAP and its Integration with Heterogeneous Inference Accelerators — *Yun-Te Lin (National Center for High-performance Computing)*

Deploying Scalable Inference Endpoints for Science at ALCF — *Venkatram Vishwanath (Argonne National Laboratory)*

Session 2 (Wed, July 30, 16:00)

Scaling In-memory Computing to Data Center Levels for Fast and Efficient Science Inference — *Satyam Srivastava (d-Matrix)*

Challenges and Strategies for Deploying Scalable AI Inference — *Panagiotis Kourdis (Intel)*

Generative AI Inference at Scale: A World of Trade-offs — *Darshan Gandhi (SambaNova)*

Model Architecture and Performance Evaluation (MAPE)

Session 1 (Thu, July 31, 8:30)

Performance Modeling and System Design Insights for AI Foundation Models — *Shashank Subramanian (Lawrence Berkeley National Laboratory)*

Characterizing GPU Memory Errors: Insights from a Cross-supercomputer Study — *Lishan Yang (George Mason University)*

Design Choices for Compute-Efficient Mixture-of-Expert Models — *Daria Soboleva (Cerebras Systems)*

X-MoE: Enabling Scalable Training for Emerging Mixture-of-Experts Architectures on HPC Platforms — *Minjia Zhang (University of Illinois at Urbana-Champaign)*

Session 2 (Thu, July 31, 11:00)

Architecture of AERIS, an Argonne Earth Systems Model — *Väinö Hatanpää (Argonne National Laboratory)*

MegaFold: System-level Optimizations for Accelerating Protein Structure Prediction Models — *Minjia Zhang (University of Illinois at Urbana-Champaign)*

Diamond: Democratizing Large Foundation Model Training for Science — *Zhao Zhang (Rutgers University)*



Communication-efficient Large Language Model Optimization — Zhao Zhang (Rutgers University)

AI-powered Performance Insights — From Data to Predictions — Ashwin M. Aji (AMD)

Session 3 (Thu, July 31, 14:00)

Agentic Systems on SN40L Dataflow Architecture — Darshan Gandhi (SambaNova)

Machine Learning-guided Memory Optimization for DLRM Inference on Tiered Memory — Dong Li (University of California Merced)

Eliminating Communication in LLM Training via Generic Tensor Slicing and Overlapping — Dong Li (University of California Merced)

Phoenix: Enabling Sparse Fine-tuning for Foundation Model Downstream Tasks on Cerebras — Wenqian Dong (Oregon State University)

Quantum/HPC Hybrid Solutions in the Cloud — Sebastian Hassinger (AWS)

APPLICATIONS TRACK (Cedar room)

AI Models for Software Engineering and Development (Wed, July 30, 14:00)

Agentic Systems for Scientific Code Generation — Tanwi Mallick (Argonne National Lab)

CelloAI: Leveraging Large Language Models for HPC Software Development in High Energy Physics — Meifeng Lin (Brookhaven National Laboratory)

ChatHPC, AI-assistance for HPC Programming and Software Ecosystem — Keita Teranishi (Oak Ridge National Laboratory)

CLEVER: A Benchmark for Two-staged, End-to-end Verified Code Generation — Amitayush Thakur (University of Texas at Austin)

LASSI-EE: Automated Energy-aware Refactoring of Parallel Scientific Codes Using LLMs — Matthew Dearing (University of Illinois Chicago and Argonne National Laboratory)

LLM Agent-based Code Translation for Low Resource Languages — Le Chen (Argonne National Laboratory)

BOF: Foundation Models for Fusion Energy (Wed, July 30, 16:00)

AI for Fusion Diagnostics, Control and Scientific Discovery — Egemen Kolemen (Princeton University and Princeton Plasma Physics Laboratory)

AI for Scientific Control in Magnetic Fusion Energy — William Tang (Princeton University)

Foundation Models in Fusion Energy Simulation and Experiment — Michael Churchill (Princeton Plasma Physics Laboratory)

Surrogate Model of First Principle Simulations of Fusion Plasma — Xishuo Wei (University of California, Irvine)

Surrogate Models as Essential Building Blocks for Fusion Foundation Models — Alvaro Sanchez-Villar (Princeton Plasma Physics Laboratory)

Tokamaks and Tokenization: Enabling the Foundation Model Scale with Electron Cyclotron Emission Imaging Data — Jesse Rodriguez (Oregon State University)

AI for Scientific Discovery in Materials Science (AI4MS) (Thu, July 31, 8:30)

Accelerating Discovery of Novel Materials Using AI — Geetika Gupta (NVIDIA)

ChemGraph: Automating Computational Chemistry with Agentic AI — Thang Pham (Argonne National Laboratory)

Domain-aware Data Compression for Scientific Instruments — Amarjit Singh (RIKEN (R-CCS))

How Can AI for Materials Science Reach Internet Scale? — Anuroop Sriram (Meta)

Machine Learning Force Fields and Generative Models for Atomistic Simulations: Navigating Speed, Accuracy, and Scalability Trade-offs in the Age of (Some) Large-scale Scientific Data — Aditi Krishnapriyan (UC Berkeley)

Scaling Deep Learning for Materials Discovery — Simon Batzner (Google DeepMind)

AI Education and Outreach (Thu, July 31, 11:00)

ADAPT PA: Giving Students Computing Skills for Every Career Path — Barr von Oehsen (Pittsburgh Supercomputing Center)

Code Green Jam: Peer Student Learning to Write Energy-efficient Parallel Code — Matthew Dearing (University of Illinois Chicago and Argonne National Laboratory)

Lessons Learned from Vibe Coding Using Warp — Arvind Ramanathan (Argonne National Laboratory)

Lessons Learned From the Helsinki Hackathon — Miguel Vazquez (Barcelona Supercomputing Center)

Earth and Environment (AI for Digital Earth) (Thu, July 31, 14:00)

Characterizing Extreme Weather in a Huge Ensemble of Machine Learning Weather Forecasts — Ankur Mahesh (UC Berkeley and Lawrence Berkeley National Laboratory)

Climate in a Bottle: Steerable, Foundational Climate State Sampling at 13 Megapixel Ambition — Mike Pritchard (NVIDIA)

Fairness of Geospatial Foundation Models — Kyoungsook KIM (National Institute of Advanced Industrial Science and Technology (AIST))

Foundation Models for Earth Systems — Hendrik Hamann (Stony Brook University)

CONFERENCE MAP

