ANL-22/91

ADVANCED RESEARCH DIRECTIONS ON

AI FOR SCIENCE, ENERGY, AND SECURITY

Report on Summer 2022 Workshops

Jonathan Carter
Lawrence Berkeley National Laboratory

John Feddema
Sandia National Laboratories

Doug Kothe
Oak Ridge National Laboratory

Rob Neely
Lawrence Livermore National Laboratory

Jason Pruet
Los Alamos National Laboratory

Rick Stevens
Argonne National Laboratory

U.S. DEPARTMENT OF ENERGY
U.S. DEPARTMENT OF ENERGY | Office of Science
NNSA National Nuclear Security Administration

May 2023

ABOUT ARGONNE NATIONAL LABORATORY

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

DOCUMENT AVAILABILITY

Online Access: U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free at OSTI.GOV **(www.osti.gov)**, a service of the US Dept. of Energy's Office of Scientific and Technical Information.

Reports not in digital format may be purchased by the public from the National Technical Information Service (NTIS):
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312
www.ntis.gov
Phone: (800) 553-NTIS (6847) or (703) 605-6000
Fax: (703) 605-6900
Email: **orders@ntis.gov**

Reports not in digital format are available to DOE and DOE contractors from the Office of Scientific and Technical Information (OSTI):
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
www.osti.gov
Phone: (865) 576-8401
Fax: (865) 576-5728
Email: **reports@osti.gov**

Cover design by Argonne National Laboratory. Cover imagery via Shutterstock.

ERRATA

After initial publication of this report the authors were made aware of some minor errors. These errors were corrected in the present version on May 9, 2024.

Location	Error	Correction
Page 11	Missing citation to [14] in **Accelerators** section	Text corrected to "…in many cases [13], promising initial results have been observed by developing customized AI/ML methods that automatically compensate for unknown time-varying changes to accelerator components (such as magnets, and to unknown changes in the accelerator's input beam distribution) [14]."

[This page left intentionally blank]

Advanced Research Directions on AI for Science, Energy, and Security

Report on the U.S. Department of Energy (DOE) Summer 2022 Workshop Series on Artificial Intelligence (AI) for Science, Energy, and Security

Program Committee

Jonathan Carter	Associate Laboratory Director, Lawrence Berkeley National Laboratory
John Feddema	Program Lead, Sandia National Laboratories
Doug Kothe	Associate Laboratory Director, Oak Ridge National Laboratory
Rob Neely	Program Director, Lawrence Livermore National Laboratory
Jason Pruet	Program Director, Los Alamos National Laboratory
Rick Stevens	Associate Laboratory Director, Argonne National Laboratory

U.S. Department of Energy Contact

Ceren Susut-Bennett	Program Manager, U.S. Department of Energy

Key Contributors

Argonne National Laboratory
Prasanna Balaprakash, Pete Beckman, Ian Foster, Kamil Iskra, Arvind Ramanathan, Valerie Taylor, Rajeev Thakur

Lawrence Berkeley National Laboratory
Deb Agarwal, Silvia Crivelli, Bert de Jong, Damian Rouson, Mike Sohn, Michael Wetter, Stefan Wild

Lawrence Livermore National Laboratory
Timo Bremer, Michael Goldman, Ana Kupresanin, Luc Peterson, Brian Spears, Dave Stevens, Brian Van Essen

Los Alamos National Laboratory
Russell Bent, Mike Grosskopf, Earl Lawrence, Galen Shipman

National Energy Technology Laboratory
Kelly Rose

National Renewable Energy Laboratory
Ray Grout

Oak Ridge National Laboratory
Nicholson Kouakpaizan, Femi Omitaomu, Slaven Peles, Pradeep Ramuhalli, Arjun Shankar, David Womble, Guannan Zhang

Sandia National Laboratories
Tommie Catanach, Ron Oldfield, Siva Rajamanickam, Jaideep Ray

Sustainable Horizons Institute
Mary Ann Leung

Editorial

Charlie Catlett	Senior Computer Scientist, Argonne National Laboratory
Emily M. Dietrich	Strategic Program Communications Lead, Argonne National Laboratory

Special Thanks

To the Argonne National Laboratory Communications and Public Affairs Division’s Writing Center of Excellence, including key support from Andrea Manning and Lorenza Salinas.

CONTENTS

EXECUTIVE SUMMARY

Over the past decade, fundamental changes in artificial intelligence (AI)—from foundational to applied—have delivered dramatic insights across a wide breadth of U.S. Department of Energy (DOE) mission space. AI is helping to augment and improve scientific and engineering workflows (e.g., for control, design, and dramatic performance gains through surrogate models) in national security, the Office of Science, and DOE's applied energy programs. The progress and potential for AI in DOE science was captured in the 2020 "AI for Science" report from the DOE laboratory community in collaboration with academia and industry. Specific scientific areas ready to further leverage the power of AI ranged from the scale and performance of computational models to data analysis to creating new classes of observations using computer vision. Since that report, the scale and scope of scientific AI have accelerated, revealing new, emergent properties that yield insights that go beyond enabling opportunities to being potentially transformative in the way that scientific problems are posed and solved.

Thus, under the guidance of both the Office of Science (SC) and the National Nuclear Security Administration (NNSA), the DOE national laboratories organized a series of workshops in 2022 to gather input on new and rapidly emerging opportunities and challenges of scientific AI. This 2023 report is a synthesis of those workshops. The scientific community believes AI can have a foundational impact on a broad range of DOE missions, including science, energy, and national security. Further, DOE has unique capabilities that enable the community to drive progress in scientific use of AI, building on long-standing DOE strengths and investments in computation, data, and communications infrastructure, spanning the Energy Sciences Network (ESnet), the Exascale Computing Project (ECP), and integrative programs such as the NNSA Office of Defense Programs Advanced Simulation and Computing (ASC) and the SC Scientific Discovery through Advanced Computing (SciDAC) programs.

Today, the urgency to undertake a major and transformational initiative in AI is increasing, fueled both by the acceleration of AI advancements and the robust international activity and investments to capture these advancements. Moreover, the introduction of powerful language models in public-facing Internet services such as those from OpenAI, Microsoft, Meta, and Google have revealed a pressing need for fundamental understanding of new, emergent capabilities of these models and the associated risks to society. This report details the criticality of harnessing AI to advance science and address national imperatives such as energy and security, laying out a research agenda that is equally relevant and desperately needed, while also addressing challenges such as those discussed in an April 2023 open letter from the Association for the Advancement of Artificial Intelligence (AAAI), including "*the potential for AI systems to make errors, to provide biased recommendations, to threaten our privacy, to empower bad actors with new tools, and to have an impact on jobs*" [1].

Fields such as natural language processing (NLP) and image recognition have shown game-changing promise, as have the design, engineering, deployment, and operation of complex systems—especially those lying at the heart of DOE's core science, energy, and security mission areas. Progress in designing and deploying supercomputers in China, Japan, Europe, and other nations has resulted in a competitive AI position that cannot be ignored. As AI capabilities begin to transform nearly every aspect of science, energy, and security, establishing leadership in AI and in the underlying capabilities, including high-performance computing (HPC), will be intimately tied to the nation's future and its role in the global order. This race is arguably deeper and more consequential than any the nation has seen in the past 60 years. As noted in the Final Report of the National Security Commission on Artificial Intelligence:

> *No comfortable historical reference captures the impact of artificial intelligence (AI) on national security. AI is not a single technology breakthrough, like a bat-wing stealth bomber. The race for AI supremacy is not like the space race to the moon. AI is not even comparable to a general-purpose technology like electricity. However, what Thomas Edison said of electricity encapsulates the AI future: "It is a field of fields … it holds the secrets which will reorganize the life of the world*." [2].

Similarly, global forces threaten the nation's leadership in semiconductors, despite promising results from ECP, HPC, and data infrastructure. These areas are inextricably tied to leadership in AI, where the most revolutionary advances are empowered by computation and unprecedented volumes of data. The extreme scales offered by exascale systems represent the global stakes for AI competitiveness, but *leadership* will hinge on developing sustainable exascale and beyond-exascale (zettascale) computing environments along with the underlying theory, mathematics, and software systems necessary to exploit the power of those systems. As such, this report lays out six crucial foundational AI methodologies; elucidates their potential to transform DOE's science, energy, and security mission areas; sets forth a broad architecture of crosscutting technology areas that must be advanced to enable those transformations; and assesses the state of DOE's workforce and the scale, computational capability, and data infrastructure with respect to the department's ability to affect those advancements.

Global leadership—empowered by comprehensively and aggressively embracing and advancing AI across DOE—will also require bold initiatives in at least three dimensions. The first is to address increasingly disruptive workforce challenges, notably the diversion of talent from fundamental and applied sciences at DOE laboratories and academia toward supporting commercial applications where only a subset of incentives and goals align with DOE missions. The second is to capitalize on lessons learned through designing and deploying exascale systems, from semiconductors and HPC system co-design efforts through computing and storage system integration to system and application software, along with the need in future systems for much closer, and nontraditional, partnerships with industry providers. The third is to fully embrace the nascent potential to harness emergent capabilities of deep learning—exemplified in the AI approaches outlined in the first section of this report—by investing in focused campaigns targeted at DOE mission challenges, all of which are nationally strategic.

This report lays out a comprehensive vision for DOE to leverage and expand new capabilities in AI to accelerate the progress, and deepen the quality of mission areas spanning science, energy, and security. Equally important, the vision and blueprint align precisely with the pressing need for scientific grounding in areas such as bias, transparency and explainability, data security, validation and accuracy, and grappling with the impact of AI on jobs. Much of the most dramatic progress being made in AI comes from industry and defense in the U.S. and other nations, whose objectives and incentives only partially align with DOE's mission. These advances also reflect the migration of AI and computer science talent to industry, creating a workforce disruption that DOE must address with a sense of urgency. Nevertheless, DOE's investments in exascale systems, infrastructure, software, theory, and applications—combined with unique, multidisciplinary co-design approaches scaled to thousands of experts—uniquely position the DOE complex to extend its global leadership in science, energy, and security. Concurrently, these DOE assets and capabilities are uniquely suited to address new, AI-related challenges faced by society today—creating not only opportunity, but the responsibility, to lead the nation and to creatively engage U.S. industry to address those challenges. Focused, sustained campaigns toward the development and application of new AI methods are required, along with their integration into (and in some cases replacement of) the tools and infrastructure supporting DOE mission areas and leveraging of the world-leading human, computational, and data science infrastructure created through the ECP and foundational DOE programs and integrative infrastructure including ESnet, SciDAC, and others.

AI for Science, Energy, and Security: Report Overview

Well over a thousand researchers participated in seven workshops in 2019 and 2022. The workshops in 2019 and the resulting report, “AI For Science,” detailed the opportunities for applying new AI and machine learning (ML) techniques to the DOE enterprise, spanning 16 application areas, including science, energy, security, facilities, and other facets of the complex. Building on this application roadmap, the 2022 workshops were organized around *significant advances in AI that represent emerging challenges and opportunities*, focusing on (1) six broadly applicable *AI building block approaches* with potential to transform the department's modeling, simulation, and experimental processes; (2) the *domain-specific opportunities* they represent for science, energy, and security grand challenges; (3) *crosscutting technologies* that must be adapted or created to enable those opportunities while also addressing significant new challenges associated with emergent properties in AI such as those that are demonstrated with large language models; and (4) the current *state of readiness* in workforce, data, infrastructure, and scale.

Section 01 details six new AI-empowered computing paradigms—AI Approaches (see the Introduction's sidebar). These form a set of building blocks that combine and scale fundamental AI functions, such as inference from large-scale and often unstructured and multi-modal data sources, NLP, and object recognition. These building blocks can be integrated to generate transformational capabilities, from surrogate and foundation models; to digital twins; to automated, real-time control and optimized instruments, experiments, or complex infrastructure, and ultimately autonomous experiments, laboratories, and instruments; to automated software engineering and programming. Executing on these AI-empowered computing paradigms is timely given recent discoveries of e*mergent capabilities that represent new classes of AI models*, such as large language models (underlying products like ChatGPT,[1] Bard, and Bing) and foundation models, and accelerated progress in capabilities such as physics-informed surrogate models.

In Section 02, we show the impact of applying these new AI approaches to the unique challenges of DOE's application and program areas in *basic science*, *energy*, and *national security programs*, as well as the emerging *Energy Earthshots*. Achieving these transformations will require fundamental changes in the nature of computational workloads, significantly

[1] The remarkable popularity and societal concern regarding OpenAI's ChatGPT—growing to 100M users during the several months' time that this report was compiled—illustrates the urgency and criticality of the research and development outlined in this report.

increasing the scale of computational and data resources needed as workloads shift to encompass model training as well as exploring a broader range of model scenarios. Transforming our effectiveness in addressing DOE science, energy, and security challenges requires rethinking foundational concepts, including traditional simulation, modeling, and data analysis approaches and meeting new and rapidly evolving demands placed upon underlying physical and software infrastructure.

Section 03 describes five key crosscutting technology challenge areas that must be addressed to bridge the gap between model-driven methods and data-driven methods; develop the underlying mathematical and foundations of scientific machine learning; and create new integrative systems—themselves empowered by the new approaches outlined in Section 01. These demand advances in theory and foundational mathematics and computer science methods. The importance of these multidisciplinary challenges is illustrated by the paradigm-shifting opportunities outlined throughout Sections 01 and 02, but these core capabilities are also manifest, and their importance amplified, in precisely the concerns expressed today with respect to AI safety and ethics, including a proposed framework for an *Ethics Framework to Guide AI RD&D* [1].

We conclude with Section 04 by assessing the current state and highlighting the challenges, opportunities, and strategies necessary to advance and leverage new AI capabilities, translating decades of investment and advancement of DOE's world-leadership in modeling, simulation, and infrastructure into world-leadership in AI-empowered science, energy, and security systems. This will require the DOE workforce, scale of operation, computational and data resources, and instrumentation to be similarly transformed to meet the challenges and achieve the vision captured in this report.

ES.1 References

[1] Association for the Advancement of AI, 2023. Working together on our future with AI, April 5. https://aaai.org/working-together-on-our-future-with-ai/, accessed May 12, 2023.

[2] National Security Commission on Artificial Intelligence, 2021. *Final Report*, October. https://www.nscai.gov/2021-final-report, accessed December 16, 2022.

[3] Grout, R., Rose, K., Taylor, V., and Essen, B., 2022. *AI@DOE Interim Executive Report*, United States. https://doi.org/10.2172/1872103, https://www.osti.gov/servlets/purl/1872103, accessed May 9, 2023.

INTRODUCTION: ADVANCED RESEARCH DIRECTIONS ON AI FOR SCIENCE, ENERGY, AND SECURITY

Within the backdrop of recent developments—for example, AI's broad and fast-paced advances and potential impact on society, the rising tide of experimental and observational data, and availability of extreme-scale compute systems such as those deployed through the U.S. Department of Energy's (DOE's) exascale computing programs—DOE's core missions in science, energy, and security stand at an inflection point. Decades of investments in world-class physical experimental, observational, and computational infrastructure; the underlying theory, modeling, and software necessary for the design, operation, and optimization thereof; and the diverse design, operational, and scientific expertise and experience necessary to use that infrastructure all provide the nation with world-leading capabilities. These foundational human and technology infrastructure assets uniquely position DOE to harness and drive new and emerging capabilities in artificial intelligence (AI), directly addressing research questions that we now see thrust into the public discourse regarding the benefits and dangers of powerful AI.

AI APPROACHES

New AI-Empowered Computing Paradigms, known in this report as AI Approaches

The scale of data and computation for training AI models is opening the potential today for new paradigms in computation, including the following AI Approaches:

01. AI and Surrogate Models for Scientific Computing
02. AI Foundation Models for Scientific Knowledge Discovery, Integration, and Synthesis
03. AI for Advanced Property Inference and Inverse Design
04. AI-Based Design, Prediction, and Control of Complex Engineered Systems
05. AI and Robotics for Autonomous Discovery
06. AI for Programming and Software Engineering

Harnessing DOE Leadership in Computation and Data

The most promising advances in AI result from scale, thus computational capacity and capabilities are central to driving the future of AI [1]. The nationwide Exascale Computing Project (ECP) team of over 1,000 scientists, engineers, and program support staff from DOE laboratories, academia, and industry has positioned DOE uniquely in this respect, having created a vision for exascale computing and then developing, organizing, and executing a DOE complex-wide campaign to not merely lead the world but to *redefine the field*. In 2022, the Exascale Computing Initiative (ECI) demonstrated this paradigm shift, deploying the world's first exascale supercomputer—the highest ranked world-wide, with more capability and capacity than the next four ranked systems *combined.*[2] In 2023, the second DOE exascale machine will provide twice this capacity.

ECP leveraged decades of investment in software, facilities, and scientific workforce, including programs such as the National Nuclear Security Administration (NNSA) Defense Programs Advanced Simulation and Computing (ASC) and Office of Science (SC) Scientific Discovery through Advanced Computing (SciDAC) and Energy Science Network (ESnet) programs. DOE's science, energy, and national security mission areas have relied on this infrastructure supporting physics-based modeling and simulation as an underlying paradigm for discovery and design as well as for operations. This paradigm spans every facet of computation, from basic mathematical algorithms and libraries to system software; workflow and data management to applications; and encompassing processing, memory, storage, and communications hardware and system architectures.

Nevertheless, the dividends of these intellectual and financial investments have also exacerbated growing challenges in model, code, and workflow complexity. Similarly, the enormity of the data produced by models, even on sub-exascale systems, has outstripped traditional data management, curation, and analysis capacity, which are similarly complex and reliant on human experts. Exploiting the potential of emerging, extreme-scale AI models such as surrogate or foundation models will place the entire data management infrastructure in the critical path for computation rather than the traditional role of repositories. Moreover, these challenges are central to a critical concern facing society today: understanding the data used to train large language

[2] The November TOP500 rankings show the ECP system, Frontier, at 1102 Petaflop/s, leading Japan's Fugaku (442), EU's Lumi (309) and Leonardo (174), and DOE's Summit (149), also an ECI system.

models [2]. Here, AI also offers new approaches to managing scale and complexity for both the data and computational software infrastructure. The resulting transformation will yield complex models that retain resilience and robustness yet are more agile and flexible. This outcome will bring deeper integration of complex workflows combining experiment and computational models.

DOE's world leadership in exascale computing and the broader aspects of computation and related infrastructure directly translate to leadership in science, engineering, and security—all of which rely on computational modeling and simulation. But ECP also accelerated progress in pioneering applications, such as ExaLearn and CANDLE (both described in Chapter 17), that leverage unprecedented advances in AI and machine learning (ML), including those that are *only unlocked through exascale computation and commensurate scales of data.* There remains much untapped potential for these innovations to drive new science, energy, and security discoveries but also to accelerate the pace of discovery itself [3]. Moreover, implementing new AI models within traditional modeling and simulation approaches has resulted in both entirely new large-scale, data-driven workflows for exascale systems and extraordinary improvements in computation rates, ultimately multiplying the capacity of those systems.

ECP also revealed challenges that are amplified by the scale of computation and data necessary to fully embrace new AI methods, which will require sustained growth in both the capabilities of individual exascale and post-exascale simulation technologies and the overall capacity of computational and data resources supporting DOE mission areas. Addressing the prodigious costs of design, deployment, and operation of exascale systems will itself require AI models. Simply put, these investments have the potential—through nontraditional DOE-industry partnerships—to impact AI and computation like the impact observed in using reusable rockets for satellite communications and ultimately space travel. Consequently, urgent and immediate action is critical to capturing and extending the alignment of insights, community, infrastructure, and momentum created with the ECP program.

Leveraging Industry Advances to Extend U.S. Leadership

The incredible pace of innovation in AI is fueled by enormous investments by industry and nation states, primarily focused on applications central to industry and national security. Underlying techniques and methods, as well as infrastructure design and investment strategies from industry and defense applications will provide indispensable inputs for DOE mission areas in science, energy, and security. The same was true for exascale computing—industry and the work of other U.S. agencies provided important technologies and strategies. But without DOE leadership and sustained investment, today's most capable systems would be operating in China, Japan, and the European Union in support of *their* leadership in science, energy, and security.

Despite the rapid progress being made in industry and defense in the U.S. and other nations, many of the objectives central to DOE's mission are not being addressed by industry or defense activities. However, DOE's investments in exascale systems, large-scale data infrastructure, software, theory, and applications—combined with unique co-design approaches now scaled to thousands of experts—uniquely position the DOE complex's use of AI to extend its global leadership in science, energy, and security. New AI approaches, outlined in this report, can transform DOE's mission areas, particularly through the enabling capabilities of DOE's exascale and beyond computational infrastructure. But these new methods and resulting applications will not self-assemble through incremental progress—they demand a complex-wide, integrated initiative with a scale and vision that will impact every aspect of not only computational applications but the design, optimization, and even the assembly and operation of scientific instruments, user facilities, and both experimental and operational infrastructure.

Embracing Fundamental AI Approaches: Building Blocks

Six major AI "approaches" have emerged and solidified even during the three years since the initial DOE AI workshops were conducted [4]. Thus, in 2022 the DOE laboratories organized a second set of AI workshops, which examined this set of conceptual building blocks—each grounded in fundamental AI capabilities such as inference, optimization, and deep learning (Section 01). In domain areas spanning the DOE complex (Section 02), opportunities are identified to apply these approaches, challenges that must be overcome to do so, and specific advances that will be required. In turn, these Advanced Research Directions (ARDs) reveal crosscutting technology requirements in DOE's infrastructure and computational methods, such as scientific workflows and the data lifecycle (Section 03). Finally, in Section 04, we assess the readiness of the DOE complex—from hardware to the workforce—to implement the methodological, logistical, and cultural changes necessary to not only adopt new AI capabilities in support of the DOE mission but to provide the scientific leadership necessary to advance national competitiveness in science, energy, and security.

The scientific community created a comprehensive report following the 2019 workshops, laying out opportunities and challenges across 16 domains and technology areas comprising the DOE complex, from material science to complex engineered systems to mathematics and computer science [4]. However, the approaches addressed in that report were either in early formative stages or, in some cases, had not yet revealed the potential associated with scale. For instance, the confluence of advances in ML (particularly self-supervised, transfer, and deep learning) with extreme-scale data and enormous investments in computation time has only recently resulted in emergent capabilities in natural language processing (NLP) that reveal strategies for application in science and engineering. Such “foundation models” are “trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks” [5].

The state of AI methods in 2019 suggested that substantial gains would accrue through grassroots adoption and exploration across the many scientific, energy, security, engineering, and infrastructure facets of the DOE complex. Although this path to progress remains in place today, this kind of incremental investment and organic activity would also limit DOE—and by extension the nation’s science, energy, and security initiatives—to incremental advances at a time when other global leaders are investing in transformational AI agendas. Since 2019, early and entirely new capabilities associated with large-scale AI models represent an inflection point where there is opportunity for DOE to embrace pathfinding rather than adopting AI as followers. Simply put, *global leadership cannot be achieved through incremental nor solely grassroots progress*.

Seizing Opportunities; Addressing New Challenges

Success in this transformation will also exacerbate existing challenges. Harnessing the creativity and effectiveness of a truly diverse scientific workforce raises the bar on training multiple generations of researchers in AI methods. Many of today’s entry-level research functions may be eliminated to realize the top-to-bottom shift from traditional modeling and simulation to the AI approaches, including autonomous discovery and robotics (Chapter 05), along with associated reinventions in crosscutting areas such as those outlined in Section 03 of this report (data infrastructure, workflows, programming tools, etc.). Here, ethics must play an important proactive role in guiding workforce reinvention and be a central element in the formation and execution of the work outlined throughout this report—in contrast to a reactive or passive role. These challenges are discussed in Chapter 16.

Concurrent with these workforce challenges and rigorous data curation and associated tasks necessary for training AI models, new facets for consideration have emerged, such as new vulnerabilities related to training data—for example, the intentional or unintentional insertion of data that would undermine the correctness of the trained model. The need for new mathematics, theory, and foundational methods and approaches to data and model evaluation is highlighted throughout this report and emphasized in Chapter 12.

A Blueprint for Leadership

A *leadership strategy* for developing, advancing, and harnessing the potential power of the six AI approaches will require nothing less than a coordinated and comprehensive, sustained series of scientific, engineering, and infrastructure campaigns. This report specifies a blueprint for those campaigns, anchored in grand challenges that are central to the DOE mission. Certainly, industry progress will continue to be useful and relevant to DOE mission areas, but industry incentives are tied to market forces and business growth, and their data are often quite different in nature and content relative to DOE science, energy, and security data. Nevertheless, DOE AI initiatives must proceed in coordination with macrotrends in AI, many of which are industry-led and supported by increasing private sector investments. This approach fundamentally differs from the modeling and simulation technologies, methods, and related infrastructure that DOE has invented, invested in, and led for the past 50 years.

Five such AI macrotrends are reflected throughout this report:

1. A trend toward *larger-scale models, with new and emergent capabilities whose training requires computational resources that eclipse even the largest modeling and simulation efforts in the ECP.*
2. This training itself relies on *significant preparation and encoding of enormous multimodal data streams and sources.*
3. A shift from a 1:1 relationship between data and simulation models to a 1:n relationship where the resulting *AI model can be adapted and applied to many (“n”) modeling tasks*.
4. These trends introduce the need for *extensive and rigorous evaluation suites,* well beyond those necessary for current modeling and simulation projects.
5. The scale of software engineering and programming efforts to harness these trends is substantial and is balanced with the potential for *generalizable foundation models that can support large communities*—in contrast to the current modeling/simulation paradigm of many individual research teams creating bespoke models and evaluation suites.

The remaking of DOE’s science, energy, and security landscape with respect to computation, data, and experiments will of course create improved versions of contemporary modeling and simulation systems; but more importantly, it will result in a new class of applications that integrate AI capabilities in multiple steps. This migration will begin with hybrid applications (AI and traditional) and similarly mixed workflow tools, yielding to end-to-end replacements over the next several years. Moreover, advanced modern simulation and visualization tools such as “digital twins” established in engineering-based applications are rapidly moving into complex scientific-based domains [6].

The focus on AI and the need for a comprehensive revitalization of DOE’s scientific enterprise reflects the growing evidence that AI is intimately tied to the nation’s future and its role in the global order. This reality is woven throughout reports from the White House [7] and National Academies [8], as well as industry [9] and nongovernmental sources. Each of these and other reports convey a similar message, that AI is one of only several competitive areas that “tell the story of a nation (and its allies) coming perilously and unwittingly close to ceding the strategic technology landscape and along with it the capacity to shape the future” [10].

Table Intro-1 Summary of the AI for Science, Energy, and Security – Expected Outcomes.

01. AI and Surrogate Models for Scientific Computing	02. AI Foundation Models for Scientific Knowledge Discovery, Integration, and Synthesis
• Fusion Energy	• Stockpile Modernization
• Predictive Multiphysics Simulations	• Knowledge Distillation (unstructured to structured knowledge) and Hypothesis Formation
• Cosmology	• Digital Twins for Engineering Complex Scientific Domains
03. AI for Advanced Property Inference and Inverse Design	**04. AI-Based Design, Prediction, and Control of Complex Engineered Systems**
• Materials, Chemistry, and Biology Design (atomic / molecular scale)	• Hi-rep Rate Laser
• Engineered Structures / Systems (continuum scale)	• Accelerators
• Non-proliferation / Decision Superiority (process / protocols)	• Reactors (Fusion and Fission)
05. AI and Robotics for Autonomous Discovery	**06. AI for Programming and Software Engineering**
• Nuclear Weapons Design Transformation	• Adaption of Codes for New Computational Targets
• Accelerated Discovery in Materials, Chemistry, and Biology	• Discovering Quality Control Algorithms and Quality Control Optimization
• Advanced Manufacturing	• AI-Driven Co-design

Expected Outcomes

We highlight key expected science, energy, and security outcomes around the six AI approaches detailed in Section 01, which hold transformational potential both individually and in combination. More detailed descriptions of these and other expected outcomes are included throughout the report, particularly in Section 02: Domains.

Harnessing the five macrotrends comprising the leadership blueprint above will require a set of focused campaigns at the scale of the ECP project—hundreds of participants from DOE laboratories, universities, and industry, working together to co-design major instruments (in this case, exascale computers) and the software and applications that unlock the power of that instrument. Each of the six new AI paradigms described in Section 01 require organizing DOE data to build and train large-scale AI models targeting specific domain areas and involving model design and evaluation. The scale required for these campaigns is illustrated by industry efforts such as OpenAI’s development of GPT-4 [11] [12], whose initial training required months of dedicated time on an exascale platform. By late 2023, DOE’s Frontier and Aurora systems alone will provide nearly five times the computational capacity of OpenAI’s system.

Each campaign will demand an orchestrated team of hundreds of participants who will (a) assemble and prepare data from across the DOE complex, (b) strategically augment existing experimental data with data from current computational models, (c) design and train large-scale AI models—typically surrogates or foundation models—along with careful evaluation (e.g., validation, uncertainty quantification), and (d) develop and scale crosscutting capabilities (e.g., workflows, data and communications infrastructure) and methodologies (e.g., supporting explainability).

Here we provide highlights of the expected outcomes of these campaigns.

01. AI AND SURROGATE MODELS FOR SCIENTIFIC COMPUTING

Surrogate models, trained by the results of computational models, demonstrate orders-of-magnitude speedups over the originals. Conceptually, surrogate models represent a potential to effectively achieve zettascale performance on exascale systems by virtue of their simpler yet faithful representation of the full, complex system. Such performance is essential for AI systems that can rapidly explore a decision space or that can suggest (or actuate) decisions related to complex instruments or infrastructure, as we further outline in Chapter 04. Following are four exemplars illustrating the impact of surrogate models on the DOE mission.

Climate. Surrogates will enable a new type of AI-accelerated climate model, accelerating the model core and process physics to yield a speedup rate of at least three orders of magnitude. The surrogate model is trained by using a variety of methods, including full baseline model cases on exascale systems and by training the individual model components. A key capability that such a model would enable is that massive ensembles could be run in the same amount of time as a single model, providing the basis for uncertainty quantification in the climate simulation output. This capability has the potential to revolutionize climate predictions, improve our understanding of climate variation, and accelerate predictions of climate impacts on humans and critical infrastructure. The urgency for understanding climate impacts, devising adaptations, and evaluating mitigation strategies places high priority on this work.

Fusion Energy. Surrogates have enabled the introduction of an exciting kind of AI-accelerated fusion energy model, accelerating a validated global electromagnetic gyrokinetic code (GTC) to yield a speedup of over 5000x. This synthetic gyrokinetic surrogate model, SGTC, is trained using AI/ML methods, including full-baseline deep-learning approaches, with training carried out on current leadership-class systems, such as Summit at Oak Ridge National Laboratory and Polaris at Argonne National Laboratory. A key capability that this new model has demonstrated is carrying out massive numbers of experimentally validated cases run in the same amount of time as a single model, providing the basis for real-time output. This approach has the potential to revolutionize real-time predictions in magnetic fusion energy, accelerating progress toward favorably modifying the plasma state to a more benign thermodynamic state. The urgency of devising such mitigation strategies for dangerous disruptive events in thermonuclear burning plasmas such as the International Thermonuclear Experimental Reactor (ITER) places high priority on intensive future validation studies of this kind [13].

Predictive Multiphysics Simulations. AI capabilities for bridging temporal and length scales in multiphysics simulations hold the promise of qualitative leaps in our ability to predict and design complex physical systems. These capabilities will employ a spectrum of AI methods that includes optimization of solvers and other fine-grained elements of simulations, efficient learned representation of cluttered data with sparse true information content, and accurate surrogates for coupled partial differential equations. Without this contribution, progress in our ability to simulate physical systems for science and national security will stall because advances in processor technologies can no longer keep pace with the computational cost of increasing simulation fidelity. For example, a factor-of-ten improvement in resolution in present three-dimensional (3D) simulations (which is still far below that needed for bridging to the mesoscale) would require a computer that is ten thousand times more powerful than exists today. Advances in simulation fidelity permitting an ability to predict from mesoscale phenomena to macroscopic performance will lead to a new generation of engineered physical systems for energy, science, and national security.

Cosmology. The U.S. has invested heavily in cosmological surveys leading to discoveries that have unearthed some of the deepest mysteries in fundamental science. Because cosmology is an observational science, detailed simulations are used as forward models to understand and interpret large-scale datasets from sky surveys that cover wavebands from the radio to gamma rays. These simulations are some of the largest applications run on state-of-the-art high-performance computing (HPC) systems. Under the ECP, the simulation capability for DOE-relevant cosmological missions has been significantly enhanced. Even so, using a single large-scale simulation to directly analyze data is cost prohibitive, as it may require thousands or even millions of individual runs. Precision surrogate models for summary statistics that correspond to cosmological observables were pioneered by DOE scientists who succeeded in reducing the time by many orders of magnitude (more than a billion), thereby allowing more powerful methods of data analysis to be used. As computational power grows, it will be possible to develop powerful, effective AI-based surrogate models (digital twins) for individual simulations in analogy to the case of climate science.

02. AI FOUNDATION MODELS FOR SCIENTIFIC KNOWLEDGE DISCOVERY, INTEGRATION, AND SYNTHESIS

The recognition that exascale computing capabilities would be critically important for training AI models was well understood in 2019, but results from recent industry investments in computational resources, along with access to vast multimodal data sources, point to even greater opportunities. These industry advances have demonstrated powerful and, in some cases, new emergent (unplanned and unexpected) capabilities, such as the ability of very large models to adapt to new tasks despite relatively sparse additional training data. Leveraging DOE's investments in exascale systems, along with enormous and growing volumes of scientific data, foundation models have the potential to be trained for application on broad classes of problems relevant to equally large scientific communities. This opens the possibility for large-scale, community-built foundation models—including digital twins—that, in analogous fashion to scientific instruments, enable entire communities to perform computational experiments without the need to create bespoke infrastructure. Examples of the impact of developing foundation models for DOE science, energy, and security include the following areas:

Stockpile Modernization. The NNSA's nuclear deterrence mission requires rigorous analysis of the design, manufacturing, and surveillance of components and systems. When a component or system failure is identified, a significant findings report is created, and an extensive root cause analysis is performed. This is an extremely time-intensive process requiring person-months of effort searching through design, manufacturing, testing, and qualification documents and data. AI transformer and foundation model methods have shown tremendous advances in automatically identifying patterns in natural language, understanding relationships, and summarizing text. An urgent need exists amongst the NNSA and the DOE labs to extend these methods to specifically target the domains of math, science, and weapons system design. The current human-intensive processes in the NNSA life extension programs (LEPs) carry inherent risk of schedule slips. The ability to automatically digest technical documents, create summaries, and perform root-cause analysis could save the NNSA laboratory staff from spending enormous amounts of time on manually searching through documents for relevant information regarding significant findings. This same technology could be used in DOE Office of Science (SC) research facilities to perform literature searches, find similarities in reports, and summarize information.

Knowledge Distillation (unstructured to structured knowledge) and Hypothesis Formation. Scientists and engineers at DOE laboratories seek to understand phenomena for which explanatory theories are lacking or inadequate (e.g., how clouds affect climate, or how vortices evolve in fusion plasmas), to solve engineering problems (e.g., an energy storage mechanism that can store 10x more energy for one-tenth of the price), or, frequently, to do both at once. Regardless of the specific problem being studied, a frequent challenge is the vast amount of existing knowledge that could potentially be relevant to its solution—a quantity that typically far exceeds the cognitive capacity of any one individual or even team. The recent and considerable successes achieved with large language models suggest that a transformative solution may be on the horizon. Most of the current “knowledge” is recorded, often implicitly, in unstructured forms, whether text (e.g., published scientific papers, technical reports, unpublished documents, lab notebooks, outputs from computations) or other digital formats (e.g., images, videos, simulation outputs). Such unstructured data contain vast amounts of information about what experiments and computations have been performed (whether successfully or unsuccessfully) and the results that were obtained, and also about the inferences made, hypotheses generated, and conclusions formed by human experts from different disciplines and backgrounds. A large language model trained on large corpora of this unstructured knowledge, particularly one that incorporates knowledge about the physical world, may well be able both to distill succinct structured representations of extant knowledge (e.g., by extracting every recorded property of a specific material or structure from millions of documents) and to generate hypotheses concerning previously unobserved relationships (e.g., by observing that a certain phenomenon has been observed only under specific unusual circumstances). Such a model would be an invaluable aid for DOE researchers working in a wide range of domains.

Digital Twins for Engineering and Complex Scientific Domains. Digital twins established in engineering-based applications are rapidly moving into complex scientific-based domains [6]. DOE is undergoing a digital transformation initiative to support digital engineering within NNSA and SC. Transforming the nuclear deterrent (ND) lifecycle from a test-based process to an integrated test–modeling and simulation (ModSim)-based process presents challenges due to the complex workflows. These workflows span desktop to HPC computational resources; design to environmental specification to rigorous qualification to surveillance activities; and multiple disciplines such as electrical, electromagnetics, mechanical, thermal, and various combinations of these. Ultimately, the goal of these workflows is to build the knowledgebase necessary to support a risk assessment. Much of this kind of risk-based assessment is asserted with expert judgment and experience. AI can be used to automate these workflows and formulate the risk assessment. Digital twins of components and systems will enable shortened design, testing, qualification, and surveillance life cycles. New AI methods are needed for: (1) model preparation and setup—such as computer-aided design (CAD) geometry cleanup and simplification for meshing, interpretation of design intent already within the CAD assembly design, material model choice and uncertainty-informed attribution, etc.; (2) model design and tuning: generating reduced-order models, solver choices, solver settings, etc.; (3) simulation: setting up and executing robust ensembles, quantities of interest extraction, optimization, etc.; (4) model credibility; (5) data: needed to form the technical basis; (6) risk assessment approach(es): to assert certification from the component level through to the system level; and (7) convolving subject matter expert judgment and formal uncertainty quantification methods.

03. AI FOR ADVANCED PROPERTY INFERENCE AND INVERSE DESIGN

A third AI method leverages the application of AI models to property inference and inverse design problems. The former involves the use of AI models to predict the physical properties of a system given its design, and the latter entails models that enable scientists to determine a system design that has specified desired properties. For example, in chemistry, a property inference model might be used to predict the melting point or toxicity of a molecule, given its chemical composition, while an inverse design tool would be used to identify molecules that are liquid at room temperature, have a high heat capacity, and are not toxic. Three outcomes illustrate the promise of property inference and inverse design.

Materials, Chemistry, and Biology Design (atomic / molecular scale). Critical DOE science, energy, and security missions depend on the discovery and development of new molecules for structural materials, biological therapeutics, energetic materials, and far more. Using a combination of HPC, AI, and experimental expertise, DOE will develop a molecular discovery engine that can build critical therapeutics—either small molecules or proteins—on bold timescales relevant to emergent biothreats, for example, several weeks from threat assessment to molecule design to deployment. The molecular discovery engine for bioresponse will be composed of high-performance, AI-steered processes that predict and optimize multiple properties of the molecule–efficacy on target, safety in humans, pharmacokinetics, and manufacturability. The predictive models will provide validated measures of uncertainty and will integrate with automated chemical synthesis and experimental systems to steer the design optimization process and validate the resulting molecular designs. Early efforts to apply AI-driven models to components of the drug development pipeline have been quite successful. Growing access to the required data and models, coupled with improving AI-based and mechanistic computational models and automated precision measurement technologies have primed these applications for new capabilities. Inaction will leave the U.S. vulnerable to future pandemics and global pharmaceutical technoeconomic competition. These capabilities will accelerate and strengthen our response to biomedical threats to national security and to our economy. They will enable early threat assessment and guide the pre-positioning of data, models, and molecules to enable earlier starts for response. They will shorten development and validation timelines and potentially reduce cases, deaths, and economic impact. The capability is generally applicable to therapeutic development and will accelerate new medicines in areas such as cancer and neurological disease. However, such a molecular discovery engine has far more uses within DOE and the nation. With modification, the engine could be tuned to deliver new structural polymers, molecules for high explosives, custom metallic alloys for critical applications, and many more general material science applications. The impacts would be substantial, not only for science and security, but also for general U.S. economic and manufacturing competitiveness.

Engineered Structures / Systems (continuum scale). The use of AI-enabled property inference and inverse design to create and optimize engineered systems and structures represents an opportunity for unprecedented integration across scales, from materials to components to entire complex engineered systems—and ranging from energy storage materials to entire distribution systems or from hardened electronics to weapons systems. Using AI for inverse engineering will result in “born qualified” manufacturable material components specifically tailored for precise design specifications. The impact of this integration will encompass time and other costs, as well as the safety, reliability, and sustainability of these materials, components, subsystems, and systems. For example, the complexity of the nation’s energy systems, combined with increased prevalence of extreme weather, confounds our ability to design and operate such systems reliably and cost effectively. Advances in AI will enable us to replace today’s large simulation models with inference from AI models and support real-time decision and control through inverse design and optimization processes, while also integrating inherently multimodal, heterogeneous, and rapidly growing data from the energy infrastructure into (global) energy infrastructure models with high fidelity to provide trustworthy predictions.

Non-proliferation / Decision Superiority (process / protocols). AI-based advanced property inference methods will revolutionize capabilities for detecting, analyzing, and strategically acting on potential proliferation activities; AI-enhanced text and graph analytics tools will significantly improve our ability to locate proliferation information and identify disinformation, as well as to identify loosely coupled organizations involved in nuclear proliferation, disinformation, and weapons development. Similarly, AI models trained on data from facilities monitoring and remote sensing will strengthen our ability to rapidly detect anomalies and patterns of life analysis across multiple sources of information, including rare event detection characterizing rogue actors (nation states or non-nation states). Simply put, new AI methods as outlined in this report will lead to game-changing acceleration of data-driven, repeatable, and reliable decision-making with archived pedigree, and a reduction in the time and resources needed for situational analysis buttressed by greater confidence and analytical rigor.

04. AI-BASED DESIGN, PREDICTION, AND CONTROL OF COMPLEX ENGINEERED SYSTEMS

The use of AI for prediction and control of complex engineered systems—ranging from energy distribution grids to scientific instruments and user facilities to fusion energy systems—requires faster-than-real-time modeling. In some cases, timescales that demand highly local decisions given even the small latencies involved in sending data across a campus (or building) are also required. The concept of digital twins captures key capabilities necessary for the use of AI here—complex, multimodal representations of complex systems that operate faster-than-system-time (e.g., leveraging surrogates) to support decision-support models, and that continuously learn from operational and model data. Effective use of AI in the design and control of complex engineered systems will revolutionize science and engineering, including significantly improving the efficiency, reliability, and robustness of the energy infrastructure in the U.S.; enabling next-generation facilities and autonomous laboratories for scientific advancement; reshaping the manufacturing infrastructure to improve competitiveness; and developing advanced systems for national defense. The development and use of AI capabilities of this nature will have impacts as detailed below.

High-rep Rate Laser. DOE experimental facilities—from giant lasers to light sources to robotic chemistry systems—will greatly increase both the volume and quality of mission-critical data obtained in experiments by developing and using AI systems to simultaneously predict experiment performance and control the experimental facility. Data from laser experiments, images from accelerator light sources, and semi-autonomous manufacturing systems are all critical for the NNSA stockpile stewardship mission, but also for fundamental science that is core to the entire DOE mission. Across DOE, current experiment setup uses large-scale simulation to predict the experimental conditions and to optimally configure diagnostics to observe those conditions. The execution of these experiments and the following analysis is currently limited by using slow, conventional control systems and high-latency remote access to HPC. AI control models that combine simulation-based knowledge of experimental conditions with an ability to command AI-ready diagnostics will free scientists to explore system behavior far faster and more thoroughly than with rate-limiting traditional methods. These semi-autonomous, or “self-driving,” facilities would allow subject matter experts to explore new scientific territory with unprecedented speed. Self-driving operations based on AI control models would revolutionize experimental science across an expansive array of applications. The repetition rate and quality of design discovery on laser facilities, from the scale of the National Ignition Facility (NIF) to university lasers, would increase dramatically. Advanced accelerator systems, such as the currently upgrading SLAC National Accelerator Laboratory’s Linac Coherent Light Source (LCLS), will see dramatic throughput gains from new, rapid self-tuning and new hyper-capable laser probes for helping break new ground in High Energy Density (HED) experiments aimed at increasing our fundamental understanding of materials at extreme pressures and temperatures. Other beneficiaries include self-driving robotic chemistry systems for accelerated material science and drug discovery, advanced manufacturing platforms able to deliver real-time corrections for manufacturing errors, and comprehensive metrology systems that can analyze critical parts with increased fidelity and speed. AI-driven acceleration in these key experimental systems will shorten the time to solution for the stockpile mission, advance the pace of fundamental scientific discovery, and continue to position DOE as the premier attractor of talent in applied and fundamental science. New capabilities in AI are proliferating at the same time that computing is advancing to edge/control systems, providing a potent combination to automate experimental configurations on timescales of microseconds versus human timescales. The combination of deep expertise in HPC and large-scale experiments has positioned DOE to take an early lead in these transformations to highly automated experimental facilities. To wait would be to cede expertise here to Europe or China, with Europe already ahead in the use of small-scale, high-rep lasers and China coming on strong in both manufacturing and fundamental science.

Accelerators. Particle accelerators and accelerator-based photon sources are key components of scientific discovery and are used in applications across industry, national security, and medicine. Both extending the capabilities and reducing the size and cost of accelerators are important to progress in many areas of science, including for better understanding the structure of the universe through high-energy particle physics; creating brilliant photon sources for basic energy sciences, materials, and industry; and exploring new states of matter. This requires ever more complex and precise systems for which AI/ML methods are starting to be applied in design, deployment, and operation, and will become ever more critical as AI/ML methods—and accelerator performance demands—advance. An especially crucial expected impact of AI/ML algorithms is their use in the control and parameters tuning for accelerators in real time, during operation, to maximize performance. Major challenges ideally suited to new AI/ML approaches include the ability to precisely control the properties of accelerated beams, which are a function of many device components and environmental fluctuations. Switching between different experiments—requiring large changes in beam properties—presents additional challenges and can require hours of hands-on tuning. Although first-principles modeling based on the multitude of component settings is computationally intractable in many cases [13], promising initial results have been observed by developing customized AI/ML methods that automatically compensate for unknown time-varying changes to accelerator components (such as magnets, and to unknown changes in the accelerator’s input beam distribution) [14]. In a related aspect of using AI/ML for improving the efficient operation of accelerators, early results have exploited classification and anomaly detection algorithms, with the aim of preventing accelerator damage or beam loss in the case of abnormal operation. For example, ML techniques have recently been applied to the early detection and classification of quench precursors in superconducting magnets, where conditions can build up to a circumstance where the magnetic field is suddenly lost. And, they have reduced orbit deviations in a synchrotron light source by an order of magnitude [15]. Looking to the future, a combination of effective AI/ML models and fast feedback control look to hold the keys to new generations of accelerators, for example, current work has enabled a new generation of efficient laser drivers for accelerators by combining more than eighty fibers all controlled to a fraction of the wavelength of light [16]. Development and implementation of even more accurate models will routinely be important to a broad range of future accelerators, from extracting the maximum intensity, to developing new and more compact accelerators based on laser driven plasmas, to future particle colliders.

Reactors (Fusion and Fission). High Energy Density Physics (HEDP) and fusion physics rely on multiphysics codes that model radiation-magnetohydrodynamics (radMHD) and density functional theory (DFT) calculations. These are computationally expensive calculations that display low-dimensional emergent behavior. There are also expensive experiments with multiple diagnostic measurements that are designed to test and calibrate the physical models. This calibration underscores a critical need for methods that can construct efficient, high-fidelity surrogates of the physics; identify the low-dimensional sub-manifold structure of the modeled physics and the data (reduced-order model or topology); and finally assimilate the data with the model to refine and extend the estimate of the sub-manifold structure. Simply put, the physics needs to incorporate its deep learning from the multiphysics codes and the experimental data. Solving this problem will have major impacts on the understanding, uncertainty quantification, and validation and verification of HEDP, inertial confinement fusion, magneto-inertial fusion, magnetic confined fusion, and the stockpile. This approach could also be applied to a broad range of other physical problems such as climate physics, geophysics, and astrophysics. For magneto-inertial fusion in particular, it would enable new designs and reduce the risk of designs not performing both at current scale and future scales. It would also enable a much-improved experimental design to understand the physics (hypothesis test) and to reduce the risk. Such an advance could lead to a commercial fusion energy breakthrough and a more reliable stockpile.

05. AI AND ROBOTICS FOR AUTONOMOUS DISCOVERY

The use of AI for automating discovery in laboratory and other processes, including advances in robotics, will leverage property inference and inverse design to improve each step of discovery processes, bringing AI models to bear on designs ranging from energy storage to explosives to disease treatments. Combining these with AI-enabled robotics, guided by self-learning digital twins, DOE has the opportunity to fully integrate AI computation, data, and instruments in laboratories and user facilities—including multi-instrument laboratory workflows. Additional impacts of autonomous discovery and robotics are described below.

Nuclear Weapons Design Transformation. On average, a major NNSA Alteration or Life Extension Program (LEP) typically runs 3.5 years behind its initial baseline schedule. In no small part, the first developmental half of the product development lifecycle tends to be full of requirements, architecture, design, qualification, cost, schedule, and design for manufacturing/surveillance iterations—each iteration requiring a few months of re-baselining by the core weapons system realization teams. AI methods described in this report enable models to be trained using the entire historical and current nuclear weapons data corpus (e.g., detailed design, requirements, architecture, qualification, production, surveillance, and formal and informal programmatic and technical information both labeled and unlabeled). The resulting model will propose a detailed weapons system design given a new set of requirements / architecture / funding / schedule constraint (test) data. Subject matter experts will then integrate the proposed detailed concept and leverage the high-fidelity concept to rapidly re-baseline the weapons system's detailed design. This design cycle acceleration could reduce staffing, time, cost, and scope, affording the nuclear security enterprise a high probability of successfully executing simultaneous modernization programs on time and on budget.

Accelerated Discovery in Materials, Chemistry, and Biology. All possible natural and synthetic materials are formed from 3D atomic configurations of just a few dozen different chemical elements. Ab initio computational methods can accurately predict diverse properties at the nanoscale but not on the vastly larger meso- to macro-scales on which critical performance and processing behaviors emerge (e.g., photovoltaics, metal alloy glasses, multiferroics, memristors). An AI/ML workflow that can leverage exabytes of ab initio data at the nanoscale would address this challenge, producing quantitatively predictive simulations of material synthesis processes and resultant performance properties. An AI-enabled workflow leveraging, for instance, new learning techniques would be sufficiently fast and accurate on exascale platforms to allow exploration and exploitation of vast combinatorial spaces of chemical composition and processing conditions on a timescale of days to weeks rather than over many months. This acceleration would provide real-time guidance for experimental design campaigns, where many of the nation's most urgent security challenges are attributable to limitations in current materials. Such areas include, for instance, carbon-free nuclear fusion energy production using new materials that can withstand hot plasma conditions; solar power production through advances in materials for photon capture and energy storage; nuclear deterrence, military, and space exploration advances enabled with reliable high-performance materials for extreme environments; and similar advances in computation, transport, and medicine.

Advanced Manufacturing. Direct-digital additive manufacturing (AM) platforms, while attractive from a design flexibility standpoint, are still plagued by the inability to achieve process and parts qualification for high-consequence applications. This is particularly true for metal powder-feedstock-based AM modalities, such as laser-powder bed fusion and directed energy approaches. Variabilities in powder feedstocks, the stochastic nature of laser-melt-pool-plasma interactions, heterogeneous polycrystal grain structures from solidification, and systematic and random defects in the form of porosity and distortion underpin the difficulties of achieving material and process qualification. Automating the process inputs to parts/performance integration requires data-driven computational intelligence that addresses all these stochastic variabilities. AI methods including inverse design will support the creation of workflows that can traverse the digital thread from model-based design through build and final part inspection. Deep learning capabilities will further reduce risk associated with process-variation through models that are trained on at-line and on-line sensor data and process and performance models. This advance will provide routine qualification successes and more effective application of techniques, such as powder metal AM to enable disruptive part designs, unique materials, and form factors for high-consequence national security systems in DOE SC and NNSA that cannot be produced with conventional approaches. Moreover, the use of these and other AI enablers could reduce the typical 10-year timelines associated with the insertion of new (and certified) metal AM parts into NNSA systems, in turn improving modularity and agility. Ultimately, the AM enabled by these, and other AI methods is the only way to achieve a cycle-reduction that impacts future programs and new systems. Success will also allow optimized in-situ monitoring and post-build inspection to minimize cost while maintaining product confidence.

06. AI FOR PROGRAMMING AND SOFTWARE ENGINEERING

Throughout the DOE complex and underpinning every scientific and operational process are increasingly complex software systems. The growth in scale and complexity of these systems, combined with their roles in critical systems such as instrument or energy infrastructure control, has been a recognized challenge for several decades. This situation has been particularly emphasized, given that the networked nature of these critical systems also exposes them to cybersecurity risks. Tremendous progress has been made in the use of AI to assist programmers and even to develop programs. Impacts expected through the use of AI for programming and software engineering include those listed below.

Adaptation of Codes for New Computational Targets. AI for Programming and Software Engineering promises to change how we adapt codes for new computational targets enabling the nation to answer some of our most pressing science, energy, and security questions in weeks rather than years. Using large-scale master models that are trained on both the corpus of general-purpose programming and optimization techniques alongside the wealth of DOE science and engineering algorithms, we will develop automated aids allowing computational scientists to rapidly implement and evaluate these methods for a scientific problem. These master models will also be trained with high-performance implementations of algorithms on a variety of hardware technologies and will conduct performance and robustness evaluation using an active learning approach. AI has demonstrated massive speedups in code development. What currently takes large teams of developers and scientists years to complete might be accomplished in months or less in the future. This will address what has become a concerning dilemma—the pace of hardware specialization has become faster than the ability of human programmers to adapt to the advances offered by industry. Innovation in everything from materials science to designing complex engineered systems will be improved by our ability to map applications to quickly changing hardware.

Discovering Quality Control Algorithms and Quality Control Optimization. Existing digital controls and systems for high-consequence applications, some dating back generations, are vulnerable to natural faults and adversarial attack. This applies to a wide variety of software systems that ensure the safety of the nuclear enterprise (nuclear weapons and its infrastructure) and energy systems (nuclear reactors, dams, oil refineries/pipelines, and electrical generation / transmission). Discovering faults and vulnerabilities in control and system software governing these applications before they are exercised is critical. Performing the usual by-hand assessments is not tractable because the type, quantity, and diversity of these installations are vast. What is needed is a systematic and automatic methodology for discovering faults and vulnerabilities in black box digital systems. The ability to quickly diagnose issues with high-consequence controls in the nation’s nuclear deterrent and energy infrastructure is important today and will likely increase in the future. An automated way to reconstruct (learn) a digital system and then use it in a formal analysis to check safety, security, and reliability properties would go a long way to securing the nation’s assets. A robust AI model learning and proof-finding assistant could dramatically change the scope and applicability of formal verification for national defense systems, which is currently limited by the time required for human-driven proof search. By effectively automating the process of complex formal verification, this research would enable DOE and NNSA to verify more requirements for digital national defense systems and the energy infrastructure, and more complex properties, potentially including cybersecurity properties. This advance would reduce overall design time while increasing assurance that the resulting systems are safe and reliable. The past few years have seen a number of academic and industry efforts working on how to apply recent advances in deep learning to formal methods-based verification in general and proof search in particular [17][18][19][20]. These efforts have demonstrated the feasibility of the approach, but it remains to be shown that it can translate to success in practice for real-world problems such as those faced by NNSA.

AI-driven Co-design. The exponentially growing demand for computing and the end of Moore’s Law in microelectronics have resulted in an urgent call for microelectronics-compute co-design, in which each level of the “stack” co-evolves, sometimes radically. The co-design knowledge space, however, is enormous, spanning materials to algorithms over vastly different conceptual scales. Future HPC systems will benefit from leveraging a far more heterogeneous assortment of microelectronics technologies than today's systems have, but achieving this increased diversity, which may include conventional complementary metal-oxide semiconductor (CMOS) accelerators, analog computing, quantum computing, and neuromorphic computing, presents several implementation challenges. DOE requires an ability to design such heterogeneous computing systems effectively with U.S. industry partners and needs the ability to jointly configure systems based on application requirements while tailoring the applications to these systems. For a given computational application (or assortment of applications), this heterogeneous design should be able to effectively identify a desirable customized combination of these computing technologies that implements desired functions while maximizing the advantages of each technology for maximal time, space, and energy efficiencies. Further, this approach should be able to forecast how these emerging technologies will evolve to make solutions flexible going forward. An AI system to solve this task would have to explore a very large combinatorial space of interactions between potential components, with the main data for this effort being simulation and benchmark data from different existing and proposed microelectronics platforms. The AI methods required for this problem could leverage recent advances in reinforcement learning (which has been used for optimizing circuit design, but not yet full computing systems), adaptive AI frameworks, model-based learning, and stochastic AI methods that make complex strategic decisions from a large search space. Solving this problem will provide a significant step forward in maintaining U.S. leadership in microelectronics technologies and will help reduce the energy requirements of computing systems. The growing challenges of improving conventional computing technology present a risk to U.S. leadership in microelectronics, which poses significant economic and national security challenges. By maximally leveraging these emerging computing technologies, the U.S. and DOE have an opportunity to extend their leadership in HPC technologies overall and achieve more impactful capabilities in computing for science, energy, and national security while meeting energy efficiency and cost requirements.

References

[1] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., and Chi, E.H., 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682.*

[2] Bender, E.M., Gebru, T., McMillan-Major, A., and Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March, pp. 610–623.

[3] Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., et al. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* 2 (4), 100179. DOI: https://doi.org/10.1016/j.xinn.2021.100179

[4] U.S. Department of Energy-Office of Science, 2020. *AI For Science: Report on the Department of Energy (DOE) town halls on artificial intelligence (AI) for science*, Stevens, R., Taylor, V., Nichols, J., Maccabe, A. B., Yelick, K., and Brown, D. (eds.), Feb. https://publications.anl.gov/anlpubs/2020/03/158802.pdf and https://doi.org/10.2172/1604756, accessed November 30, 2022.

[5] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., and Brynjolfsson, E., 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[6] Kapteyn, M., Pretorius, J., and Willcox, K., 2022. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, Jan. 31 (special one-year anniversary collection).

[7] National Security Commission on Artificial Intelligence, 2021. *Final Report*, October. https://www.nscai.gov/2021-final-report, accessed December 16, 2022.

[8] National Academies, 2022. Machine Learning and Artificial Intelligence to Advance Earth System Science: Opportunities and Challenges – A Workshop, February. https://www.nationalacademies.org/our-work/machine-learning-and-artificial-intelligence-to-advance-earth-system-science-opportunities-and-challenges---a-workshop, accessed December 16, 2022.

[9] Artificial Intelligence and Business Strategy, 2022. *MIT Sloan Management Review*. https://sloanreview.mit.edu/tag/artificial-intelligence-business-strategy/, accessed December 16, 2022.

[10] Schmidt, E., Schawlow, N., Work, R.O., Thornberry III, W., and Flournoy, M., 2022. Mid-decade challenges to national competitiveness. *Special Competitive Studies Project (SCSP)*, September.

[11] Park, Y.J., Kaplan, D., Ren, Z., Hsu, C.-W., Li, C., Xu, H., Li, S., and Li, J., 2023. Can ChatGPT be used to generate scientific hypotheses?, arXiv:2304.12208.

[12] OpenAI, 2023. GPT-4 System Card, March 23. https://cdn.openai.com/papers/gpt-4-system-card.pdf, accessed May 12, 2023.

[13] Dong, G., et al., 2021. Deep learning-based surrogate Model for first-principles global simulations of fusion plasmas. *Nuclear Fusion* 61, 126061.

[14] Scheinker, A., Cropp, F., Paiagua, S., et al., 2021. An adaptive approach to machine learning for compact particle accelerators. *Sci. Rep.* 11, 19187. https://doi.org/10.1038/s41598-021-98785-0

[15] Leeman, S., et al., 2019. Demonstration of machine learning-based model-independent stabilization of source properties in synchrotron light sources. *Phys. Rev. Lett.* 123, 194801. https://doi.org/10.1103/PhysRevLett.123.194801

[16] Wang, D., Du, Q., et al., 2021. Stabilization of the 81-channel coherent beam combination using machine learning. *Optics Express* 29 (4), pp. 5694–5709.

[17] Tactician, undated. A seamless, interactive tactic learner and prover for Coq. https://coq-tactician.github.io/, accessed February 13, 2023.

[18] Crouse, M., 2021. A deep reinforcement learning approach to first-order logic theorem proving. *AAAI*.

[19] Loos, S., Irving, G., Szegedy, C., and Kaliszyk, C., 2017. Deep network guided proof search. https://arxiv.org/abs/1701.06972.

[20] Bansal, K., Loos, S.M., Rabe, M.N., Szegedy, C., and Wilcox, S., 2019. HOList: An Environment for Machine Learning of Higher-Order Theorem Proving. https://arxiv.org/abs/1904.03241.

SECTION 01: AI APPROACHES

This section details six new AI-empowered computing paradigms, or AI Approaches. These approaches form a set of building blocks combining and scaling fundamental AI functions, such as inference from large-scale and often unstructured and multi-modal data sources, natural language processing, and object recognition. These building blocks create transformational capabilities, from surrogate and foundation models to digital twins to automate real-time control of instruments, experiments, or complex infrastructure; inverse design systems and ultimately autonomous experiments, laboratories, and instruments; and automated software engineering and programming. Making this report particularly timely are relatively recent discoveries of emergent capabilities that represent new classes of AI models, including foundation models and physics-informed ML surrogate models.

01. AI AND SURROGATE MODELS FOR SCIENTIFIC COMPUTING

The U.S. Department of Energy (DOE) has been a world leader in scientific computing for decades. DOE's use of scientific computing has helped the nation meet many mission challenges, advancing the state of the art in science, engineering, energy, and national and global security. Growing computing power has enabled increased complexity and fidelity in simulations and their expansion into new scientific frontiers. However, the computational cost to capture these details has grown to consume the largest supercomputing resources. While these full-scale simulations lead to important discoveries and enhanced understanding, only a limited number are possible given that they require the use of entire machines. This high computing cost significantly limits the questions we can ask and the science we can do.

Advances in science and engineering require extensive use of "many-query applications" (e.g., parameter sweeps, inverse problems for parameter estimation, and model-based design optimization). These applications require multiple computationally expensive model invocations, often called sequentially rather than concurrently. This demands many simulations of a model in rapid succession. Simply put, while "hero" simulations are good demonstrations of results of many-simulation efforts, but they are often insufficient to drive large-scale scientific advancement, complex systems control, and autonomous science.

Artificial intelligence (AI) and machine learning (ML) have demonstrated the ability to create accurate, fast-running *surrogate models* for computationally expensive simulations. Using a limited number of evaluations of the simulation, AI/ML methods learn to accurately predict the output for new scenarios with quantification of the prediction uncertainty, allowing researchers to get an accurate approximation of the full simulation in a fraction of the time. Early work by many groups has demonstrated the enormous potential of using these methods to accelerate scientific computing applications. Groups have demonstrated speedups from 100 times to over one billion times in diverse applications, such as Density Functional Theory (DFT) simulations of electronic structure, molecular dynamics of protein complexes, cosmology, earthquakes, and computational fluid dynamics. AI-based surrogate models have unlocked new frontiers in prediction and are trained with complex, diverse data structures such as images, text, and networks. This has catalyzed new opportunities for scientific impact for DOE computing capabilities.

The increase in speed, which will result in evaluations in fractions of a second instead of in days, is critical to leveraging DOE's world-class computing capabilities to meet grand challenges. The utility of surrogate models is best highlighted by the answer to the question: "*What could happen with world-class simulations if they could be evaluated in fractions of a second instead of days or weeks?*" This chapter outlines how surrogate models will be a key component in the integration of AI into DOE scientific and engineering workflows. These workflows may range from (a) learning control laws and providing data augmentation for AI-enhanced real-time controllers of engineered systems to (b) embedding surrogates in real-time monitoring, forecasting, and data assimilation of digital twins of complex systems to (c) integrating a hierarchy of surrogates as "closures" or "constitutive models" in multi-scale, full-system simulations to represent unresolved physical processes. Such AI/ML applications will ensure that the highest-quality information generated by high-fidelity scientific simulations can be shared transparently across scales in practical engineering simulations. Advancing our ability to connect surrogates of complex data from simulations to similar data structures generated by experimental diagnostics will be critical for allowing surrogate-simulator-AI systems to be integrated into experimental workflows to meet autonomous science goals.

This chapter addresses opportunities across the DOE mission space for acceleration; the needs for future computer architectures to support AI-accelerated, high-performance computing (HPC); and needed advances in applied mathematics, algorithms, AI, and software frameworks.

PROJECT SPOTLIGHT

Project Name: Black-box optimization for scientific machine learning models

PI: Guannan Zhang

Organizations Involved: Oak Ridge National Laboratory

Goal: Develop black-box optimization methods for inverse problems that involves non-automatically differentiable simulators.

Significant Accomplishment: Application of our surrogate-based black-box optimization method to calibrate a constitutive material model (mercury) for a neutron target.

In the News: Radaideh, M., Tran, H., Lin, L., Jiang, H., Winder, D., Gorti, S., Zhang, G., Mach, J., and Cousineau, S., 2022. Model calibration of the liquid mercury spallation target using evolutionary neural networks and sparse polynomial expansions, *Nuclear Instruments and Methods in Physics Research B*, 525(15), pp. 41–54.

1.1 State of the Art

Surrogate models are data-driven, AI/ML-based approximations of physical, chemical, or biological processes, trained on measured data and/or data generated by executing high-fidelity (and computationally expensive) simulation models. Although AI-based surrogates can be used for many purposes (e.g., to discover unknown constitutive laws), the most straightforward use of AI-based surrogates for HPC is to serve as fast-running, accurate proxies of computationally expensive, high-fidelity models. These proxies can then be used in many-query applications (e.g., design optimization, uncertainty quantification, real-time control, and digital twins), without compromising on the fidelity of the calculations. Using AI-based surrogates enhances the reliability and realism of decision-making applications. Many have been successfully applied to various physical simulations. For example, Figure 1-1 shows examples of an open-source code for AI-based surrogates [1], applied to many different physics codes to accurately accelerate physical simulations.

The different types of AI-based surrogates can be categorized by their incorporation of physics constraints and their interpretability (i.e., by their sophistication and physical realism). Surrogate models with no physics-based constraints**,** typically known as *black box models***,** do not explicitly incorporate the underlying governing equations but instead relate input and output (I/O) data directly using statistical or machine-learned relationships. Some examples are Gaussian process or neural network models. On the other hand, *physics-informed surrogates* take advantage of both known governing equations and data. Between the black box and physics-informed surrogates, *interpretable surrogates* have been developed, in which I/O data are related by well-understood forms, such as differential equations, or explainable AI algorithms, such as linear models.

Black Box Models. Examples of *the black box approach* include the Gaussian process [2, 3, 4, 5, 6], radial basis functions [7, 8], Kriging [9, 10], and neural networks [11, 12, 13]. The black box approach is attractive because it requires no prior knowledge in the AI/ML model regarding how HPC-based physical simulations are implemented. Black box models are quick and easy to train and can be applied to any field, providing that sufficient data are available. Examples are numerous, including porous media simulations [13, 14, 15], nuclear DFT [4, 5], cosmology [16], turbulence modeling in compressible flow [17], and climate science [18]. Despite the popularity of black box approaches, model accuracy depends heavily on the quality and amount of data used for training the model (henceforth, “training data”). Note that the generation of training data can itself be very computationally expensive, requiring sweeps through the parameter space of the high-fidelity model, although many sparse sampling schemes (e.g., Clenshaw-Curtis grids [19]) and approaches to active learning have been invented to reduce the computational burden. Sole dependence on training data, with no inclusion of scientific “smarts” in the black box model, is the crux of such models’ limitations. For example, if a prediction is required outside of the region covered by the training data, (i.e., extrapolation), model accuracy tends to be poor. In addition, physics constraints (e.g., symmetry, positivity, or conservation) are not typically satisfied by black box approaches. Also, although they generally perform well, when these models fail, it is hard to analyze when and why because of their black box nature.

Interpretable Surrogate Models. To overcome these issues, *interpretable surrogates* have emerged, in which I/O data are related by known forms—typically differential equations—which are easier to analyze than neural networks. For example, eigenvalue analysis of the underlying system tells us whether the dynamics will be stable or not. The governing differential equations can be discovered by means of several mechanisms (e.g., using sparse or dense regression [20, 21, 22, 23], symbolic regression [24, 25], or Koopman operators [26, 27], and neural networks [28]). These approaches have shown promising results. For instance, they have accurately identified some known partial differential equations, such as Lorenz equations, using noisy data (i.e., simulation data corrupted by synthetically generated noise).

Physics-Informed Surrogate Models. Recently, advances in *physics-constrained, data-driven modeling* have emerged. Physics-informed neural networks [29], for example, embed the underlying differential equations within the training of a neural network by adding the residual term in the objective function. This innovative method of solving inverse problems does not rely on forward simulations. However, training the neural network is computationally expensive, is not very scalable across processors, and requires significant human intervention. The method’s accuracy is also not as robust as that of classical numerical methods. Developing a deeper understanding of the convergence behavior of physics-informed neural networks is a critical and active

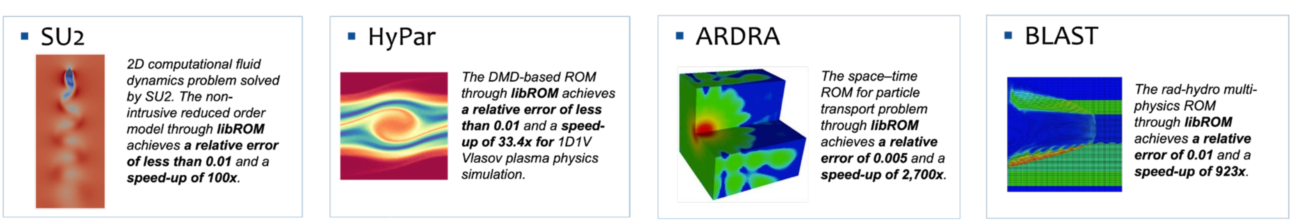

Figure 1-1. AI-based surrogates can accelerate computational fluid dynamics, plasma physics, particle transport, and multiphysics simulations with a high accuracy.

research area [30, 31, 32]. Another approach to physics-constrained, data-driven modeling is to directly learn infinite dimensional operators with neural networks, such as DeepONet [33] and the Fourier neural operator [34]. However, for problems with complex spatial domains and those of a multi-scale or multiphysics nature, classical numerical discretization methods provide more robust and accurate solutions.

Closures (or constitutive models) are a special type of physics-informed surrogate model that are used to simulate the effect of fine-scale physical processes in system-level models. Because large-scale, system-level simulations (e.g., complex engineered systems and Earth-system simulations) cannot afford the high computational cost of modeling fine-scale physics, closures are used to approximate them. Several theories regarding the structure of closures have been proposed, but they include unspecified constants, which are traditionally calibrated to simple experiments but tend to be inaccurate in realistic situations. Recently, these closures have been learned from high-fidelity simulations (as well as experimental data), with the constants replaced by functions (usually neural networks) of the state of system-level simulations. The innovation lies in devising transformations of the state so that the inputs into the neural network preserve invariance properties [35, 36, 37, 38]. It is also possible to reconstruct the fine-scale processes (and not just their effect) from the system-level information using spatial patterns learned from training data [39].

Reduced-Order Models. The next natural category of physics-constrained, data-driven methods consists of *projection-based reduced-order models (ROMs)*, in which the known physics constraints are explicitly used to relate data by projecting high-fidelity governing equations to a low-dimensional manifold. These approaches take advantage of not only the available first principles, but also the classical numerical discretization of the governing equations. Because of the explicit use of first principles, each step of a ROM systematically builds on the previous steps, exposing all relationships. Therefore, as a final product, ROMs can deliver tunable accuracy with adjustable speed-up, providing great flexibility and robustness. However, the development time for a projection-based ROM is longer than that for a black box approach because such models require significant human ingenuity to architect prior to training, and the training is computationally expensive.

Projection-based ROMs have been used for Euler equations [40, 41, 42, 43], Navier–Stokes equations [44, 45, 46], large-scale Boltzmann problems [47], lattice-type structure response problems [48, 49], digital twins of a fixed-wing unmanned aerial vehicle [50], and design optimization problems [51, 52, 53, 54, 48, 49]. However, these traditional linear subspace-projection-based ROMs are often inaccurate in low-dimensional solution representation for problems with slowly decaying Kolmogorov's width (e.g., advection-dominated moving-shock problems). To overcome this issue, efficient nonlinear manifold ROMs [55, 56] have been developed, in which the nonlinear manifold solution representation through neural networks is used to effectively capture solution dynamics with low-dimensional latent spaces. Figure 1-2 illustrates the accuracy and robustness in extrapolation and speed-up trends measured against the level of physics embedded in the ROM.

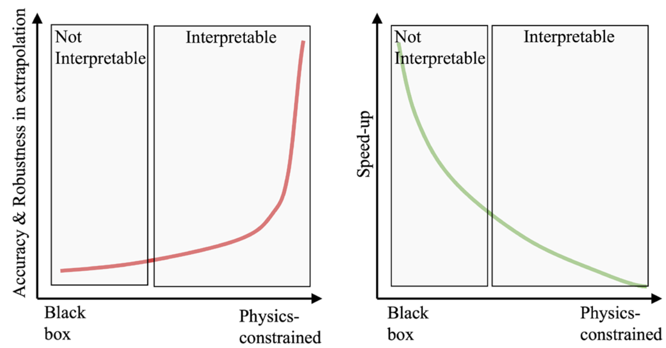

Figure 1-2. Categorization of surrogates according to their incorporation of physics-constraints (the underlying governing equation or the existing numerical discretization methods for the governing equation). Having more physics constraints often means increased accuracy and robustness in extrapolation. Conversely, it also means decreased speed-up.

Researchers have completed deep theoretical work over the past several decades to develop projection-based ROMs. With the emergence of ML, projection-based ROMs and AI are expected to merge and improve the field of surrogates for HPC simulations.

To summarize, surrogate models can be categorized by their sophistication and adherence to physics (by construction). The simplest type, black box models, are purely statistical and machine-learned constructs, without any physics/scientific "smarts" built into them. They are easier to train, but their failures are difficult to diagnose and fix. "Interpretable" surrogate models introduce a degree of causality and physical constraint in the surrogate model's architecture, allowing easier diagnoses of failures. Physics-informed surrogate models are statistical constructs that honor (approximately) the governing equations of the phenomena being modeled but are very difficult to train and use. ROMs, the last category, are not statistical models but they are rather derived from the governing equations via controllable approximations. They are difficult to formulate, but once trained, allow users to trade off complexity (and computational speed) versus approximation error.

DOE has been a world leader in construction and integration of surrogates into various applications, particularly for uncertainty quantification (UQ). However, there is fertile ground for improved surrogate methods using AI and for surrogate integration into AI workflows to meet future needs. The experience from these early adoptions allows us to chart a course over the next decade to integrate surrogate model approaches to address computational challenges throughout DOE mission areas.

1.2 Grand Challenges

In order to meet the next generation of its mission challenges, DOE must establish world leadership in development and use of AI surrogates. This will mean (1) building general purpose, multi-simulator surrogate models for specific domains, (2) establishing self-guided surrogate model construction from highly complex data structures and with physical constraints, and (3) developing infrastructure to smoothly plug surrogate models into HPC simulations, AI training workflows for autonomous systems, and monitoring/digital twins for seamless integration, agnostic to the software frameworks used to build the surrogate.

Building general purpose, domain meta-surrogates to combine data from across scales, from diverse models, and of varying fidelities. Predominantly, state-of-art, simulation-based science is at a point where substantial progress relies on the development of surrogates for specific applications, often relying on single models of a fixed fidelity. This fails to fully leverage the modeling capabilities in DOE science.

For many DOE applications, there are competing simulation models, each of which may exist for different levels of cost/fidelity trade-off. Typically, independent surrogate models are built for single simulators on specific problems, even for very similar tasks. This “siloed” behavior is inefficient and costly. Instead, building general purpose domain surrogates that can learn from data and sub-surrogates from all models, across fidelities, will make more accurate and robust predictions. This will reduce the impact of model-form error in individual codes.

An example would be a meta-surrogate for high-energy-density (HED) hydrodynamic systems that can take in data from simulating heterogeneous HED systems with multiple codes (i.e., xRAGE and Hydra) and can build a meta-surrogate that can give accurate prediction of individual code output for new cases, but also give a prediction for the physical system leveraging information from all codes simultaneously.

Establishing self-guided meta-surrogate model construction from complex data structures and with physics constraints. Building high-quality surrogate models is data-expensive and requires both detailed understanding of the problem structure and substantial effort to identify weaknesses in simulator or surrogate predictive capability with iterative model refinement. Research advances in AI for surrogate models will lead to AI-driven, intelligent data collection for adaptive training of surrogate models and iterative model criticism and improvement for stronger surrogate predictive capability. Advances in cost-aware active learning will allow meta-surrogates to:

a. Dynamically identify and collect data across models and experimental space to ensure predictions are accurate with controlled uncertainties.
b. Test and improve model structure to reduce model-form error in simulation and surrogate performance.
c. Dynamically expand and contract the model parameter space to adapt to effect sparsity for insensitive model inputs while growing to handle prediction for new scenarios and system designs.

Developing infrastructure to smoothly integrate surrogate models into HPC simulations, AI training workflows for autonomous systems, and monitoring/digital twins that are agnostic to the software frameworks used to build the surrogate. To fully utilize surrogates to meet the needs of the other AI building blocks and for the domain goals, surrogates must be smoothly, easily integrated into DOE workflows. The software and hardware infrastructure to make a portable, performant, platform-agnostic framework for composing, combining, and adapting AI surrogates will enable their smooth assimilation and allow for quick incorporation and testing of new methods and approaches as surrogate technology advances.

For example, as mentioned in 2.1, large-scale HPC simulation can benefit from leveraging surrogate models as “closures.” The communication layer to easily incorporate surrogates of specific closure models or meta-surrogates encompassing multiple closure models, regardless of the framework for the trained surrogate, does not currently exist but would be critical for wide-spread adoption of closure surrogates.

Even further, this infrastructure will allow multi-scale surrogate-simulator-AI systems to meet DOE goals. Figure 1-3 shows an abstracted diagram of a multi-scale system where the full system is used for monitoring and control with a micro-scale closure model that also allows for micro-scale system monitoring. Integrating the full micro-scale simulator into the full-scale code would be infeasible, as would running the full-scale simulation in the monitoring and control loop. A surrogate of the micro-scale, using active learning to guide evaluation of the micro-scale simulation, then is used as a full-scale closure. The full-scale surrogate is used for fast-querying for control and monitoring tasks. This diagram shows a general form of multi-scale surrogate-simulation-AI system that would be unlocked by the development of this infrastructure.

Meeting these grand challenges is fundamental to other AI building blocks described in this section as well as domain needs. *Composable* surrogate models are necessary for capturing all scales for building a full-scale digital twin of the power grid for monitoring, testing “what-if” scenarios, and generating AI strategies for handling disruptions in real time (these are further discussed in Chapters 03 and 04).

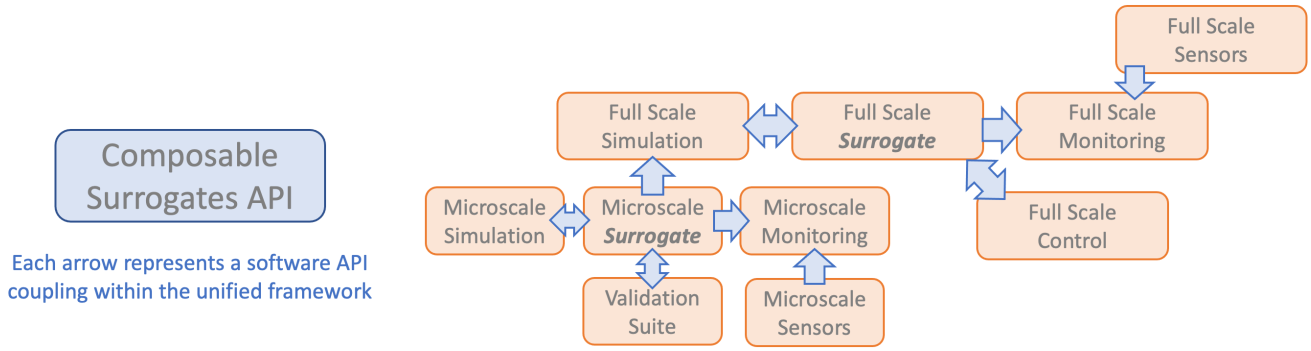

Figure 1-3. Diagram of infrastructure allowing surrogate model integration into a multi-scale monitoring and control problem with multiple HPC simulations and real-world data integrated into a single system.

Multi-scale simulation for DOE applications, from stockpile management to fusion energy systems to climate science, is founded upon modeling heterogeneous processes at high fidelities. Doing so, while leveraging the best DOE scientific computing capabilities, requires surrogate models to provide fast, accurate prediction with well-quantified uncertainties for sub-scale physics to large-scale models.

Surrogate models are also critical for fast iteration to guide optimal, autonomous design across the DOE mission space, from National Nuclear Security Administration (NNSA) interests to energy production and storage, to Energy Earthshots (see Section 02 of this report).

1.3 Advances in the Next Decade

These Grand Challenges motivate three surrogate-specific capabilities that we seek to develop within the DOE community in the next decade. Underlying these are additional requirements for advances in cross-cutting technologies: data management infrastructure for large training data and hardware platforms for heterogeneous workflows. Thus, we follow with a set of challenges organized by the five crosscutting technologies outlined in Section 03 of this report.

1.3.1 SURROGATE-SPECIFIC CAPABILITIES

1. **Building AI-enhanced surrogate models that handle complex data structures into technical workflows**: Future workflows will require the capability to handle complex image, natural language, and graph/network data structures while mixing surrogate models with conventional models in the form of networks or hierarchies. New research incorporating advances in AI with these data modalities as well as new workflow tools to assist with the assembly handling the disparity in scales will be required.
2. **Constructing trusted surrogate models**: DOE's established practice of making critical decisions based on model predictions requires understanding and communicating bounds on model-based predictions. This problem becomes exponentially more difficult with the development of a networked hierarchy of models in meta-surrogates. Building the foundations to provide meaningful UQ bounds when the assumptions of traditional methods must be violated and when generalizing to data beyond that previously seen in training will be critical to building trust in AI surrogate use.
3. **Training surrogate models in a scalable and sustainable manner**: The capability to scalably (in terms of processors, platforms, and users) train surrogate models will require a new AI-enhanced, multi-platform software framework. Such a framework does not exist and will require significant research to define an appropriate architecture before it can be implemented and tested.

We expand on these three surrogate-specific decadal advances next.

Building AI-enhanced surrogate models that handle complex data structures into technical workflows. Scientific workflows consist of a series of transitions to data by modules. These modules have traditionally consisted of physics-based models or post-processing scripts. The workflows are shallow, and workflow automation tools such as Sandia Analysis Workbench (SAW, [57]) can address them. In the future, however, some modules may be entirely AI/ML-based, while others may be hierarchies of surrogate models or have closures embedded in a physics-based model. Such hybrid networks of surrogates, which may embody physics at disparate time/length scales, are invariably "stiff" systems; no scalable methods can address them. Advances in fundamental mathematics that can either address the contrast in scales or smooth them over will be a prerequisite for achieving this capability.

Inference from observation (i.e., inverse problems, property inference, inverse design) plays a large role in scientific research; the existence of (networks of) fast-running surrogate models can enable the solution of high-dimensional inverse problems (e.g., estimation of multi-dimensional fields rather than scalar parameters). However, in many DOE applications, observations are sparse, and scientific research in DOE has typically employed Bayesian methods that quantify uncertainty in the outputs of such inverse problems. However, scalable, *high-dimensional Bayesian inversion* *solvers* that are performant are rare, and fundamental research in Bayesian mathematics will be required to exploit the power of composable networks of surrogate models.

The Idea of *composable surrogates* also raises several challenges: how to achieve the compositions; how to detect and avoid surrogate-to-surrogate incompatibilities during assembly; and how to compile, maintain, and integrate a repository of composable surrogates with tools that compose a scientific workflow. Constructing such mix-and-match workflows will require concepts borrowed from component-based software design, which DOE has explored in the past [58], but which have not been widely accepted in scientific simulations. Scientific workflows of the future, with a mixture of physics-based and data-driven models, will thus require workflow tools that do not exist today.

Recent advances in generative models for text, images, video, and networks hint at the great potential for surrogate modeling of complex data structures generated by big physics facilities—radiograph images, temporally and spatially-resolved spectroscopy, etc.—that are currently converted to scalar or low-dimensional vector summaries for surrogate modeling. Advances in connecting surrogates of complex data from simulations to similar data structures generated by experimental diagnostics will be critical for allowing surrogate-simulator-AI systems to connect into experimental workflows to meet autonomous science goals.

Constructing trusted surrogate models. Despite their predictive skill, surrogate models are approximate and can fail in myriad ways, the most common being out-of-distribution (OOD) use (i.e., outside the feature space spanned by the training dataset). This affects the generalization ability, uncertainty assessment, and robustness of surrogate models. One strategy to improve the trustworthiness of surrogate modeling results is to impose physical realizability constraints, either during model construction or training. The discovery of *causal relationships* in training data, by assembling/integrating fundamental relationships predicted by physics (in contrast to relying on correlations discovered in data) can also be called “trust by construction” (see example in [59]). Such assembly will be necessary for AI-enhanced control laws used in complex engineered systems such as scientific instruments or autonomous vehicles, which may have to function in contested environments in which they may encounter scenarios outside their training data (see Chapter 04). Today there are no general methods by which physics can be included in the architecture of an arbitrary surrogate model, although much work has been done for specific types of surrogates such as closures [35, 38]. Thus, endowing surrogate models with trust during construction will require further work in the mathematics of OOD detection, causality, and other aspects of surrogate modeling.

A second approach to building trust in surrogate models is to provide uncertainty bounds with their predictions. Minimizing uncertainties increases the requirements for the quantity and diversity of training data. An example of this approach is to assemble training data from various sensing modalities (e.g., images, time series, and tensors). However, this is not currently used because of our fundamental ignorance of how *multimodal data* may be assimilated into surrogate models, given that there will be wide disparities in their fidelities, quantities, and forms. Transfer learning could potentially address this problem, but we currently lack the mathematical basis for learning from multimodal data.

The crudest—but perhaps the most effective—way of endowing surrogate models with trustworthiness is to qualify them (i.e., determine the types of physics/processes present in their training data and demarcate the feature-space where the surrogate model may be used). However, such qualification requires that developers create unsupervised or semi-supervised methods to characterize the training dataset, which in turn necessitates that they incorporate physics/domain information into the unsupervised methods. Some preliminary work has been done [60], but general techniques that will scale to multiple types of physics have not been developed. Fundamental algorithmic research is thus necessary to enable developers to qualify surrogate models, as well as to construct the software frameworks with such hybrid unsupervised learning methods.

Training surrogate models in a scalable and sustainable manner. The widespread use of surrogate models across the DOE complex will require automating their construction, likely via AI/ML agents. Because surrogate models are first trained on traditional simulation model-generated data (and in some cases further tuned using experimental data), automation of the (adaptive) sampling of the input space (to generate informative datasets) and selection and tuning of the surrogate model architecture will be required. The training process may also span multiple hardware architectures, each optimized for the disparate tasks involved in constructing the surrogate. Significant research has gone into specific tasks such as active learning and adaptive design of experiments to efficiently generate training data, and automated tuning of machine-learned model architectures. But other key tasks remain, including AI-based orchestration of the training process and embedding those orchestration tools in an *AI-enhanced software framework* that constructs surrogates.

Such a software framework, supporting the multiple platforms where the training process is executed, must be designed and developed. Preliminary work suggests that such a framework is possible. Dakota [61] automates the process of sampling an input space, running simulations to generate training data and training a surrogate on these data, but is limited to conventional surrogate models and does not span platforms. SAW [57] is a workflow automation tool that maintains the provenance of all simulations within its purview and integrates with Dakota, but it is limited to conventional platforms. In both Dakota and SAW, the workflow is automated via expert-driven scripts rather than AI. Thus, although they may form the starting points of the AI-enhanced software framework we envision, the final product will require significant research in the appropriate framework architecture and design, followed by implementation and evaluation.

1.3.2 CROSSCUTTING TECHNOLOGY CAPABILITIES

To meet the challenges outlined above, advancements must also be made in the technical crosscuts detailed in the chapters comprising Section 03 of this report. These include (1) the fundamental mathematical underpinnings of surrogate modeling; (2) the software frameworks for building and training surrogates; (3) corresponding frameworks for integrating surrogates into workflows; (4) data handling for implementation and integration of surrogates and AI into DOE infrastructure; and (5) hardware architectures that provide scalability, flexibility, and composability—from HPC to edge. Next, we outline priority directions in these areas that will bridge the technological gap previously discussed regarding the development and application of surrogate models.

Mathematics and fundamental research. Research to address the current shortcomings of surrogate models, as identified above, will require advances along four fronts. First, we need *a new theory of surrogate models to establish when such a model is ready for production use*. This theory would be similar in character to proofs of convergence of statistical models, replacing the standard notions of convergence in some metric with some quantitative measure of correctness and consistency. Proofs of consistency and correctness [62, 63] would also facilitate detection of outliers and rare events. The mathematical properties of surrogates (e.g., stiffness), and restrictions on their use cases (e.g., detection of OOD use) that bound their generalizability, must also be identified and quantified.

Second, research is needed to create a *new framework that extracts surrogate models from multimodal data* without imposing a fixed architecture. This will require methods to impose priors/constraints/regularizations in such a setting (e.g., to embed physical models, conservation laws) with the algorithm that discovers the model. It is currently unclear how this might be done, beyond the obvious method of including the constraints in the loss function.

Third, we need *training algorithms that can fit models to data under prescribed requirements for accuracy, cost, and resources*. These algorithms must learn and exploit the geometry of the training data and select training samples where needed (i.e., active learning), requiring research on how a finite set of samples needs to be distributed within a high-dimensional feature space to maximize the extraction of information [64, 65].

The final mathematics and fundamental research thrust regards the *development of verification and validation methods for surrogate models, so that they will be trustworthy*. This must encompass new methods for explainability [66, 67] and interpretability, as well as unsupervised/semi-supervised methods that quantify the information content of a training dataset (e.g., identify the types of physics it has). Methods that extract a set of representative prototypes from a training dataset (to allow deep dives) are part and parcel of the methods for trustworthy AI [60].

Software frameworks for training AI surrogates. In order to meet the software needs for AI-surrogate modeling and to leverage the power of surrogates in DOE computing, we need to invest in software development for (1) a portable, performant, platform-agnostic framework for composing, combining, and adapting AI surrogates; (2) a software framework for deploying reproducible, verified, and validated surrogate libraries; and (3) software for AI-driven, automated surrogate construction with a high-level front end to enable domain scientists to build surrogates without requiring extensive AI expertise. Together, these three priority research directions impact all aspects of leveraging surrogate modeling to achieve autonomous and AI-accelerated scientific discovery. The framework for composing, combining, and adapting AI surrogates will ensure that advances in surrogate modeling can be integrated with heterogeneous codes and executed on diverse hardware for AI and autonomous systems. An infrastructure for verified and validated surrogate libraries will ensure trustworthy, reliable deployment of surrogates across domains, while AI-driven surrogate construction will reduce barrier-to-entry for domain experts to utilize surrogate model technologies to make scientific advancements. Further, these advancements will integrate smoothly into autonomous workflows to accelerate experimentation and discovery by minimizing human-in-the-loop factors, as we discuss next.

Workflows for integrating surrogates and AI. Making these software advances will facilitate necessary advancements in building AI workflows for DOE science. Leveraging the power of AI to create workflow composition assistants that translate scientific problems into workflows without requiring complex domain or computational knowledge will streamline the paths to solutions, leveraging DOE's diverse leadership computing architectures and experimental facilities. Moreover, it will broaden and diversify participation in the DOE science mission, catalyzing new ideas and strategies. As we advance autonomous science across the DOE complex, *self-healing workflows that auto-detect and correct errors* (malicious or unintentional) at scale will be necessary to ensure robust operation. Such workflows will include automated detection of surrogate failure and degradation (e.g., when surrogate systems leave domains of trustworthiness) to ensure that the AI agents are relying on accurate approximations of computational science models and not untrusted extrapolations. By the end of the decade, we foresee an intelligent, AI-driven, federated workflow scheduler that dynamically executes workflows from the exa-cluster to the edge, integrating scientific instruments and self-driving laboratories (as discussed in Chapter 05) to accelerate science and engineering breakthroughs across the DOE.

Data management for integration of surrogates. Advances in data management will also be critical to the next decade of AI research and development toward AI-enabled science. Data wrangling (e.g., finding, cleaning, feature engineering) represents a significant fraction of the process of building, training, and improving surrogates. The use of AI systems to reduce human involvement is urgently needed to dramatically reduce the time and cost of data wrangling. Investment in methods for storing, sharing, and finding heterogeneous data sources, along with automated data preparation and augmentation, will ensure data availability with high throughput for AI training, testing, and operational tasks. Fully leveraging the wealth of data generated by the DOE scientific enterprise, will demand infrastructure for efficiently sharing data, including real-time continuous data, using intuitive queries both across the DOE complex and with academic and industry partners. Beyond simply making the data available, leveraging AI to interrogate available data for data selection, recommendation, classification/labeling, and generating configurable data preparation and augmentation pipelines will reduce the data processing overhead necessary in AI workflows.

Hardware architectures for AI surrogate integration. Operating AI systems across the computational continuum of exascale to edge will require not only advances in software and workflows to bridge the heterogeneous scales, but new frontiers in flexible, composable hardware to reach the potential that autonomous science offers. Investment in composable hardware accelerators will ensure that surrogates can be built, trained, and tested throughout the AI pipelines, allowing adaptive scalability as requirements for model size and resource vary. Large computational facilities will be critical for providing data from world-class simulation for training surrogate models, as well as providing “exascale as a service” capability for AI surrogates to execute high-fidelity simulations “as needed” for active learning. It will also be essential to develop hardware features to deploy trained AI agents seamlessly and robustly from leadership-class HPC with bespoke accelerators to low-power edge devices sensitive to SWAP (size, weight, and power). These advanced hardware architectures will require native support for UQ, and they must be built for robustness (physical robustness for edge devices, and for long, stable operation for training surrogates).

1.4 Accelerating Development

To enable the long-term achievement of the research thrusts described above, we propose several candidate pilot projects that have the potential to jump-start some of the theoretical and algorithmic development outlined, using existing data and incremental extensions of existing tools. These pilots illustrate a pathway to immediately begin making progress toward achieving the identified grand challenges.

- *ML closures for plasma turbulence for fusion learned from high-fidelity sub-scale simulation and experimental training datasets:* This pilot will require the development of (1) high-dimensional, high-order, scalable optimization algorithms for fitting ML surrogate models, and (2) development of composable infrastructure between sub-scale simulation, full-scale simulation, and inference using experimental data. These directly connect to the first and third grand challenges.
- *ML surrogate for the ocean model in Earth system models to accelerate spin-up*: This pilot could develop techniques to impose stability in networks of surrogates operating at different length and/or time scales in multiphysics and/or multi-fidelity networks, and efficient design and training of surrogates relevant to the second and third grand challenges.
- *Discover biological mechanisms that link environmental forcing (hyperspectral data) to biological response (omics data):* This pilot would investigate how surrogates could be trained by assimilating multimodal data (hyperspectral data and omics data). In addition, the pilot could investigate whether the architecture of the surrogate model could be discovered from data. These goals would be applicable to the first and second grand challenges.
- *Generative ML surrogate for large-scale cosmology simulations*: This pilot includes ML-accelerated subgrid physics and AI/ML methods for increasing dynamic range by incorporating learning from high-resolution simulations (“super-resolution”). This pilot essentially involves construction of AI-enhanced closures for cosmological simulations and embedding them in a physics model that is part of a complex cosmological simulation workflow relevant to the third grand challenge.

1.5 Expected Outcomes

Advancements in AI surrogate models and their integration into science, engineering, and autonomous workflows will accelerate science and engineering to meet the grand challenges we face, many of which are detailed in Section 02 of this report. Building surrogates into AI workflows will ensure that trained AI agents are able to explore and learn from the highest-quality computational approximations to physical systems, unlocking the potential of autonomous systems and AI for DOE science. These capabilities can revolutionize power generation, storage, and delivery for the 21st century; manage the nation's nuclear stockpile so experts can evaluate weapons performance with confidence and make informed decisions without relying on nuclear testing; drive advancement in fusion energy science, ensuring that the U.S. leads the way to fusion power generation; and provide insights needed to address a rapidly changing climate and avoid or mitigate environmental catastrophes.

DOE has invested over decades to become the world's leader in scientific computing, creating physics simulations that can represent complex processes in real-world systems with unmatched fidelity. AI surrogates represent a unique opportunity to increase the impact of these investments by enabling improvements in model execution time by factors ranging from 100 to 1B. By investing in research and development for AI surrogate technology and building surrogate models of high-performance physics simulation into AI workflows, we will leverage DOE's expertise to solve the big problems that impact our nation and the world.

1.6 References

[1] Choi, Y., Arrighi, W.J., Copeland, D.M., Anderson, R.W. and Oxberry, G.M., 2019. *libROM*. Lawrence Livermore National Laboratory, Livermore, CA (United States).

[2] Kennedy, M.C., and O'Hagan, A., 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3), pp. 425–464.

[3] Heitmann, K., Bingham, D., Lawrence, E., Bergner, S., Habib S., Higdon D., Pope, A., et al., 2016. The Mira–Titan Universe: precision predictions for dark energy surveys. *The Astrophysical Journal* 820(2).

[4] Schunck, N., McDonnell, J.D., Higdon, D., Sarich, J., and Wild, S.M., 2015. Uncertainty quantification and propagation in nuclear density functional theory. *The European Physical Journal A*, 51(12), pp. 1–14.

[5] Schunck, N., O'Neal, J., Grosskopf, M., Lawrence, E., and Wild, S.M., 2020. Calibration of energy density functionals with deformed nuclei. *Journal of Physics G: Nuclear and Particle Physics,* 47(7).

[6] Tapia, G., Khairallah, S., Matthews, M., King, W.E., and Elwany, A., 2018. Gaussian process-based surrogate modeling framework for process planning in laser powder-bed fusion additive manufacturing of 316L stainless steel. *The International Journal of Advanced Manufacturing Technology*, 94(9), pp. 3591–3603.

[7] Daniel Marjavaara, B., Staffan Lundström, T., Goel, T., Mack, Y. and Shyy, W., 2007. Hydraulic turbine diffuser shape optimization by multiple surrogate model approximations of Pareto fronts. *Journal of Fluids Engineering,*129(9), pp. 1228–1240.

[8] Huang, F., Wang, L., and Yang, C., 2015. Hull form optimization for reduced drag and improved seakeeping using a surrogate-based method. In: *The Twenty-fifth International Ocean and Polar Engineering Conference,* OnePetro, June.

[9] Han, Z.H., Görtz, S., and Zimmermann, R., 2013. Improving variable-fidelity surrogate modeling via gradient-enhanced kriging and a generalized hybrid bridge function. *Aerospace Science and technology*, *25*(1), pp. 177–189.

[10] Han, Z.H. and Görtz, S., 2012. Hierarchical kriging model for variable-fidelity surrogate modeling. *AIAA Journal*, *50*(9), pp. 1885–1896.

[11] Guo, X., Li, W., and Iorio, F., 2016. Convolutional neural networks for steady flow approximation. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 481–490. August.

[12] Zhang, Y., Sung, W.J., and Mavris, D.N., 2018. Application of convolutional neural network to predict airfoil lift coefficient. In: *2018 AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference,* p. 1903.

[13] Kadeethum, T., O'Malley, D., Fuhg, J.N., Choi, Y., Lee, J., Viswanathan, H.S., and Bouklas, N., 2021. A framework for data-driven solution and parameter estimation of PDEs using conditional generative adversarial networks. *Nature Computational Science,* 1, pp. 819–829.

[14] Kadeethum, T., Ballarin, F., Choi, Y., O'Malley, D., Yoon, H., and Bouklas, N., 2022. Non-intrusive reduced-order modeling of natural convection in porous media using convolutional autoencoders: Comparison with linear subspace techniques. *Advances in Water Resources*, 160, p. 104098.

[15] Kadeethum, T., O'Malley, D., Choi, Y., Viswanathan, H.S., Bouklas, N., and Yoon, H., 2021. Continuous conditional generative adversarial networks for data-driven solutions of poroelasticity with heterogeneous material properties. *arXiv preprint,* arXiv:2111.14984.

[16] Heitmann, K., Bingham, D., Lawrence, E., Bergner, S., Habib, S., Higdon, D., Pope, A., Biswas, R., Finkel, H., Frontiere, N., and Bhattacharya, S., 2016. The Mira–Titan universe: Precision predictions for dark energy surveys. *The Astrophysical Journal*, 820(2), p. 108.

[17] Ray, J., DeChant, L., Lefantzi, S., Ling J., and Arunajatesan, S., 2018. Robust Bayesian calibration of a k-e model for compressible jet-in-crossflow simulations. *AIAA Journal*, 56(12), pp. 4893–4909, December.

[18] Huang, M., Ray, J., Hou, Z., Ren, H., Liu, Y., and Swiler, L., 2016. On the applicability of surrogate-based MCMC-Bayesian inversion to the Community Land Model: Case studies at flux tower sites. *Journal of Geophysical Research – Atmospheres*, 121(13).

[19] Smith, R.C., 2014. *Uncertainty Quantification*, SIAM Computational Science and Engineering Series.

[20] Brunton, S.L., Proctor, J.L., and Kutz, J.N., 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. In: *Proceedings of the National Academy of Sciences*, 113(15), pp. 3932–3937.

[21] Fries, W.D., He, X., and Choi, Y., 2022. LaSDI: Parametric latent space dynamics identification. *arXiv preprint,* arXiv:2203.02076.

[22] He, X., Choi, Y., Fries, W.D., Belof, J., and Chen, J.S., 2022. gLaSDI: Parametric physics-informed greedy latent space dynamics identification. *arXiv preprint*, arXiv:2204.12005.

[23] Qian, E., Kramer, B., Peherstorfer, B., and Willcox, K., 2020. Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems. *Physica D: Nonlinear Phenomena*, 406, p. 132401.

[24] Schmidt, M., and Lipson, H., 2009. Distilling free-form natural laws from experimental data. *Science*, 324(5923), pp. 81–85.

[25] Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S., 2020. Discovering symbolic models from deep learning with inductive biases. *arXiv preprint,* arXiv:2006.*11287*.

[26] Mezić, I., 2013. Analysis of fluid flows via spectral properties of the Koopman operator. *Annual Review of Fluid Mechanics*, 45, pp. 357–378.

[27] Huhn, Q., Tano, M.E., Ragusa, C.R., and Choi, Y., 2022. Parametric dynamic mode decomposition for reduced order modeling. *arXiv preprint,* arXiv:2204.12006.

[28] Koch, J., 2021. Data-driven surrogates of rotating detonation engine physics with neural ordinary differential equations and high-speed camera footage. *Physics of Fluids,* 33, p. 091703. https://doi.org/10.1063/5.0063624

[29] Raissi, M., Perdikaris, P., and Karniadakis, G.E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, pp. 686–707.

[30] Wang, S., Yu, X., and Perdikaris, P., 2022. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449, p.110768.

[31] Wang, S., Teng, Y., and Perdikaris, P., 2021. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5), pp. A3055–A3081.

[32] Shin, Y., Darbon, J., and Karniadakis, G.E., 2020. On the convergence of physics-informed neural networks for linear second-order elliptic and parabolic type PDEs. *arXiv preprint,* arXiv:*2004.01806*.

[33] Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G.E., 2021. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3), pp. 218–229.

[34] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A., 2020. Fourier neural operator for parametric partial differential equations. *arXiv preprint,* arXiv:*2010.08895*.

[35] Ling, J., Kurzawski, A., and Templeton, J., 2016. Reynolds-averaged turbulence modeling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics,* 807, pp. 155–166. doi:10.1017/jfm.2016.615

[36] Singh, A.P., Medida, S., and Duraisamy, K., 2017. Machine-learning-augmented predictive modeling of turbulent separated flows over airfoils. *AIAA Journal,* 55 (7), pp. 2215-2227.

[37] Frankel, A.L., Safta, C., Alleman, C., and Jones, R., 2022. Mesh-based graph convolutional neural networks for modeling materials with microstructure. *Journal of Machine Learning for Modeling and Computing*, 3(1).

[38] Frankel, A.L., Jones, R.E., and Swiler, L.P., 2020. Tensor basis Gaussian process models of hyperelastic materials. *Journal of Machine Learning for Modeling and Computing*, 1(1).

[39] Kim, H., Kim, J., Won, S., and Lee, C., 2021. Unsupervised deep learning for super-resolution reconstruction of turbulence. *Journal of Fluid Mechanics,* 910, p. A29. doi:10.1017/jfm.2020.1028

[40] Copeland, D.M., Cheung, S.W., Huynh, K., and Choi, Y., 2022. Reduced-order models for Lagrangian hydrodynamics. *Computer Methods in Applied Mechanics and Engineering*, 388, p. 114259.

[41] Amsallem, D., and Farhat, C., 2008. Interpolation method for adapting reduced-order models and application to aeroelasticity. *AIAA journal*, 46(7), pp. 1803–1813.

[42] Cheung, S.W., Choi, Y., Copeland, D.M., and Huynh, K., 2022. Local Lagrangian reduced-order modeling for Rayleigh-Taylor instability by solution manifold decomposition. *arXiv preprint,* arXiv:*2201.07335*.

[43] Lauzon, J.T., Cheung, S.W., Shin, Y., Choi, Y., Copeland, D.M., and Huynh, K., 2022. S-OPT: A points selection algorithm for hyper-reduction in reduced-order models. *arXiv preprint,* arXiv:2203.16494.

[44] Xiao, D., Fang, F., Buchan, A.G., Pain, C.C., Navon, I.M., Du, J., and Hu, G., 2014. Non-linear model reduction for the Navier–Stokes equations using residual DEIM method. *Journal of Computational Physics*, 263, pp. 1–18.

[45] Stabile, G., and Rozza, G., 2018. Finite volume POD-Galerkin stabilized reduced-order methods for the parametrized incompressible Navier–Stokes equations. *Computers & Fluids*, 173, pp. 273–284.

[46] Veroy, K., and Patera, A.T., 2005. Certified real-time solution of the parametrized steady incompressible Navier–Stokes equations: Rigorous reduced-basis a posteriori error bounds. *International Journal for Numerical Methods in Fluids*, 47(8-9), pp. 773–788.

[47] Choi, Y., Brown, P., Arrighi, W., Anderson, R., and Huynh, K., 2021. Space–time reduced-order model for large-scale linear dynamical systems with application to Boltzmann transport problems. *Journal of Computational Physics*, 424, p. 109845.

[48] McBane, S., and Choi, Y., 2021. Component-wise reduced-order model lattice-type structure design. *Computer Methods in Applied Mechanics and Engineering*, 381, p. 113813.

[49] McBane, S., Choi, Y., and Willcox, K., 2022. Stress-constrained topology optimization of lattice-like structures using component-wise reduced-order models. *arXiv preprint*, arXiv:2205.09629.

[50] Kapteyn, M.G., Knezevic, D.J., Huynh, D.B.P., Tran, M., and Willcox, K.E., 2022. Data-driven, physics-based digital twins via a library of component-based reduced-order models. *International Journal for Numerical Methods in Engineering*, 123(13), pp. 2986–3003.

[51] Choi, Y., Oxberry, G., White, D., and Kirchdoerfer, T., 2019. Accelerating design optimization using reduced-order models. *arXiv preprint,* arXiv:1909.11320.

[52] Choi, Y., Boncoraglio, G., Anderson, S., Amsallem, D., and Farhat, C., 2020. Gradient-based constrained optimization using a database of linear reduced-order models. *Journal of Computational Physics*, 423, p. 109787.

[53] Choi, Y., Oxberry, G., White, D., and Kirchdoerfer, T., 2019. *Accelerating topology optimization using reduced-order models.* LLNL-CONF-771564. Lawrence Livermore National Laboratory, Livermore, CA (United States).

[54] Amsallem, D., Zahr, M., Choi, Y., and Farhat, C., 2015. Design optimization using hyper-reduced-order models. *Structural and Multidisciplinary Optimization*, 51(4), pp. 919–940.

[55] Kim, Y., Choi, Y., Widemann, D., and Zohdi, T., 2021. A fast and accurate physics-informed neural network reduced-order model with shallow-masked autoencoder. *Journal of Computational Physics*, p. 110841.

[56] Kim, Y., Choi, Y., Widemann, D., and Zohdi, T., 2020. Efficient nonlinear manifold reduced-order model. *arXiv preprint,* arXiv:*2011.07727*.

[57] Sandia Analysis Workbench, 2022. National Technology and Engineering Solutions of Sandia, LLC. https://www.sandia.gov/saw/, accessed May 12, 2023.

[58] Allan, B.A., Armstrong, R.C., Wolfe, A.P., Ray, J., Bernholdt, D.E., and Kohl, J.A., 2002. The CCA core specification in a distributed memory SPMD framework. *Concurrency Computat.: Pract. Exper.,* 14, pp. 323–345. https://doi.org/10.1002/cpe.651

[59] Schmelzer, M., Dwight, R.P., and Cinnella, P., 2020. Discovery of algebraic Reynolds-stress models using sparse symbolic regression. *Flow Turbulence Combust* 104, pp. 579–603. https://doi.org/10.1007/s10494-019-00089-x

[60] Barone, M., Ray, J., and Domino, S., 2022. Feature selection, clustering, and prototype placement for turbulence data sets. *AIAA Journal*, 60(3), pp.1332–1246.

[61] Dakota Web page, 2021. National Technology and Engineering Solutions of Sandia, LLC. https://dakota.sandia.gov, accessed May 12, 2023.

[62] DeVore, R., Hanin, B., and Petrova, G., 2021. Neural network approximation. *Acta Numerica*, 30, pp. 327–444. doi:10.1017/S0962492921000052

[63] Petersen, P., 2022. *Neural Network Theory*. University of Vienna. http://pc-petersen.eu/Neural_Network_Theory.pdf, accessed May 12, 2023.

[64] Schmelzer, M., Dwight, R.P., and Cinnella, P., 2020. Discovery of algebraic Reynolds-stress models using sparse symbolic regression. *Flow Turbulence Combust* 104**,** pp. 579–603. https://doi.org/10.1007/s10494-019-00089-x

[65] Boullé, N., Earls, C.J., and Townsend, A., 2022. Data-driven discovery of Green's functions with human-understandable deep learning. *Sci Rep.* 12**,** p. 4824. https://doi.org/10.1038/s41598-022-08745-5

[66] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D., 2019. A survey of methods for explaining black box models. *ACM Comput. Surv.,* 51, 5, Article 93, September, 42 pp. https://doi.org/10.1145/3236009

[67] Burkart, N., and Huber, M.F., 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, pp. 245–317.

02. AI FOUNDATION MODELS FOR SCIENTIFIC KNOWLEDGE DISCOVERY, INTEGRATION, AND SYNTHESIS

Many of the U.S. Department of Energy (DOE) scientific domains and mission spaces contain precious few samples of interest that are properly labeled by subject matter experts (SMEs) but have vast troves of unlabeled datasets. To apply artificial intelligence (AI) methods in these areas will require breakthroughs in the field of low- or zero-shot learning to overcome the challenge of sparse labels. We define the concept of a master model as a class of models that demonstrate emergent behavior and can solve new tasks after "seeing" only a limited number of examples. Foundation models are a cutting-edge approach to developing master models.

Foundation models—built specifically for DOE missions—hold impressive promise for transforming both the way the DOE does its science and the impact and reach of that science. The concept of a foundation model is one of the most significant AI approaches derived from the scale of computation and data combined with the new computing and data systems being deployed through the DOE Exascale Computing Project (ECP), which are ideally suited for it. Per [1], a foundation model is one that is "trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks."

Foundation models are intended to become the digital equivalent of an SME; they will have a deep understanding of a particular domain, displaying the ability to develop keen insight and discover meaningful patterns in vast troves of data, that may initially seem uncorrelatable. The significance of foundation models cannot be overstated, as they "are based on standard ideas in transfer learning and recent advances in deep learning and computer systems applied at a very large scale, demonstrate surprising emergent capabilities [2] and substantially improve performance on a wide range of downstream tasks" [3].

Foundation models should be seen as a critical piece of a national science transformation, driven by the DOE AI for science, energy, and security mission areas, that will accelerate our pace of discovery for basic science, applied science, national security, and broader economic impact. Foundation models represent a pinnacle in inductive reasoning (models learned from data), and provide a significant, complementary asset to the standard deductive reasoning that is used to create the DOE's traditional modeling and simulation capability. Constructed using a transformer model architecture (Figure 2-1) [4], the key promise of foundation models is that they offer to extract previously unseen correlations and patterns within existing datasets, and to span the gap for application domains where there are no explicit governing equations or physical rules. As the term "foundation" implies, the DOE science community has the opportunity to create models analogous to large-scale instruments, enabling many individual teams to work together with many other teams to contribute data and expertise to build, and then incrementally train, a shared foundation model for their downstream tasks.

DOE has a clear mission-driven need to produce foundation models for science, energy, and national security. Across DOE missions, these models are likely to transform what is scientifically achievable. We expect that the combination of wide, downstream functionality with emergent capabilities will allow researchers to incorporate wide ranges of scientific knowledge and correlation, synthesize that knowledge to formulate profound new questions to set scientific scope, and rapidly find answers to previously unsolvable questions.

PROJECT SPOTLIGHT

Project Name: Transfer learning for inertial confinement fusion

PI: Luc Peterson

Organizations Involved: Lawrence Livermore National Laboratory, Weapons and Complex Integration – Inertial Confinement Fusion (ICF) and Advanced Simulation and Computing programs

Goal: Evaluate transfer learning as a method of calibrating a simulation-based neural network to experimental data, creating a model that is predictive of ICF experiments.

Significant Accomplishment: Leveraging just 19 Omega ICF experiments and 30,000 low-fidelity simulations in a 9D design space, we were able to create transfer learning neural network models that could predict the outcome of future experiments with significantly lower error than the simulations alone.

In the News: Humbird, K. D., Peterson, J. L., Spears, B. K., and McClarren, R. G., 2020, "Transfer learning to model inertial confinement fusion experiments," in *IEEE Transactions on Plasma Science* 48 (1), pp. 61–70, Jan. doi: 10.1109/TPS.2019.2955098*. This paper was a 2022 IEEE Transactions on Plasma Science (TPS) Best Paper Award winner.*

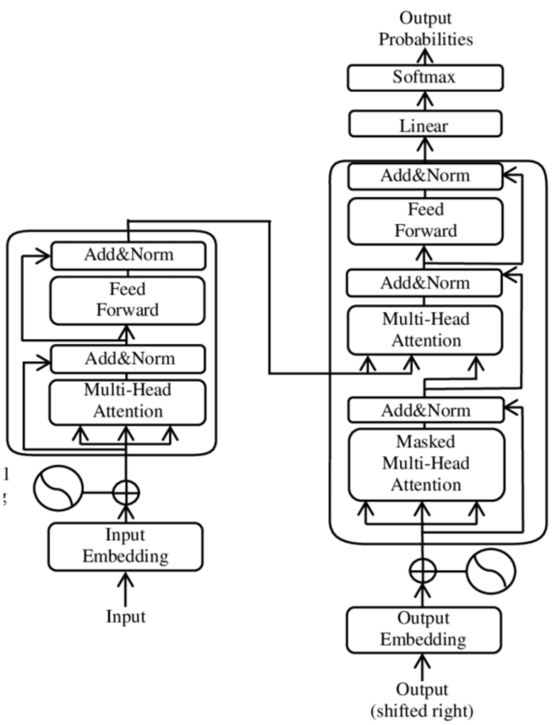

Figure 2-1. Transformer model architecture underlying foundation models [4].

Mature foundation models will be able to ingest a large range of data types—images, time series, scalars, and more—to produce a concise and never-before-seen representation of scientific phenomena. These representations will enable vastly improved predictive capabilities, greatly deepen the insight extracted from observations, again enabling scientists to ask and answer questions that are intractable today.

For example, a foundation model could be built on broad datasets, both experimental and simulated, from the spectrum of fusion energy research to answer fundamental questions essential to fusion energy production, such as building on recent DOE breakthroughs toward sustainably achieving and surpassing “breakeven” with respect to energy produced versus energy used to drive the experiment [5]. In fact, we might expect such a model to suggest and evaluate a range of strategies: from preventing energy-sapping instabilities, to guiding safe device and complex system operation, to expediting the design of commercially viable systems in public–private partnerships.

In addition to the missions that motivate a need for foundation models, *no entity exists—other than DOE—that is fully positioned to undertake such an endeavor* in nationally critical sciences including fusion science, nuclear stockpile stewardship, climate preparedness, bio-responsiveness, and many others, such as those described in Section 02 of this report.

DOE also has clear advantages for building world-leading applied-science foundation models by virtue of its key structural and cultural features: mature interdisciplinary teams spanning the computational and physical/life sciences, enormous amounts of data (simulated and empirical), and world-leading high-performance computing (HPC) capabilities—in full display with the >1000-person ECP. This combination of enterprise resources, human capacity, and scientific mission is unique to DOE. To create and exploit foundation models will require intentional and deliberate government investment in building the AI excellence that must accompany the teams, data, and compute power.

We detail below the current status of foundation models while emphasizing the bold, nation-scale challenges that can be launched and completed with the assistance of strong DOE investment and guidance.

2.1 State of the Art

Foundation models are being researched in a variety of scientific areas such as computer vision [6], protein structure prediction [7], organic molecules [8], biological literature mining [9], and DNA sequence analysis [10]. Simultaneously,

GAME-CHANGING OPPORTUNITY

Advances in natural language processing demonstrate the potential for generative AI models—those that can design and create systems—to be built and trained with an unprecedented scale of data and computation. GPT-3, a large language model (LLM) created by OpenAI.com, generates text based on user prompts.

Prompt: Write a paragraph on the importance of AI to the future of the United States and the Department of Energy's mission.

GPT-3 AI Model Response: *Artificial intelligence (AI) is rapidly transforming our economy, society, and environment. AI has the potential to revolutionize the way the Department of Energy works, from energy efficiency to renewable energy to national security. AI can be used to improve prediction models, automate data analysis and decision-making, and provide insights into energy markets and other complex systems. AI can also be used to develop better predictive models of energy supply and demand and to enable more efficient and cost-effective energy management. This could help the Department of Energy manage energy efficiency and renewable energy programs more effectively and efficiently, helping to reduce energy costs and emissions. AI could also help the Department of Energy monitor and protect the nation’s energy infrastructure from physical and cyber attacks. AI has the potential to be a powerful tool for the Department of Energy in its mission to secure, clean, and affordable energy for all Americans.*

in industry, several startups are heavily prioritizing these models (e.g., Hugging Face [11], AI21 Labs [12], Cohere [13], Aleph Alpha [14]); and big tech companies such as Google [15], Facebook [16], and Microsoft [17] are increasingly investing in developing and using these models in products that impact billions of people. Given the scale at which these models are being created—current foundation models have hundreds of billions, or even trillions, of parameters and they are trained on gigabytes or terabytes of data—their development is only possible for a few.

Global crises such as the COVID-19 pandemic, climate change, and international conflicts highlight the urgent need to find innovative AI-enabled solutions that can sift through vast amounts of diverse data from multiple sources to quickly find patterns, detect threats, identify correlations, make predictions, develop new vaccines and materials, and allocate and optimize resources. Furthermore, they can interact with external sources and self-update, thus taking advantage of the continuously expanding and increasing scale of diverse data sources.

As illustrated in Section 01 of the report, foundation models will be part of a constellation of interrelated technologies that will be leveraged to advance multiple aspects of national security, from material design to climate change, to healthcare, to food production, to power distribution. They have already shown promise in compressing large amounts of data and deriving new information from the knowledge they ingest. These capabilities pose an important opportunity to drive and accelerate national security, and their successful adoption will determine how quickly the United States can respond to disasters and drive economic and strategic competitiveness for the future.

The critical property of foundation models—support for many different downstream tasks, including those not initially contemplated—can be illustrated in materials science. Here, the difficulty of integrating diverse data sources (e.g., material properties, natural language, chemical structures, process flows) in forms consumable by neural networks, has confounded the adoption of machine learning (ML) in the field, because individual models can typically predict only several parameters. In 2020, scientists at Waseda University developed and applied a graph-based data representation approach to overcome this limitation, using 14 data sources from 10 individual material science project teams to train a single (large) neural network to predict more than 40 parameters [18]. More recently, a team at Microsoft Research developed a foundation model for climate and weather modeling [19]. The potential for such shared models—developed, trained, and used by dozens of scientific teams—illustrates the role that very large models can play in revolutionizing DOE's traditional modeling and simulation approaches.

Once trained, foundation models will provide new tools for rapid and targeted multimodal data acquisition, integration of private and public data, and modeling and decision support analysis within their domain. Possible sources of integrated multimodal data will include large-scale scientific experiments coupled with exascale modeling and simulation campaigns, molecular design for advanced manufacturing and drug design, public health mitigations, delivery of care, satellite imagery, communications signals, social and environmental indicators, social media data, and other large sources of information.

> *"Given this potential, we see foundation models as the subject of a growing paradigm shift, where many AI systems across domains will directly build upon or heavily integrate foundation models. Foundation models incentivize homogenization: the same few models are repeatedly reused as the basis for many applications. Such consolidation is a double-edged sword: centralization allows us to concentrate and amortize our efforts (e.g., to improve robustness, to reduce bias) on a small collection of models that can be repeatedly applied across applications to reap these benefits (akin to societal infrastructure), but centralization also pinpoints these models as singular points of failure that can radiate harms (e.g., security risks, inequities) to countless downstream applications."* R. Bommasani et al [1] on the opportunities and challenges of creating and using foundation models.

2.2 Grand Challenges

We propose two grand challenge problems that will greatly advance the state of play for AI in applied science while also capitalizing on existing DOE strengths. Together, these two challenges will advance the state of the art for foundation models in applied science, drive critical infrastructure and techniques to keep these models current and valid, and build key controls to ensure responsible use, to limit risk, and to detect security/accuracy threats. They are to:

1. **Build a set of carefully selected world-class, transformational foundation models for key scientific domains with expertise similar to that of an SME** (e.g., materials, high-energy-density physics [HEDP]). Each model instance would be a massive multi-task, broad-spectrum applied science foundation model based on a broad swath of DOE data associated with a particular set of mission challenges (e.g., biology, nuclear stockpile).
2. **Deploy an AI system leveraging multiple foundation models in conjunction with traditional scientific modeling and simulations.** This will demonstrate the end-to-end integration of this approach and would represent a blending of the standard deductive reasoning with the emerging inductive reasoning of AI methods.

Each of these Grand Challenges is discussed below.

2.2.1 BUILD A SET OF CAREFULLY SELECTED WORLD-CLASS, TRANSFORMATIONAL FOUNDATION MODELS FOR KEY SCIENTIFIC DOMAINS WITH EXPERTISE SIMILAR TO THAT OF AN SME

Foundation models excel at combining streams of disparate data to produce novel, sometimes emergent, predictions based on the inherent correlations in that data. The inductive reasoning driven by these models provides a significant complement to the traditional deductive reasoning that underpins the world of scientific modeling and simulation. The DOE hosts some of the world's most valuable and high-precision scientific data across a vast number of applications.

We propose here a grand challenge problem to build a set of digital SMEs by training foundation models for multiple domains of interest to the DOE, based on *all* of DOE's available applied-science data within each domain. These data can include simulation output from enormous supercomputer simulations, experimental data from singular experimental facilities, legacy datasets from historical experiments, simulation code and programming input from a wide array of scientific computer programs, and even the totality of a field's published scientific literature.

These rich streams of data can be combined in a uniform and distilled representation that allows a number of novel tasks to be performed by the resulting foundation model, effectively creating a digital SME that can complement existing SMEs and serve as a catalyst for knowledge transfer to early-career staff. These foundation models will also serve as master models within their domain, providing a nexus for integrating multiple related data modalities that, to date, have been largely inaccessible by significant swaths of each DOE community. Foundation models, acting as digital SMEs, can be imagined for any, and perhaps all, of the following purposes:

- Make detailed predictions of physical system evolution based on synthesis of experiment data, simulation data, and even potentially codes;
- Illuminate previously undiscovered phenomena that emerge from the integration of scales and scientific disciplines;
- Detect inconsistencies in published results and measurements based on combinations of theory and observation across disciplines;
- Enable rapid design or inverse design of new systems, devices, chemicals, materials, and processes; and
- Identify and predict rare events or anomalous behavior within complex systems.

The research and development required here is substantial. It requires national-scale investment, and leverages the DOE's workforce, infrastructure, and expertise. The challenge would require:

- Development of new methods to represent multiple disparate sources of data in a robust and meaningful way for ingestion into foundation models. To successfully accomplish this for scientific applications, the DOE will need to shift from a primary focus on data representations for languages and natural images to structures such as *K*-dimensional vector fields, large and small graphs, and a greater consideration for sparse information, all of which provide a more robust encoding of complex scientific phenomena.
- Creation of self-supervised learning tasks and supervised domain adaptation tasks for scientific and national security use cases. For each new scientific data type being included as input into the foundation model, it is necessary to develop semantically meaningful learning tasks that are elevated from the simple task of auto-encoding. For language models, the task of predicting the missing word from a properly constructed sentence is a powerful, self-supervised task that allows models to autonomously learn significant components of a language's structure, grammar, and other elements. Developing analogous tasks for material design, hydrodynamics simulations, and other tasks is a necessary, but challenging task that is uniquely suited to the multidisciplinary research expertise of the DOE and will not be adequately addressed by the commercial community.
- Advances in data curation and pan-DOE connectivity.
- Qualitative leaps in foundation model development, representation learning, and transfer learning. Key opportunities abound in the ability to develop new learning methods that can combine the current results of experiments, traditional HPC modeling, and simulation with traditional scientific literature.
- Transformations of HPC-scale computing to train and deploy models with billions to trillions of free parameters. To date, the training of foundation models is the exclusive domain of a select number of organizations and national institutes. Each model represents a heroic training run that is the culmination of hundreds to thousands of staff hours. The DOE's investment in advanced- and leadership-class exascale computing places the national labs within this domain, but it will require significant investments in workforce and research and development to make the training of these models accessible to multiple research teams—and no longer the exclusive province of "hero" runs. In short, the DOE will need to democratize exaflop days of training for deep-learning models to broaden their accessibility.
- Updates of experimental and production facilities to become AI-ready in order to integrate foundation models into their operations.

These foundation models would also require associated theory with respect to learning objectives and methods so they could be specialized for targeted and critical missions.

Specifically, while large language models have evolved into today's foundation models, outside of natural imagery and text, the existence of these foundation models for scientific and national security applications has yet to be demonstrated. The ability to develop this emergent behavior with respect to zero- or low-shot learning is unproven, and a technical moonshot will be required to demonstrate this capability across multiple domains. To begin, we recommend that models based on broad, pan-DOE data be used to create models that will be trained and fine-tuned to execute important tasks for multiple DOE domain areas.

We describe three examples that leverage considerable DOE research in AI to date: molecular design, cancer treatment discovery, and national security.

1. Foundation models specialized to molecular design

The development of a master model that has the ability to generate complex molecular compounds, polymers, crystals, proteins, or synthesizable drug compounds would be revolutionary. The demands for such a model would require new representational learning that preserves both structure and function of complex three-dimensional (3-D) objects that are governed by local properties (i.e., bonds), but also micro- and macroscopic structure (e.g., folding and periodic structure). The learning tasks that are required to train these models from a self-supervised and domain adaptation perspective are unknown, as is the most appropriate representation for developing these models. Fundamental research around the characterization of known compounds is crucial to preserving a sufficient amount of information while preserving computational efficiency.

Despite the challenges, the ability for a foundation model to generate novel compounds that are optimized for user-driven specifications and properties would have significant impact on both the national economy, but also the agility of the DOE and the U.S. Government to respond to emerging threads. Such a model would impact core components of the National Nuclear Security Administration (NNSA) stockpile stewardship efforts, lead to new drug discovery, and fundamentally differentiate advanced additive manufacturing, which would in turn have impacts across the board from national security to new designs for green energy solutions.

2. Foundation model for cancer treatment

The seminal paper on foundation models [1] uses healthcare and biomedical research as a key illustration of the opportunities and challenges for these models. Foundation models built by DOE on broad-spectrum applied-science data can be specialized by ingesting the wide array of healthcare data available across the world. They can become the central reservoir of the medical and relevant non-medical knowledge needed to integrate and connect varied disciplines and diverse sources/modalities of data, distill their information into a multifaceted representation, and—if properly developed—provide an acceptable, safe way to disseminate knowledge with a proper level of encryption and de-biasing of the data.

These models have the potential to optimize the feedback loop between healthcare experts and real-world information, leading to improved decision making. They can be finely tuned for specific tasks in healthcare and biomedicine and then used by the government (e.g., for pandemics and security), medical professionals (e.g., healthcare providers and biomedical researchers), and the public. Thus, they can support multiple points of contact that efficiently connect data, tasks, and people.

Foundation models, with their ability to integrate enormous amounts of ever-increasing multimodal data at rates surpassing human expertise, could dramatically accelerate biomedical research and facilitate more effective and efficient healthcare. Expediting the development of medicines, identifying who will develop cardiovascular disease or diabetes in advance, identifying who will benefit from a highly expensive and/or toxic and/or life-changing cancer treatment, and predicting areas of high vulnerability for the next pandemic are only a few examples that could save millions of lives and dollars, and put our country at the forefront of bio-preparedness.

3. Foundation model specialized for national security

Although foundation models are being developed in a wide range of disciplines, foundation models such as those for security, biomedicine, and healthcare also have the potential to spread harm and pose a national threat, and they should only be developed as a joint effort leveraging expertise from government agencies. For example, they could be used to counterattack a pathogen specifically engineered by AI for lethality or to target a genetic profile.

DOE labs have the secure computing power, AI and decision support expertise, and interagency collaborations in place to start unlocking capabilities across science, climate, healthcare, infrastructure, manufacturing, agriculture, development of new materials, and countless other sectors. In fact, national security requires DOE to urgently develop this capability to prepare for future societal crises.

Furthermore, partnerships must be developed with multiple healthcare, energy, and climate systems, including the national Veterans Healthcare Administration, the Centers for Disease Control and Prevention, U.S. Department of Homeland Security, U.S. Department of Health and Human Services, universities, industry, and international partners to securely curate, store, and integrate relevant data streams and to accurately quantify requirements on crisis response. This will allow them to be addressed with the interrelated healthcare foundation model, climate foundation model, materials foundation model, or smart grid foundation model, among others, as shown in Figure 2-2.

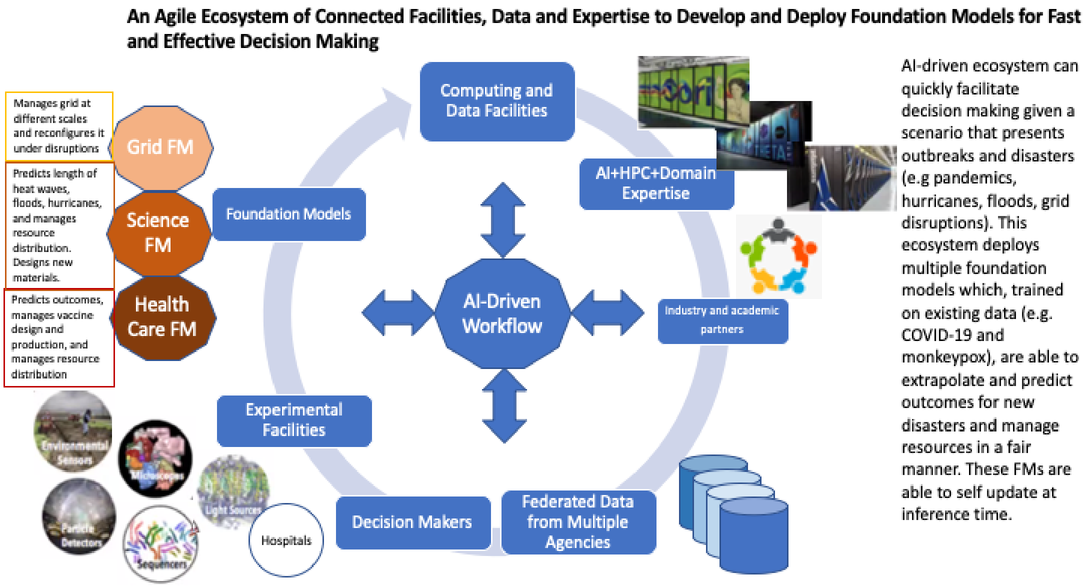

Figure 2-2. Data from multiple agencies, DOE computing and experimental facilities, and expertise connected in an ecosystem that supports the creation and deployment of foundation models that are readily available for decision making. Figure courtesy of Silvia Crivelli, Lawrence Berkeley National Laboratory.

2.2.2 DEPLOY AN AI SYSTEM LEVERAGING MULTIPLE FOUNDATION MODELS IN CONJUNCTION WITH TRADITIONAL SCIENTIFIC MODELING AND SIMULATIONS

A pilot AI system that leverages multiple foundation models should be developed in conjunction with traditional scientific modeling and simulations to demonstrate the end-to-end integration of this approach involving a blending of the standard deductive reasoning with the emerging inductive reasoning of AI methods (these concepts are detailed in Chapter 03).

The integration of emerging foundation models with existing HPC modeling and simulation approaches defines the next-generation workflow that has been described as cognitive simulation. Fundamentally, there is a significant challenge when integrating foundation models with billions to trillions of trained neural network weights, all of which can be executed at lower precision than traditional modeling and simulation workloads (ModSim). As a result, the interface between these two domains is emerging and requires a significant demonstration.

Within the cognitive simulation workflow, we expect that traditional ModSim applications will integrate responses from trained foundation models in multiple capacities, such as surrogates for in-the-loop physics calculations or generative model-driven design space exploration. In these use cases, the demands for low-latency, high-bandwidth execution of trained foundation models will become a significant percentage of the compute budget. Research into next-generation AI accelerators, novel hardware architectures, and efficient execution of large, complex neural network models will be critical to enabling these cognitive simulation workflows.

Finally, as foundation models offer an inductive approach to developing emergent behavior on new and challenging tasks, this will place an increased burden on both the interpretability and validation of trained models. All of these topics are covered in Chapters 01, 03, and 04.

2.3 Advances in the Next Decade

Over the next 1–3 years, promising data representation and self-supervised learning techniques will be identified for multiple modalities within one or two SME domain areas. Current parallel training techniques will be extended to enable scalable training of foundation models for scientific applications. Prototype, small- to moderate-scale (million- to billion-parameter) foundation models will be developed for a few domain areas. Multi-modal scientific datasets will be curated that are suitable for training combined self-supervised foundation models and supervised training for multi-task adaptation.

Over the next 3–5 years, proof of principal applications and theory will be developed for robust training of scientific foundation models. The research identified from the cross-cutting methods chapters such as Chapter 12 (Mathematics and Foundations) and 13 (AI Workflows) will be combined to improve the interpretability and robustness of model training. The size of foundation models that can be trained on a regular basis will increase.

Over the next 10 years, robust training of large foundation models will demonstrate emergent behavior across multiple modalities for several domains—essentially first-generation digital SMEs. Multiple digital SMEs will be integrated at scale into a cognitive simulation workflow to demonstrate the end-to-end coupling of both deductive and inductive reasoning approaches involving a combination of automated as well as assisted reasoning tools. Initial deployments of security applications leveraging foundation models will be made.

Over the next 20 years, there will be widespread adoption of assured, sustainable, auditable offensive/defensive applications of foundation models. Practices and tools will become well established for developing new foundation models in new domain areas, and digital SMEs will be deployed across the DOE complex.

2.4 Accelerating Development

Along with advancements in autonomous discovery, foundation models promise to deliver novel insight into complex national security and scientific areas that are underserved by traditional deductive-based modeling and simulation, because they are not well characterized by first-principles equations, are too complex to be modeled and simulated at sufficient scale and fidelity, or are subject to constraints that are not yet well understood.

Regardless of the source, development of inductive, data-driven foundation models leveraging advancements in autonomous discovery (see Chapter 03) that can act as *master models* for significant domain areas of science, energy, and security offers a revolutionary approach to engaging both the data and challenges in these fields. However, to unlock this promise, the DOE must have the ability to train and deploy these models at scale, in multiplicity, and without requiring extraordinary "heroic" effort for each and every model. In short, the DOE must democratize exascale computing for AI-ready modeling and simulation workflows as well as training and deployment of these foundation models. Furthermore, novel research and development will be required to adapt academic and industry practices to the unique science, energy, and security applications that are in the DOE's areas of stewardship.

2.5 Expected Outcomes

Much as precision medicine promises personalized medicine for each patient, the development of foundation models will offer tailored and scalable subject matter expertise in the form of digital SMEs for science, energy, and security applications.

Fundamentally, the ability to create a foundation model for a domain area will allow the DOE to bring an unprecedented amount of domain knowledge to bear on a multitude of problems. Foundation models can harness vast troves of crucial, yet unlabeled, data within the DOE complex to bring new insight to scientists, analysts, engineers, and policy makers. This insight will be available to spur U.S. economic competitiveness and accelerate the DOE's ability to discover, design, manufacture, and deploy novel innovations and new solutions to the challenges faced by our country and the world.

2.6 References

[1] Bommasani, R., Hudson, D.A., and Adeli, E., et al., 2021. On the opportunities and risks of foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2108.07258.

[2] Steinhardt, J., 2021. On the risks of emergent behavior in foundation models, Stanford University Human-Centered Artificial Intelligence. https://crfm.stanford.edu/commentary/2021/10/18/steinhardt.html, accessed December 8, 2022.

[3] Bommasani, R., and Liang, P., 2021. Reflections on foundation models, Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/news/reflections-foundation-models, accessed December 8, 2022.

[4] Jia, Y., 2019. Attention mechanism in machine translation. *Journal of Physics: Conference Series*, Vol. 1314, pp. 012186. DOI:10.1088/1742-6596/1314/1/012186

[5] Thomas, J., 2022. A shot for the ages: Fusion ignition breakthrough hailed as 'one of the most impressive scientific feats of the 21st century.' LLNL News, Dec. 14. https://www.llnl.gov/news/shot-ages-fusion-ignition-breakthrough-hailed-one-most-impressive-scientific-feats-21st, accessed Jan. 10, 2023.

[6] Chen, M., et al., 2020. Generative pretraining from pixels. In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 1691–1703. http://proceedings.mlr.press/v119/chen20s.html, accessed May 12, 2023.

[7] Rives, A., et al., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, Vol. 118, pp. e2016239118. https://doi.org/10.1073/pnas.2016239118

[8] Rothchild, D., et al., 2021. C5T5: Controllable generation of organic molecules with transformers. *arXiv*. https://doi.org/10.48550/arxiv.2108.10307

[9] Lee, J., et al., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Vol. 36, pp. 1234–1240.

[10] Yanrong J., et al., 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37, pp. 2112–2120. https://doi.org/10.1093/bioinformatics/btab083

[11] Hugging Face, undated. Transformers. https://huggingface.co/docs/transformers/index, accessed May 12, 2023.

[12] AI21 Labs, 2021. Announcing AI21 Studio and Jurassic-1 language models. https://www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1, accessed May 12, 2023.

[13] Sullivan, M., 2021. Ex-Googlers raise $40 million to democratize natural-language AI. *Fast Company*. https://www.fastcompany.com/90670635/ex-googlers-raise-40-million-to-democratize-natural-language-ai, accessed May 12, 2023.

[14] Ricadela, A., 2021. Powered by cloud, self-learning AI models are turning programming on its head. *Fast Company*. https://www.fastcompany.com/90683767/powered-by-cloud-self-learning-ai-models-are-turning-programming-on-its-head, accessed May 12, 2023.

[15] Nayak, P., 2019. Understanding searches better than ever before. *The Keyword*. https://blog.google/products/search/search-language-understanding-bert/, accessed May 12, 2023.

[16] Meta AI, 2020, AI advances to better detect hate speech. https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/, accessed May 12, 2023.

[17] Rosset, C., 2020. Turing-NLG: A 17-billion-parameter language model by Microsoft. *Microsoft Research Blog.* https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/, accessed May 12, 2023.

[18] Hatakeyama-Sato, K., and Oyaizu, K., 2020. Integrating multiple materials science projects in a single neural network. *Communications Materials*, 1, pp. 1–10. https://doi.org/10.1038/s43246-020-00052-8

[19] Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., and Grover, A., 2023. ClimaX: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343.

03. AI FOR ADVANCED PROPERTY INFERENCE AND INVERSE DESIGN

Many scientific problems can be cast as design problems, in which a model or structure is optimized to achieve certain desired behavior or characteristics. For example, the discovery of new materials that lead to solar cells with increased efficiency, better chemical processes that require less energy to produce essential industrial chemicals or can more efficiently purify water, new proteins and pathways for synthetic biology to efficiently make biofuels or new drugs, or new devices and architecture for microelectronics leading to more efficient and faster chips in products we use every day (phones, cars) can all be solved as design problems Likewise, complex experiments with many parameters can be "designed" to achieve a specific result. This process is sometimes called inverse design, whereby scientists "invert" specified (or desirable) properties to recover a new design for a complex system (e.g., a new material or a chemical process), and is fundamental in both science and engineering.

Property inference is related to inverse design, where the scientists attempt to "design" a model's parameters and properties to closely match an observation. It allows scientists to rigorously characterize a system (engineered or natural) as observed "in the wild," as opposed to its idealization obtained from purely theoretical approaches. For instance, a material property might not be directly measurable, but by "designing" a model to match an observation, the property can be inferred from data. This basis in realism allows a very direct route to predictive models and underpins fields like uncertainty quantification (UQ).

An inverse design process allows a more robust and automated approach to discovering optimal design configurations; the alternative (relying on human ingenuity and/or trial and error) does not scale to large design spaces. This means that experts must limit their design searches to relatively few possibilities. However, the use of innovative artificial intelligence (AI) methods in advanced property inference and inverse design has the potential to accelerate by orders of magnitude the traditional model computations and/or experiments that can take hours, weeks, or months.

Discussed here is an integrated roadmap for how advances in AI can enable us to model and analyze more complex systems, specifically toward creating capabilities and frameworks to drive design—and, in particular, inverse design—that can be used to advance and accelerate the science and engineering of Department of Energy (DOE) mission areas, including materials, chemistry, biology, physics, microelectronics, energy technologies, and other engineered systems.

3.1 State of the Art

Inference's fundamental goal is to reconstruct the conditions or parameters that give rise to data, observations, and/or signals. In simpler cases, exploration of this parameter space can be driven by expert scientific knowledge, by trial-and-error, or by systematic exploration using a combination of experimentation and forward models (e.g., simulations). Pushing the boundaries of modern science are cases with ever larger parameter counts, for which researchers are increasingly exploring AI approaches.

Machine learning (ML) and AI have been used for physical property inference in the sciences for some time. For example, they have been used in classification or regression based on supervised learning on simulated or experimental datasets. More recently, they have exploited deep learning to move to higher dimensionality and larger datasets [1]. However, there are limitations in the availability of data from experiments and computationally expensive high-performance computing (HPC) simulations, as well as in techniques to reliably extrapolate into new regimes that lack data. Self-supervised learning is used to address these issues in some cases [2] but more remains to be explored.

PROJECT SPOTLIGHT

Project Name: Molecular and strain design through machine learning

PIs: Hector Garcia Martin and Hans Johansen

Organizations Involved: Lawrence Berkeley National Laboratory

Goal: Create tools able to recommend molecules and bioengineered strains which meet a desired specification.

Significant Accomplishment: Created two computational tools (Macaw and A.R.T) that use novel approaches to generate molecules predicted to meet a desired property specification (e.g., a binding affinity of 50 nM which is a critical metric for the design of new medical drugs, or an octane number of 90 for the design of novel biofuels), and strain designs that optimize a desired goal (e.g., tryptophan productivity, which is critical for the economical production of animal feed).

In the News: News feature in Berkeley Lab News (https://newscenter.lbl.gov/2020/09/25/machine-learning-takes-on-synthetic-biology-algorithms-can-bioengineer-cells-for-you/), publications in *Nature Communications* (https://www.nature.com/articles/s41467-020-18008-4, Furthermore, obtaining uncertainties in inferred parameters is crucial in many areas of science, with newer methods being applied, including Bayesian deep learning [3] and likelihood-free inference [4].

Inverse design is a relatively new research direction in science and engineering. Unlike widely used design processes, which are often driven by human knowledge and intuition of the science and engineering problem, the goal of inverse design is for a scientist or process engineer to specify desired properties, behavior, or performance and then find the best available solution using an optimization algorithm that explores the design space. Current inverse design approaches rely on a random or pseudo-systematic search of the design space or a slightly more systematic search through the parameter space, for instance using genetic algorithms or Monte-Carlo Tree Search (as used by Google LLC in the AlphaFold [5] protein folding project). In addition to protein folding, early efforts to utilize inverse design have been explored in a variety of fields, from materials and chemical research to engineering and high energy density physics. These approaches are executed in an automated way until a solution that meets the specified target is found. A critical aspect of inverse design is that the assessment of a potential solution is fast, so that a large parameter space can be explored rapidly and accurately.

A fundamental obstacle to broad application of inverse design is that many scientific data are expensive to acquire (for instance a large computer simulation or costly experiment). Additionally, this process can be noisy and involve dozens or hundreds of parameters. By automating, expanding, and accelerating the design-search process, advances in AI, and its concentrated application to the problem of inverse design, could have far-reaching implications and transform several important science, engineering, energy, and security DOE missions.

1. **Creating Biomolecules On-Demand.** In nature, biological systems such as microbes or cells- have the innate ability to efficiently create new materials and molecules or absorb hazardous materials for internal use. *AI-enhanced inverse design could achieve full control of biological systems and harness their abilities for producing new products or converting waste products.*

   Industry has used the natural abilities of biological systems to make products for over a century. For example, industry has exploited the fermentation process for penicillin production. The birth of biogenetic engineering opened doors to modifying microbes or cells to create new materials and molecules not naturally produced but desired by industry at a higher yield or reduced cost. Examples for DOE missions include the conversion of plant-based materials into ethanol or fuels, and biodefense against new pathogens for national security. Biological systems can also be engineered to extract hazardous materials, such as radioactive materials or toxic metals, or to break down plastics. Modifying biological systems is extremely complex and requires a detailed understanding and precise control of a system's functions. Models of biological systems are complex, often incomplete, and computationally expensive. The potential efficacy of AI and inverse design in biomolecular design has been show in recent work, for example generating molecules [6] and functional protein sequences with specific target properties [7]. Engineering the desired behavior of a biological system requires the exploration of an exponentially large parameter space. Here, AI-enabled forward acceleration through inference and inverse design has the potential to dramatically advance the field [8].

2. **Materials by Design.** Developing next-generation materials is critical to DOE missions, national security, and U.S. technological competitiveness and leadership. *AI-enhanced design that harnesses the physics of materials would greatly accelerate the discovery and production of new materials with required properties*.

   The U.S. can increase its competitive advantage by designing materials that (1) eliminate use of rare and expensive elements not readily available domestically, (2) can easily be manufactured and require minimal energy sources, (3) are readily recycled and upcycled as part of the material lifecycle, and/or (4) have properties that are optimal for specific applications, from energy storage to safer explosives. Machine learning and simple inverse design approaches are already being explored for specific problems in materials research [9, 10]. For example, there are 10^{50} possible combinations to create alloys from nickel, ion, cobalt, and copper. Such materials are needed in multiple variations, from steel and concrete to plastics and catalysis used in the manufacturing of chemicals and fuels. Materials research and development, engineering, and manufacturing form a vast combinatorial design space that overwhelms current practices involving large-scale simulations and time-consuming experiments. The scientific community and industry have adopted AI approaches to accelerate the simulations of specific properties through inference. However, significant challenges remain in the adoption of AI, such as accuracy in the face of limited training data and the need for AI models that accurately describe many properties simultaneously. Exploration of the use of inverse design methods accelerated by simulation and AI inference models is typically limited to a single property in a given design space. The key to accelerating the design of new materials lies in developing new inverse design approaches that can explore the full design space, from fundamental materials properties to stability and manufacturability.

3. **Design of Microelectronics.** Building next-generation microelectronics also requires the exploration of an immense design space of relevant physical parameters.

These range from the choice of materials and quantum properties to macroscopic 3D geometry and continuum (electromagnetic) behavior to practical factors such as manufacturability [11]. *AI would drastically accelerate the optimal design and production of novel but practical microelectronics in critical mission areas*.

Examples include designing for radiation hardness in aerospace applications and nuclear deterrence and designing improved materials or geometric design for core components (such as memory, 6G communication, or computing chips). Fabrication of each microelectronics design is expensive and time-consuming, which limits opportunities for experimentation. Moving beyond the limits of human knowledge and the timeframes and costs of these design processes will demand AI models that accurately infer properties of microelectronic systems at multiple scales, from the atomic and microscopic scales to the macroscopic scale involving complex multiphysics behaviors. Combining AI models that accurately and rapidly infer essential and often multiscale properties of the microelectronics system with inverse design approaches will be imperative if the U.S. is to maintain leadership in the design and manufacturing of next-generation microelectronics.

4. **Stockpile Modernization and Nuclear Deterrence.** Developing and maintaining a safe, secure, and operationally ready nuclear stockpile for nuclear deterrence is of utmost importance to national security. *AI-enhanced design would enable a fast modernization program that minimizes cost, is safe, greatly reduces the time from design to "First Production Unit," and is capable of rapidly responding to evolving threats.*

   Without a return to nuclear testing and with only limited experimental data upon which to build models, much of the design optimization of nuclear weapons, their manufacturing, and stability during environmental conditions over decades relies on large and complex multiscale and multiphysics simulations commanding large HPC resources, including exascale computing platforms. These simulations must be highly accurate, with extremely well-defined and well-understood error bounds to ensure that new designs pass the extremely rigorous certification and qualification processes for nuclear technology. High accuracy and fidelity from property inference will be essential to either replace or accelerate traditional large-scale simulations. Novel inverse design approaches will be needed to guide and speed up the complex multi-scale multiphysics design optimization processes inherent to nuclear stockpile modernization.

5. **Accelerating Manufacturing with Automated Design.** Rapid and innovative design and the manufacturing of complex systems or machines are critical to ensure U.S. competitiveness. Decision-making in design and manufacturing is driven by human experience and knowledge and supported by computer simulations and experiments [12]. *AI methods could significantly accelerate the production of complex systems by automating design and manufacturing processes*.

   For example, next-generation hybrid or electric aircraft with intercontinental range would require radically new designs, yet the design and manufacturing of a new type of aircraft can take 10 to 15 years. Additive manufacturing (also known as 3D printing) has been transforming the industrial production of parts, enabling rapid prototyping, on-demand manufacturing, and creation of new parts that would not be manufacturable with traditional tooling methods. The aviation industry is an early adopter of additive manufacturing, as are the automotive and renewable energy industries. The DOE scientific community is also using additive manufacturing to build innovative new instruments.

   Utilization of AI with uncertainty quantification to better explore the design space with higher fidelity has the potential to accelerate these optimal design and manufacturing processes. For example, accelerating expensive computational fluid dynamics simulations with inference from accurate AI models—critical in many industrial design processes—will lead to reduced costs and a faster time-to-solution. Inverse design is critical to support and substantially accelerate the decision-making process in the design and manufacturing of complex systems, and it will help determine what is printable with additive manufacturing for aircraft and other machines. Manufacturing through AI requires a new breed of AI methods that are physics-informed, optimization-aware, capable of mitigating uncertainty, computationally efficient, able to address calibration through online experimental and field data, and advanced enough to enable system-level, automatic decision-making.

6. **Robust Energy Infrastructure.** Our nation is highly dependent on reliable and secure energy supply. *An AI-designed comprehensive and accurate model of the nation's energy infrastructure would enable the U.S. to make critical decisions in real-time, as well as develop medium- and long-term policy decisions that will ensure a stable energy supply now and in the future.*

   The energy infrastructure is one of the most complex ever built, and it must be managed today in terms of its unprecedented spatial and temporal extent, complexity, and interconnectedness of energy generation, storage, and transmission capabilities. Despite this interconnected complexity, control and information exist only at the local or regional levels, with limited sharing of information. Large-scale simulation models are used for decision support and control at the regional level, but even with exascale computing systems, large-scale heterogeneous coupled systems models encompassing the national grid remain intractable computationally. Moreover, the

U.S. energy infrastructure is rapidly changing in fundamental ways, becoming more complex as new types of energy sources are integrated into the system, such as large- and small-scale renewables, variable energy resources (e.g., wind and solar), and energy storage, including batteries, and buildings that shift loads and feed renewable energy back to the energy infrastructure.

This complexity, in combination with the increased prevalence of extreme weather events and the inadequate tools to manage, monitor, and control these systems, is leading to more frequent disruptions of our energy supply. The status quo—even apart from rapidly increasing complexity and disruptors—leads to poor, costly decision making, wasted resources, slow recovery from events, suboptimal planning of new resources such as additional energy storage, and greater susceptibility to catastrophic disturbances [13]. It also prevents decision makers from developing effective strategies toward a carbon-neutral energy infrastructure. Advances in AI are needed to replace larger simulation models with inference from AI models and support real-time decision and control through inverse design and optimization processes, while also integrating inherently multi-modal, heterogeneous, and rapidly growing data from the energy infrastructure into (global) energy infrastructure models with high fidelity to provide trustworthy predictions. Inverse design based on accurate inference AI models will allow stakeholders to make informed decisions, leading both to more stable energy infrastructure and to lower energy costs.

7. **Intelligent Water and Agriculture Infrastructure.** Water and agricultural products are the backbone of the U.S. They form an intricately connected complex system that is poorly coordinated due to the large numbers of stakeholders, distributed data with different ownership, and lack of spatial and temporal resolution needed for decision-making. *An AI-enabled global model of the water infrastructure that incorporates predictions of weather extremes and trends from climate models would support both short-term and long-term informed decision-making.*

   Such a model would serve as a building block for agriculture models and be of critical value to the energy infrastructure component that relies on hydropower. Yet no comprehensive water model exists today. Essential U.S. water data are fragmented, undermining our ability to effectively plan and act in the near- and long-term. Large climate models use supercomputing resources and could be accelerated with inference from AI models to allow for time-sensitive and decadal modeling to support decision-making and planning. Inverse design and related optimization approaches could be developed to optimize waterflows across the country and minimize resources needed to grow crops. Water-climate models incorporating AI could integrate data from heterogeneous data sources, satellites, sensors, and simulation models. This kind of water-climate model would be instrumental in the optimization of agricultural processes and the development of actionable precision agriculture (e.g., optimized seed placement for maximization of production, minimization of fertilizer use, and the reduction of waste). The water and agriculture infrastructure are emergent fields when it comes to the use of AI, with some early demonstrations underway [14].

3.2 Grand Challenges

Across the domains outlined above, vast and accurate inference from AI models will be critical to rapidly exploring the high-dimensional parameter spaces in design and operation of complex, multi-scale systems. Property inference from AI-based surrogate (Chapter 01) or foundation models (Chapter 02) is a fundamental building block of AI as well as a key ingredient for inverse design. Next-generation inverse design methods will rely on accurate and trusted AI models to accelerate and optimally search the parameter space for property optimization or decision-making.

Six key grand challenges that need to be addressed to enable mainstream adoption of AI for inverse design are outlined below.

1. **Inference with High Accuracy and Uncertainty Quantification.** Highly reliable inference of properties from trained models is an essential requirement for wide-scale adoption of AI in science and engineering [5]. *The grand challenge is to be able to build next-generation trust-worthy AI models that respect the accuracy of underlying experimental or computational training data and reliably provide uncertainty estimates of their predictions*. Such models can then be used as trusted sources for the simultaneous inference of multiple properties needed to meet design requirements. AI models that quantify uncertainty and fidelity of inferred properties, including at the edges/tails of the parameter space for which the model has been trained, will be vital to earning confidence and trust from industry adopters and regulators.
2. **Learning with Limited, Heterogeneous Data.** In many critical science and engineering domains, the data available to train AI models for property inference are limited, often spanning only subsets of the desired parameter space. Additionally, the data are heterogeneous, created by physical experiments, observations, and computational simulations (each with their own uncertainties) that must be integrated. For example, only a small fraction of the possible biomolecules and materials have been studied experimentally or computationally, but one would want the AI model to generalize across the whole structural space. *The grand challenge is to build accurate AI models for property inference that require minimal information to learn,*

incorporate domain knowledge, and seamlessly assimilate diverse datasets.

3. **Adaptive Learning with a Deluge of Heterogeneous Data*.*** Some science and engineering domains are confronted with a deluge of multi-modal data, which places very different constraints on the training of AI models for the inference of properties. For example, the energy infrastructure is instrumented with multitudes of sensors that are distributed across the country and that rapidly generate and return massive amounts of data across diverse temporal and spatial scales. These same challenges also manifest in water and agriculture systems, as the growing need for intelligent decision-making drives the need for more detailed monitoring and data integration. Furthermore, with the arrival of exascale computing, large-scale simulations have the ability to rapidly generate petabytes of data. *The grand challenge is to acquire, secure, curate, and contextualize data to train or update AI models needed for accurate property inference in real-time*. Solving this grand challenge will be essential for science and engineering to take full advantage of AI.

4. **Physics-Constrained Inference across Scales.** Many of the complex systems referenced are driven by multiple fundamental governing equations, predominantly physics, which span many spatial and temporal scales. For example, inference models describing the nuclear stockpile or microelectronics cover length scales from atoms to the whole system and timescales from milliseconds to days. *The grand challenge is to ensure AI models infer information consistent with governing equations, such as the laws of physics. Advances are needed to properly and simultaneously account for physics and multiphysics constraints across a hierarchy of scales.* An additional advantage of physics-constrained AI is that it will reduce the parameter space that must be explored for optimal or inverse design. This will be of great benefit for domains with scarce data for inference model training. The development of AI approaches that can self-learn the behavior of the laws of physics across multiple scales will be important for systems where the coupling of mathematical equations at different scales is not well defined, but where the flow of information across scales is critical for the accuracy of the model.

5. **Explainable, Interpretable and Trusted Inference.** Most of the AI models for property inference currently available are “black box” in that neither its developers nor users can explain why the model arrived at a specific decision. *The grand challenge is to build AI models that are explainable and interpretable, and in which humans can understand the decisions, predictions, and inferences made by the AI model as well as quantify the trustworthiness of the AI model for a given problem.* Trustworthy and understandable AI inference and predictions will drive actionable design processes and decisions for domains such as precision agriculture and energy system control [15], but only to the extent that trust in AI models can be reliably quantified.

6. **Inverse Design in Complex Design Spaces with Actionable Outcomes.** Fast and accurate inference with AI models will markedly accelerate the search for optimal designs in science and engineering domains, particularly where the parameter optimization space is exponentially large. *The grand challenge is to develop methods that rapidly search large parameter design spaces in a systematic and rational way, supported by domain or physics knowledge.* Solutions for this grand challenge should be capable of handling complex, often competing objectives and constraints, such as desired material property versus manufacturability, cost, safety, and recyclability. That is, advanced AI inverse design systems must make decisions based on the full process cycle. This will require the development of representation learning to constrain and create more flexible design spaces and novel reward functions capable of handling the complexity and interactions among diverse requirements, from physics to business constraints.

3.3 Advances in the Next Decade

To tackle the grand challenges of advanced property inference and inverse design, significant investment and progress in AI mathematics, algorithms, software, and infrastructure is required in multiple, cross-cutting technology areas.

1. **Mathematics and Algorithms.** The most important requirement for widespread adoption of advanced property inference and inverse design methods is the ability to build highly accurate uncertainty-aware AI-based surrogate (Chapter 01) and foundation (Chapter 02) models. Mathematical approaches and algorithms will need to be developed to seamlessly merge diverse heterogeneous datasets and train AI models that achieve the desired accuracy. New developments are needed to integrate uncertainty quantification with property inference to enable reliable decision-making for control and design.

   DOE science, engineering, energy, and security mission areas generate data from multiple modalities, with vastly different acquisition rates and fidelities [16]. For many mission areas, the available data are limited, and the rate of data generation is low. This sparsity of data drives the need for new ML algorithms that can infer accurately from minimal information and can be rapidly updated when new data become available through active or adaptive learning approaches.

   Significant development efforts will be needed for data representations and AI models that can properly encode and operate across multiple length and time scales, including the hierarchical and multiphysics information characteristic of DOE mission areas. Finding optimal representations is prerequisite to making AI models explainable, interpretable, and trusted by humans. Explainable AI (XAI) has been pursued by the Defense Advanced Research Projects Agency (DARPA) [17] as well as the National Science Foundation (NSF) [18, 19], yet with only partial overlap with DOE mission areas. Major advances are also critically important in the ability of AI models to discover data representations themselves, and to adapt the model in concert with the data representation [20].

New mathematical approaches and algorithms are also needed to enable inverse design approaches that can explore the design/parameter space rapidly and intelligently, producing optimal solutions and control decisions at scales not presently tractable. Optimal representations, combined with the integration of domain knowledge, can create essential constraints and flexible design spaces. New approaches are needed to design complex reward functions for AI system optimizers that take into account often conflicting constraints ranging from desired properties to manufacturability, cost, and safety.

2. **Data Infrastructure.** Progress in the DOE science, engineering, energy, and security mission areas increasingly requires large multi-disciplinary teams at experimental facilities, in the field (e.g., energy, water, and agricultural infrastructures, or urban integrated field laboratories), and/or at computational centers. These teams and facilities create complex multi-modal datasets with hundreds of different data types, with varying size, and many acquisition rates. New development is needed to create data infrastructure that can acquire, curate, and manage this data in an automated fashion [21]. It will be important to develop comprehensive and inclusive data standards that can facilitate the integration of these diverse data sources into training sets for AI models, such as the surrogate and foundation models discussed in Chapters 01 and 02.

   For many mission areas where movement of data is precluded, data privacy and security are important. Advances in the development of federated data and learning systems will be needed to address these challenges.

3. **AI Software and Workflows.** Development will be needed to develop robust, modular, composable software and workflow components that can manage evolving, heterogeneous datasets that are inherently distributed and, in many cases, constrained in movement and access by privacy and security requirements. Advances are needed to ensure that robust workflows can incorporate active/adaptive learning within such ecosystems and integrate advanced property inference and inverse design approaches with uncertainty quantifications to analysis.

4. **AI Hardware.** A major driver for the inference and inverse design building blocks driving DOE science, engineering, energy, and security mission areas is the need for data infrastructures and workflows that can leverage near real-time performance of emerging hardware infrastructure. New hardware components must be explored and developed to rapidly and continuously ingest data from multiple modalities, update AI models, and provide real-time inference for inverse design, decision-making or high-speed control of complex systems. Research is needed to evaluate AI hardware accelerator technologies on data acquisition latencies and time-to-solution for AI model training and inference.

3.4 Accelerating Development

Pilot projects over the next decade will drive the development of and demonstrate the utility of newly developed mathematics, algorithms, and data, software, and hardware infrastructure. The success of these pilots will provide a framework for advancements in other DOE mission areas.

Pilots that could be used to accelerate progress with respect to the grand challenges and advances needed in the next decade are described below.

1. **Rational Design in Biochemistry, Chemistry, and Materials.** Several national laboratories have projects developing ML approaches for materials and biochemical process discovery, with singular or narrow application/property areas. A pilot in each of the biochemistry, chemistry, and materials domains will drive urgently needed progress in adaptive learning from multi-modal data and build foundation models for inferring a wide range of properties with experimental accuracy; develop models that span multiple application domains and couple these models with inverse optimization approaches for end-to-end rational design.
2. **Automated Design and Optimization of Engineered and Manufacturable Systems.** An initial series of small-scale pilots should focus on the development of physics-informed AI models that are optimization-aware, capable of mitigating uncertainty, computationally efficient, able to address calibration through online experimental and field data, and capable of enabling system-level algorithms for key application targets. Building on the accurate, trustworthy, proof-of-principle AI models resulting from this pilot, the next phase will be to enable the automated design of specific engineered and manufacturable systems.
3. **AI for Energy Resilient Infrastructure.** This DOE mission area can be used to demonstrate AI inference at multiple time and length scales as well as the integration of various models for optimization with inverse design approaches. A pilot should build a carefully selected set of AI-based surrogate and foundation models for control,

optimal design, and inverse optimization at scales not tractable by current energy system models. Such models should integrate transient models in decadal design for cost efficiency, resilience, and reliability. The AI models developed by this pilot should capture the challenge of utilizing high order multi-modal datasets (>100 different data types) from local, regional, and national levels. Trust, data privacy and security, federated data approaches, and real-time data generation, such as from increasingly powerful measurement systems using edge computing, should be considered in the AI model development.

4. **Resilient Water and Agriculture Resources.** No comprehensive regional, much less national, water model exists, and such types of models would require integration of fragmented, distributed data sources. An initial pilot should focus on the development of a federated data capability with workflows and data formats from multimodal data, providing mechanisms to ensure data security and privacy. This data capability should in turn be combined with the development and training of a comprehensive and dynamic AI water model such as at the scales of the major metropolitan areas targeted by DOE's Urban Integrated Field Laboratories program. The outcomes of this water model, which should be trustworthy, will form a foundation for the next series of pilots and the integration of water and climate models for optimization and decision-making on water management, such as to address severe weather-related flooding in vulnerable urban communities. Eventually, this model should also enable the integration of agriculture models to design precision agriculture strategies that optimize resource utilization.

3.5 Expected Outcomes

Advanced property inference and inverse design are essential components to accelerating design and optimizing control in science, security, engineering, and manufacturing with AI. Advanced AI models that are accurate, optimally use available data, and are explainable and trustworthy will be interrogated and used for "what if" scenarios in design processes and decision-making. Inverse design with reduced (or even without) humans-in-the-loop interaction will have a transformational impact on the U.S. economy as it will accelerate the development of new products and processes both directly and by enabling critical infrastructure—particularly energy—to operate more reliably, with greater resilience, and at lower cost.

The use of AI will accelerate the pace and turnaround of the design of new, sustainable products and processes with greatly reduced cost. This will secure U.S. leadership in key economic growth areas spanning biomolecular and materials engineering and manufacturing. It will also help the nation secure a modern and safe nuclear stockpile, rapid response capabilities against future biological threats, and stable and integrated electric, water, and agriculture infrastructures.

3.6 References

[1] Carleo, G., et al., 2019. Machine learning and the physical sciences. *Rev. Modern Phys.* 91(4), 045002. DOI 10.1103/RevModPhys.91.045002

[2] Hayat, M.A., Stein, G., Harrington, P., Lukić, Z., Mustafa, M., 2021. Self-supervised representation learning for astronomical images. *ApJL* 911(2), L33. https://doi.org/10.3847/2041-8213/abf2c7

[3] Charnock, T., Perreault-Levasseur, L., Lanusse, F., 2022. Bayesian neural networks. In *Artificial Intelligence for High Energy Physics,* pp. 663–713. https://doi.org/10.1142/9789811234033_0018

[4] Cranmer, K., Brehmer, J., Louppe, G., 2020. The frontier of simulation-based inference. *PNAS*, 117(48), 30055-30062.

[5] Jumper, J., Evans, R., Pritzel, A. et al., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature,* 596, pp. 583–589. https://doi.org/10.1038/s41586-021-03819-2

[6] Blay, V., Radivojevic, T., Allen, J.E., Hudson, C.M., Garcia Martin, H., 2022. MACAW: An accessible tool for molecular embedding and inverse molecular design. *J. Chem. Inf. Mod,* 62(15), pp. 3551-3564. DOI: 10.1021/acs.jcim.2c00229

[7] Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos Jr, J.L., Xiong, C., Sun, Z.Z., Socher, R., and Fraser, J.S., 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp.1–8.

[8] Volk, M.J., Lourentzou, I., Mishra, S., Tung Vo, L., Zhai, C., Zhao, H., 2020. Biosystems design by machine learning. *ACS Synth. Bio.* 9(7), pp. 1514–1533. DOI: 10.1021/acssynbio.0c00129

[9] Alberi, K,. et al., 2019. The 2019 materials by design roadmap. *J. Phys. D: Appl. Phys.* 52, 013001. https://doi.org/10.1088/1361-6463/aad926

[10] Choudhary, K., DeCost, B., Chen, C., et al., 2022. Recent advances and applications of deep learning methods in materials science. *npj Comput Mater,* 8, 59. https://doi.org/10.1038/s41524-022-00734-6

[11] DOE ASCR Report, 2018. *Basic Research Needs for Microelectronics*, Oct. 23–25, https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN_Microelectronics_rpt.pdf, accessed May 12, 2023.

[12] Max Mowbray, M., Vallerio, M., Perez-Galvan, C., Zhang, D., Del Rio Chanona, A., Navarro-Brull, F.J., 2022. Industrial data science – a review of machine

learning applications for chemical and process industries, *React. Chem. Eng,* 7, pp. 1471–1509. DOI: 10.1039/D1RE00541C

[13] DOE Office of Electricity Report, 2019. *North American Energy Resilience Model,* July https://www.energy.gov/sites/prod/files/2019/07/f65/NAERM_Report_public_version_072219_508.pdf, accessed May 12, 2023.

[14] IAW Report, undated. *Digital Water: Artificial Intelligence Solutions for the Water Sector*, https://iwa-network.org/wp-content/uploads/2020/08/IWA_2020_Artificial_Intelligence_SCREEN.pdf, accessed May 12, 2023.

[15] Freiesleben, T., König, G., Molnar, C., Tejero-Cantero, A., 2022. Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. arXiv:2206.05487 [stat.ML]. https://doi.org/10.48550/arXiv.2206.05487

[16] DOE ASCR Report, 2019. *Data and Models: A Framework for Advancing AI in Science*, Dec. 16. https://www.osti.gov/biblio/1579323, accessed May 12, 2023.

[17] Gunning, D., Vorm, E., Yunyan Wang, J., Turek, M., 2021. DARPA's explainable AI (XAI) program: A retrospective. *Appl. AI Lett.* (2)e61. https://doi.org/10.1002/ail2.61

[18] McGovern, A., 2021. NSF AI institute for research on trustworthy AI in weather, climate, and coastal oceanography. *AI Matters*, *6*(3), pp. 14–16.

[19] Bates, J., 2021. Expanding the geography of innovation: NSF AI Research Institutes 2021. *NSF Science Matters*. https://beta.nsf.gov/science-matters/expanding-geography-innovation-nsf-ai-research, accessed May 12, 2023.

[20] DOE ASCR, 2019. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*, PRD #2, https://www.osti.gov/biblio/1478744, accessed May 12, 2023.

[21] DOE ASCR Report, 2021. *Toward a Seamless Integration of Computing, Experimental, and Observational Science Facilities: A Blueprint to Accelerate Discovery*, March 8. https://www.osti.gov/biblio/1863562, accessed May 12, 2023.

04. AI-BASED DESIGN, PREDICTION, AND CONTROL OF COMPLEX ENGINEERED SYSTEMS

Complex engineered systems refer to systems designed and constructed by people and comprising many subsystems whose behaviors cannot be separated or isolated from the full system or the environment in which the system operates. Complex systems are characterized by nonlinearities, interactions, connected multiscale components, sensitivity to initial conditions, and emergent behaviors [1].

Complex engineered systems underpin much of the critical infrastructure in the U.S., ranging from the energy network (encompassing power generation, distribution, storage, and consumption) to transportation systems to supply chains. Complex engineered systems can also include those that must operate in and/or interact with complex environments, such as autonomous vehicles, the U.S. Department of Energy's (DOE's) leadership-class computing environments, nuclear power (including fusion power) systems, large-scale scientific instruments (such as light sources or accelerators), advanced manufacturing facilities [2, 3, 4], and advanced turbine engines. Biological systems display many of the same behaviors and challenges as complex engineered systems and are covered more extensively in Section 02: Domains.

4.1 State of the Art

Changing contexts in recent years (for example, changing climate conditions, evolving population dynamics, and water availability) are creating stresses in the nation's critical infrastructure. There is a real and urgent need to understand the impacts of these changes on our infrastructure and to quantitatively assess and deploy solutions (e.g., control systems) to mitigate current and projected negative impacts—while being flexible enough to adapt to future demands on these complex systems. These infrastructure systems must be designed to incorporate new capabilities based on predicted changes (prognostics) as well as observed status. While the scale of the challenges may differ from system to system, each system faces a growing number of demands beyond those that the system was initially designed to support.

Three examples of complex engineered systems illustrate their centrality within the DOE mission: the electricity grid, including generation, distribution, storage, and consumption; large-scale science facilities, including high-performance computing and accelerator facilities; and magnetic confinement fusion (tokamak) reactors.

1. **U.S. power infrastructure.** Electricity is a visible and critical part of the U.S. power infrastructure and includes generation, distribution, storage, and consumption. Key components of this infrastructure are highlighted below. The power (electricity) grid integrates these components and is designed to deliver electricity reliably from generators to consumers.

   a. **Distribution.** The grid must balance base load generation—each source being a complex engineered system in its own right—with load demand at all times, with very limited capabilities for centralized control. Effective grid management must integrate capabilities across many scales, from individual homes (including electric vehicles, home energy storage, and micro-generation) to large-scale generation facilities. The grid must also respond to wide variations in demand and changing environmental conditions while also being robust and resilient to disruptions, such as cyber threats and space weather (Figure 4-1). Today's electricity grid is aging, which presents new opportunities to

PROJECT SPOTLIGHT

Project Name: Deep learning progress in fusion research

PI: William Tang

Organizations Involved: Princeton University, Princeton Plasma Physics Laboratory

Goal: Deliver AI/HP–enabled advanced warning for avoidance/mitigation of dangerous disruptions before critical damage can be done to the international burning plasma experiment scheduled to begin in 2028 for the International Thermonuclear Experimental Reactor.

Significant Accomplishment: We've moved beyond passive prediction of disruption for huge observational databases to active control; the key to the advance is the introduction of an innovative deep-learning surrogate model capable of carrying out validated first-principles simulations as a "real-time simulator" leading to a "digital twin" for tokamaks.

In the News: The PI received the 2018 NVIDIA Global Impact Award with citation "for groundbreaking work in using GPU-accelerated computing to unleash deep learning neural networks for dramatically increasing the accuracy and speed in predicting dangerous disruptions in fusion systems" and subsequently published the work in Nature (Kates-Harbeck, J., Svyatkovskiy, A., and Tang, W., 2019, Predicting disruptive instabilities in controlled fusion plasmas through deep learning, Nature, 568, 526–531, https://doi.org/10.1038/s41586-019-1116-4).

incorporate advanced artificial intelligence (AI) capabilities, including advanced system designs, new modeling and prediction capabilities, AI-based control and decision making, and improved robustness to severe disruptions—whether related to natural forces (e.g., weather, wildfire, earthquake, flood) or human adversaries. Additionally, the underlying composition of the electricity grid is evolving, for example, through the addition of distributed generation (e.g., private photovoltaic panels) and storage (e.g., whole-home batteries), as are the dynamics of use (e.g., electric vehicles, their associated charging loads, and capacity to support storage for other uses). These facets of the evolution of energy technology bring even greater complexity and nonlinearity to the electricity grid.

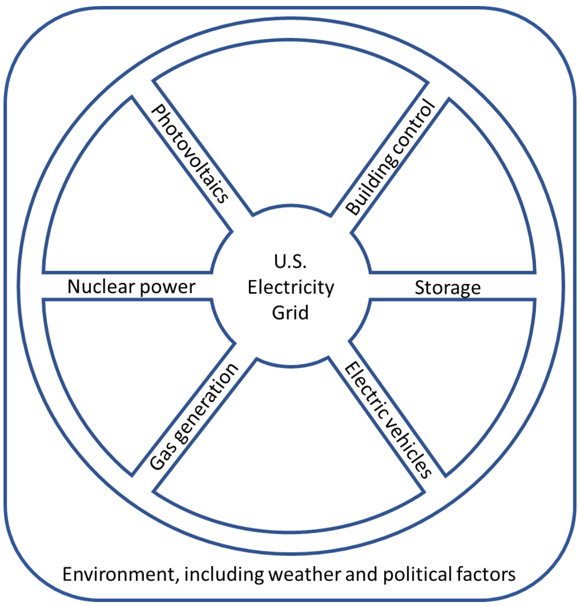

Figure 4-1. AI-based design, prediction, and control are critical for the U.S. electricity infrastructure.

b. **Generation.** The electricity infrastructure also includes inertial-based generation, such as nuclear power plants, and inverter-based generation, such as wind and photovoltaics. Baseload generation systems (fission, gas turbine)—operating as peaking systems—require new control paradigms to maximize efficiency and economics. Each of these generation systems (subsystems within the overall electric infrastructure) is a complex system in its own right. That is, they exhibit nonlinearities, with feedback both internal to the system and through its interaction with the grid.

   Nuclear power plants are one of the few zero-carbon options for electricity generation that can also support process heat applications. As such, nuclear power systems are important for district heating, water desalination, and hydrogen production. Future nuclear power systems include microreactors, which are expected to solve local or regional short-term energy needs, such as for electric power after natural disasters or at forward operating bases. AI systems, such as inverse design and property inference (Chapter 03), will be critical for the design of future plants, as will surrogate and foundation models (Chapters 01 and 02) for control and maintenance. These AI capabilities will also be critical in extending the life of the current nuclear power facilities through better prognostics and monitoring of health, fuels, and components. These and similar challenges are predominantly driven by the high cost of construction and operation, along with lifecycle issues such as the disposition of used fuel. Individually and collectively, these challenges affect the nation's ability to advance and rapidly deploy future nuclear power systems, including small modular reactors and microreactors. They also present opportunities to incorporate advanced AI-based capabilities across the nuclear power lifecycle (from design through licensing, construction, operations, and maintenance, to decommissioning and fuel disposition).

   Inertial-based generation also encompasses gas turbines. In these systems, the occurrence of low-probability but high-impact rare (abnormal) events poses critical challenges to performance and reliability. For example, in these energy systems combustion instabilities (such as lean blowout, flashback, and thermoacoustic instabilities) can cause catastrophic failure and damage. These risks are further exacerbated when gas turbine engines are operated on low/zero carbon fuels (hydrogen, sustainable aviation fuels) as opposed to regular jet fuels or natural gas. It is of great importance to understand and predict such rare events in order to avert their occurrence.

   AI offers capabilities for automated discovery and assessment of the underlying precursors and causalities governing rare events encountered in energy systems. Such capabilities are essential for the development of prognostic and control strategies to enable safe operation of these engines in high-efficiency mode while preventing rare combustion events.

c. **Consumption.** Advanced heating and cooling systems for high-performance buildings have become increasingly complex. Decarbonizing them requires new design and operational approaches that must be deployed in millions of buildings. Transformational approaches are needed as buildings move from being passive energy consumers to being not only active consumers in a dynamic energy market—in which they need to provide reliable and dispatchable load flexibility to the grid—but also active prosumers—a role where they provide heating, cooling, and energy storage for district-scale systems that integrate buildings, manufacturing infrastructure, and mobility systems.

Artificial intelligence methods and approaches discussed here and in previous chapters have the potential to streamline design approaches, support system-aware operational optimization, and automate deployment of advanced analytics and control methods.

2. **Large-scale science facilities.** DOE's Office of Science is responsible for designing, building, and operating large-scale facilities for scientific discovery. Such facilities include, for example, leadership (high-performance) computing facilities and their internal and international connectivity; accelerators; light sources; and instruments, facilities, and field laboratories supporting nanoscale science (e.g., electron microscopes), bioscience (e.g., plant phenotyping), and earth systems monitoring.

    a. **High-performance computing (HPC) facilities and their connectivity.** AI capabilities described throughout this report will rely on next-generation HPC capabilities for scientific and engineering research and development. The current exaflop-class facilities require millions of dollars of infrastructure investment, tens of megawatts of energy for power and cooling, and include millions of electronic components ranging from computational cores and accelerators to storage devices to communications infrastructure. Overlying these components are millions of lines of software, including complex application codes, operating systems, runtime systems, input/output controllers, workflow frameworks, data management utilities, and scientific simulation models. These components are interconnected at every scale, from the system networks that interconnect processor nodes to machine-room networks integrating storage to the national ESnet infrastructure and its domestic and international connections. Current challenges for AI in managing HPC facilities include performance modeling, performance optimization, scheduling, power management, prognostics and maintenance management, and proactive resource management. The critical importance of each of these will increase significantly as the next generation of HPC facilities is increasingly tightly integrated over campus and wide-area ESnet networks (themselves complex engineered systems) with edge devices, including scientific instruments, and with other complex engineered systems that will be connected through different modes, such as dedicated wired networks, beyond-fifth-generation (5G) networks, and quantum networks. Effective use of AI will also be critically important for exaflop and larger systems, where power optimization and management at the application level can translate to a significant difference in operational costs. Concurrently, the scale and complexity of these HPC systems introduces nonlinearities and system availability challenges that will require AI methods for control and optimization.

    b. **Particle accelerators.** Particle accelerators are complex multisystem machines that include many variables with nonlinear dynamics. DOE has invested hundreds of millions of dollars in the design, construction, supporting infrastructure, and operations of multiple accelerator facilities, which are integral to many aspects of the DOE scientific mission, from exploring fundamental physics to material studies. In recent years, the use of machine learning for particle accelerators has grown to include, but is not limited to, diagnostics, anomaly detection, forecasting, and AI-based controls. Integrating these methods into a comprehensive DOE effort for advance AI-based controls—including synergies with similar challenges faced by HPC facilities—will be necessary to enable better use of the facilities, including more efficient operations and improved/accelerated science discovery.

3. **Magnetic confinement fusion (tokamak) reactors.** Tokamak reactors use extremely powerful magnetic fields to confine plasmas that are a potential means of sustaining and controlling fusion in power plants. The $25B International Thermonuclear Experimental Reactor (ITER) [5] burning plasma experiment is the clearest example. These complex devices involve physics at many scales. Challenges include the development of efficient surrogate models for use in design, control, and prognostics; the avoidance or mitigation of plasma disruptions (to avoid damage to the device); and control of the power generation process [6, 7]. Deploying AI to control the plasma is the most promising strategy to increase the chance of sustained energy generation within two decades.

These examples highlight the different temporal and spatial scales inherent in complex systems and the complexity of individual components that make up most complex systems. Improving the design and control of such systems will require the ability to model the individual components and understand the interactions that occur within and among components across these scales. Time scales, for example, cover microseconds to decades and spatial scales include anything from single components to regions or nations. Although traditional modeling and simulation can provide insights, models at the necessary resolution are effectively computationally intractable, especially when considering the need for real-time control. Traditional models, constrained by computational capacity, also fail to adequately capture temporal and spatial interactions between subsystems and between the system and the environment. Moreover, even with these compromises, the models lack sufficient speed and accuracy to allow users to understand these systems, predict system behavior, or build control systems. For these reasons, traditional approaches to modeling each part of a connected system separately are no longer sufficient when dealing with changing contexts (for instance, changing climate and population dynamics) and their interactions. Wholistic, first-principles modeling is not feasible even with the exascale computing resources recently delivered by DOE. Indeed, the nature of a complex engineered system is such that modeling its components alone is insufficient (even if it were computationally feasible) because system behavior is driven not only by components themselves but also by nonlinear interactions and associated dynamics among the components.

AI building blocks discussed in earlier chapters, such as surrogate and foundation models and capabilities such as inverse design, provide the keys to the successful design, prognostics, and control of complex engineered systems. Initial AI tools are already pervasive in many such systems, including those used in science and engineering, but despite their impact to date, the full potential of AI in this context remains unrealized in the face of challenges such as those described next. Specifically, the effective application of AI to complex engineered systems is reliant on advances in both data acquisition and computing.

1. **Data.** Sensors are becoming ubiquitous and can provide accurate, real-time information about complex systems and an ever-expanding volume of historical data on system behavior, from which new AI models can be created and trained. In the next decade, advances in intelligent and autonomous sensors will amplify the need and impact of AI systems for data management and analysis, creating orders of magnitude more (and more complex) data. The introduction of AI in sensors creates a computational continuum from edge to HPC systems, which will both catalyze new AI capabilities for optimization and control while also generating new, larger, and more complex datasets. The models developed and trained with these new data sources will also be instrumental in addressing inherent uncertainties and errors in sensor data.
2. **Computing.** Even as true exascale has been reached in centralized HPC facilities, other advances, including the computing continuum just discussed, have leveraged powerful, energy-efficient, and cost-efficient computing capabilities that can be embedded in facilities and edge devices. The continuum bounded by these two very different computing modes—centralized and edge computing—also introduces opportunities for AI models (built and trained using HPC facilities) that implement lifelong AI-enabled learning (executed across the continuum) to provide AI-supported real-time control and prognostics in complex systems.

The vision of AI-enabled design, prediction, and control of complex engineered systems is captured effectively by the concept of the digital twin (DT). This term has been used in many contexts but can be generically defined as *a digital representation of an engineered system having the ability to respond to a current or predicted state of a physical system, where this representation is specifically used to facilitate better control, prognostics, and maintenance of the complex system over its lifecycle* (Figure 4-2). These characteristics, in particular the ability to control complex systems, mean that digital twin must operate several orders of magnitude faster than the physical system. Thus, the challenges of developing and integrating the necessary AI capabilities for complex engineered systems can be described in the context of the creation, optimization, and use of digital twins.

The DT concept is one that many companies have embraced over the past decade, particularly in the context of manufacturing and assembly. However, these investments are generally focused narrowly on individual systems (e.g., an assembly line) and for specific products. In contrast, the development of DT systems encompassing experimental facilities or national-scale infrastructure, or those required for national security, have not been sufficient developed with regard to the use of AI and machine learning [8].

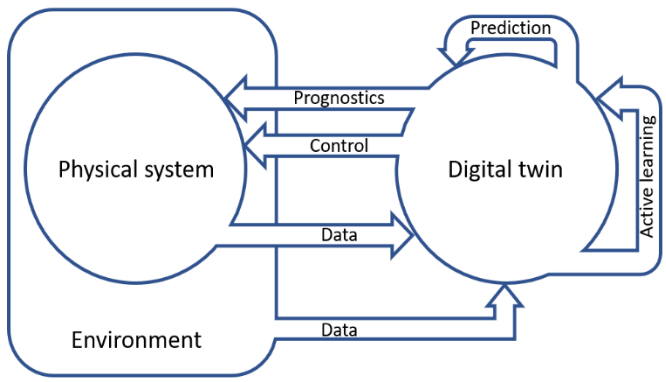

Figure 4-2. Conceptual flow diagram for a digital twin.

4.2 Grand Challenges

The following six research challenges, presented in the context of digital twins, must be addressed in order to leverage the promise of AI capabilities in the design, prediction, and control of complex engineered systems [7, 8, 9, 10, 11]. These six challenges are not independent of each other, nor are they independent of the other approaches described in this section, such as the development of AI-based surrogates (Chapter 01).

1. **Assurance.** The complex engineered systems most visible within the DOE mission space involve critical infrastructures. Here, assurance of the correctness of AI models used for the design and control of these systems is a central grand challenge. At an intuitive level, assurance addresses the question of whether an AI is making the right decision for the right reason, ensuring trustworthiness. Assurance encompasses many more factors, however, from uncertainty quantification to causal inference [12, 13]. Five facets of assurance are essential:

a. **Uncertainty quantification.** Uncertainty is an intrinsic part of both the data and the model built and trained by the data. Rigorous bounds must be computed to guarantee a robust and reliable DT and, by extension, a robust and reliable complex engineered system. The capability for uncertainty quantification will help users know when to trust the predictions from the model and greatly enhance the correct use of the AI system.

b. **Validation.** Validation of a DT considers the appropriateness of the model and can only be considered in the context of the intended application. Validation must be a continuous process that follows the evolving state of the physical system. Any validation process must consider the appropriateness of training and inference data, the specific model form and hyperparameter choices for the AI model, and the training process.

c. **Robustness.** Robustness can be characterized as how the machine-learning model responds to small changes in the data. Robustness for the DT and for AI models more generally depends as much on the selection of the data and measures of closeness as on the model design and training. For the DT, provision for robustness must be expanded to consider the full workflow, adversarial attacks, and unexpected occurrences in the environment or the data. The control system must respond robustly to noisy data, distribution shift in the data, and other normal fluctuations, while identifying anomalies that need to be investigated further and addressed.

d. **Explainability and causal analysis.** Explainability is the ability of an AI system or DT to explicitly associate a decision with a specific meaningful correlation identified in the data. Developing explainable AI models is a challenge that must be addressed before an DT can be fully deployed in the operation of many, if not most, engineered systems, particularly those associated with critical infrastructure. Causal analysis goes beyond correlation to identify the causal relationships that underlie the identified correlations. Causal analysis gives the system the ability to respond appropriately to disruptions; establishing these relationships will require the ability to test hypotheses by running experiments on the physical system. Explainability should not be an afterthought but rather should be built as a part of the AI systems in a DT. The DT should also provide mechanisms to incorporate physics knowledge and domain constraints.

e. **Anomaly detection.** Anomaly detection is the ability to identify or predict system behaviors or environments that were not considered “normal” or “usual” in designing or training the DT. Examples include system failures, system state shift, and adversarial attacks against the physical system (e.g., a cybersecurity intrusion, as distinct from adversarial attacks against the training and operation of the DT). Detection of anomalies is a critical part of any assurance effort but is identified as a separate area here because of the challenge of detecting and identifying system states that do not appear in training data.

2. **Model construction and the machine learning process.** The core of the DT is a model (or a set of models) that is built and updated on the basis of data from the physical system and the environment in which it operates. These systems span multiple spatial and temporal scales, data from many measurement modes, and large numbers of parameters, many of which cannot be measured directly. Furthermore, the process of creating the model itself entails significant challenges in data reduction, a process that focuses on identifying and representing the information contained in data. AI models are often updated continually and thus require continual or lifelong learning capability, in contrast to many existing modalities, in which sufficient training data is available prior to training the model. This challenge also includes federated learning, mechanisms to protect privacy or intellectual property or to reduce data transfer bandwidth requirements by distributing training to include edge devices. Challenges include identifying the model form, the training data and training process, and the appropriate prior information needed to construct the DT.

3. **AI-based control systems.** Control is a mature area of research, and most engineered systems have effective control systems. Nevertheless, AI presents new possibilities, many of which are increasingly important given the complexity of today’s complex systems and the demands placed on them by changing technology, environmental conditions, and usage patterns. Exploiting these possibilities will require advances in AI approaches, such as reinforcement learning and neuromorphic systems. There are many technical challenges, including the use of data-efficient learning, incorporating physical constraints, learning in partially observable large-scale complex systems, data and learning methods for distributed control, decision making under uncertainty, transfer learning, and power and speed requirements for real-time control, as well as the challenges in assurance presented previously and discussed by Sutton and Barto [14].

4. **Co-designed software and hardware ecosystem and workflow.** This challenge has three aspects. First, the physical system must be engineered to interact with the DT. This aspect of integration includes sensor design and placement, power management, and incorporation of edge-based, resource-efficient, and energy-efficient computational capabilities and control systems that can interface with the DT. Integrating the physical system with its digital representation also includes designing the

capability to deal with robustness and resilience issues, including data issues and adversarial attacks that could be introduced through the DT. Second, the DT itself introduces challenges for the hardware and software ecosystem, including issues of communication and bandwidth and the challenge of DT hierarchies (DTs for components and overall systems) and federation (i.e., when many instances of a physical system—each with an individualized DT—are deployed). Third, the ecosystem must include hardware that is embedded in the physical systems, which often imposes severe power, latency, bandwidth, and speed constraints [3, 15].

5. **Data quality, availability, and governance.** One challenge is the quality of data, which can be approached from a technical perspective. Specifically, improved sensors, combined with algorithmic advances in the placement of these sensors and processing of their data, will do much to address this challenge. The availability and accessibility of data is also a significant challenge, particularly because many of these systems are distributed (both geographically and across diverse businesses). In many, if not most, engineered systems, the availability of data can be severely restricted by the regulatory environment, privacy concerns, and intellectual property concerns. This challenge can be partially addressed through the development of an improved workflow, supported by technical solutions to issues such as privacy and equity. Digital twins and simulators that can take advantage of the DOE supercomputers can be used to generate synthetic data. However, regulations and requirements will also play a major role in making data available. Data constraints also pose a challenge to the efficiency of AI learning. Provenance of data used to train AI systems, particularly DTs associated with operational systems, will also be important, for instance to mitigate adversarial attacks through "poison" data schemes [16].
6. **Standardization and metrics.** The development and deployment of control systems with DTs is currently very system- and application-specific. As DTs become common in engineered systems, standard protocols for their design, production, deployment, certification, and maintenance will become necessary. Safety issues that arise when using DTs will drive a regulatory environment that will require standardization and guarantees. In addition to being necessary for safety and regulatory purposes, standardization and metrics will improve interoperability and performance and enable an overall increase in efficiency in the design and deployment of DTs, while also enabling the rapid growth of an industrial base to support this emerging technical area. Moreover, as complex systems are formed through vertical and horizontal composition of subsystems, DTs must be constructed in a way that allows DT-equipped components and subsystems can be composed in the same hierarchy as the engineered physical systems. This requires standardization and advances in composability.

4.3 Advances in the Next Decade

To realize the vision of robust and reliable control of the nation's critical infrastructure and other complex engineered systems, such as DOE facilities and instruments, significant investment and progress will be required in each of the crosscutting technology areas (see Section 02).

Mathematics and algorithms. The most immediate need for implementing a DT approach is to have robust and reliable surrogate models that can be used to construct and train control systems and on which DTs can be based. Surrogate models are described at length in Chapter 01, and for the prediction and control of complex engineered systems, knowledge-informed models and encodings for partially observed systems are the highest immediate priority, with significant progress expected in the next five years. Application- and use-specific surrogate models must be developed while addressing severe power, accuracy, and speed requirements in real-time control applications.

In addition to physics-informed surrogates, robust, data-efficient, and distributed reinforcement learning (e.g., for control algorithms) methods must be developed. Reinforcement learning is notorious for being a computationally difficult problem for complex systems, requiring significant volumes of data and robust test environments. Here, challenges include large state spaces, effective data representation and transformation approaches, widely varying time scales between control signals and system response, reinforcement learning that incorporates constraints, continual learning with the ability to update control policies, and robustness to limited or missing data. From a foundational perspective, the ability to generate theoretically sound convergence guarantees of optimality in online setting and verification is among the critical challenges. Significant progress must be made in each of these areas within the next 10 years to enable effective use of AI in the robust control of complex engineered systems.

To move DTs from demonstration to operation, there must be an explicit focus on assurance, with significant progress specifically in the areas of validation, causal inference, and security and privacy. Simply put, there must be an evidence-based methodology for assuring that AI can be used securely and robustly in control systems for critical infrastructure.

Data infrastructure. Data are the key to machine learning, and several challenges have been described. Within a decade, there should be common, secure data infrastructures for systems critical to our nation's infrastructure, such as the electricity grid and for DOE science and security infrastructure, including complex engineered instruments and facilities. This agreed structure would include standard data formats and expectations regarding provenance, availability, security, and privacy of data. Technical progress is also needed in providing a distributed infrastructure and tools for curation and maintenance of data throughout its lifecycle. In addition to providing a data infrastructure, DOE must work with industry to make critical data available for training the models necessary to enable DTs for the nation's critical infrastructure, from the components level to the local and regional grids operated by private entities.

AI software and workflow. Two technology crosscuts are combined into this topic. The focus for the next decade must be research as described in this chapter and pilot projects with preliminary demonstrations of validated control systems. Progress will be needed in software and workflows, with an explicit focus on validation, security, privacy, and the management of data and models. New software design paradigms that are secure, robust, resilient, composable, and analyzable by design are needed.

AI hardware. Successful deployment of AI for the control of complex engineered systems will require embedded, AI-specific hardware deployed in control systems "at the edge," as well as significant computational resources in the form of either dedicated, AI-specific HPC or HPC with AI accelerators. Existing and future AI hardware accelerators, such as graphics processing units, tensor processing units, and field-programmable gate arrays, are diverse and provide different characteristics with respect to training time, inference time, latency, power, and energy demands. A promising avenue in improving AI hardware is in customizing DT models to the given hardware platform. During the next decade, it will be necessary to leverage recent investment in semiconductor technology and to work with industry to develop and evaluate both embedded controllers and HPC capabilities. New sensors and actuators, and the computing and communications hardware necessary for edge AI computation, must be hardened and packaged to function reliably in extreme environmental and operating conditions common to most complex engineered systems, from particle accelerators to the power grid. As with the models and data, there is a need over the next 10 years to develop and evaluate methods to provide assurance on AI hardware as well.

4.4 Accelerating Development

Significant investments will be needed to achieve the decadal advancements described above, and these must be matched with investments in pilot projects and demonstrations. Five areas in which significant pilot demonstrations could be used to accelerate progress are described below. These pilots will drive specific technology advances both in the general areas described above and in specific application needs.

Potential research activities to accelerate development might focus on achieving the following outcomes:

1–3 years:

- Common data infrastructure solutions
- Foundations of assurance—knowledge-informed models at scale; metrics and validation

3–5 years:

- Demonstration of AI-based control at a pilot scale, including single-facility and local (e.g., single "neighborhood") demonstrations.

5–10 years:

- Scale up of algorithms/models to incorporate distributed control and the use of HPC resources.
- Scale-up demonstrations to manage critical infrastructure at the regional or national level.

In addition to these target outcomes, there is an immediate need to develop the basic infrastructure required to develop and demonstrate new capabilities. This infrastructure will range from the assembly and curation of datasets to the design and deployment of various scales of test beds, including instrumentation of existing test infrastructure to support the integration of new sensors, edge computation, and actuation associated with DT systems.

These capabilities could be demonstrated by the following potential pilots.

1. **The U.S. power grid.** This is a high-priority target application for accelerated research, development, and demonstration of AI for robust forecast and control. A pilot should involve a significant geographic area that includes both business and residential users, features significant penetration of generating capacity (including, for example, residential solar generation), and is susceptible to weather-related outages. The pilot will also need to be defined in terms of integration phases with respect to scope (moving from small district to regions) and functionality (e.g., initially limited to providing advisory information and gradually incorporating operational decisions and ultimately control actions).
2. **Control, including resource management, of a DOE leadership computing facility and/or scientific instrument.** The objective of this pilot would be to increase overall facility throughput and availability for a scientific workload. There are a limited number of possibilities, ranging from control of an individual DOE leadership-class computing center to a pilot involving the integration (over ESnet) of a distributed system connecting a DOE instrument (e.g., a light source) with a leadership-class machine.
3. **Transportation and mobility systems.** Several national laboratories have established projects in transportation systems. For example, Oak Ridge National Laboratory has already deployed optimized signal controllers in

Chattanooga, Tennessee, and Argonne National Laboratory has developed and demonstrated technology in Chicago, Illinois, to gather detailed traffic flow and vehicle mix data to train transportation models. A pilot would expand beyond these capabilities to provide system-wide optimization by coupling autonomous connected vehicles with traffic signal control in a major metropolitan area and providing data to supply-chain operations to allow improved fleet planning.

4. **Control of a tokamak fusion reactor.** A key goal associated with the 21st-century grand challenge for magnetic fusion energy concerns the control of the international ITER burning plasma experiment scheduled to begin in 2028. The need is to deliver advance warning for avoidance/mitigation of dangerous disruptions before critical damage occurs. A forward-looking pilot would build on the ability to use AI for passive prediction of disruption and to extend this capability to active control using a DT for a tokamak [17].
5. **Early detection of rare events in turbine engines.** As part of DOE's deep decarbonization goals for the transportation and land-based stationary power generation sectors, a major focus is on demonstrating high-efficiency and safe gas turbine operation on 100% renewable fuels (hydrogen and sustainable aviation fuels). A pilot project, in this regard, would develop and demonstrate a robust AI/ML framework, coupled with either high-fidelity simulations or real-time experiments, that is capable of causal representation learning and prognostics of rare events.

4.5 Expected Outcomes

Engineered systems are becoming more complex and interconnected. While this evolution presents tremendous possibilities for improved efficiency and effectiveness leading to economic competitiveness, it also makes such systems more susceptible to disruption, whether through adversarial attacks (e.g., cyber-attacks), environmental events, or component failure.

The use of AI for control of complex engineered systems will advance U.S. economic competitiveness in a number of critical areas, such as manufacturing, computing, renewable and green energy generation, and energy storage. These advances will also support U.S. energy security in the face of emerging threats. Critical to these advances are a research agenda developing the fundamental advances in AI and demonstrating these advances in multiple pilot-scale projects on an accelerated timeline.

4.6 References

[1] Thurner, S., Klimek, P., and Hanel, R., 2018. *Introduction to the Theory of Complex Systems*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780198821939.001.0001

[2] Chai, T., Qin, J., and Wang, H., 2014. Optimal operational control for complex industrial processes. *Annual Reviews in Control*, 38, pp. 81–92.

[3] Gao, R.X., et al., 2020. Big data analytics for smart factories of the future. *CIRP Annals*, 69(2), pp. 668–692. https://doi.org/10.1016/j.cirp.2020.05.002

[4] Ran, Y., et al., 2019, "A survey of predictive maintenance: Systems, purposes and approaches. arXiv:1912.07383v1. Submitted December 12. https://doi.org/10.48550/arXiv.1912.07383

[5] Aymar, R., Barabaschi, P., and Shimomura. Y., 2002. The ITER design. *Plasma physics and controlled fusion* 44 (5), p. 519. (See also https://www.iter.org.)

[6] Dong, G., et al., 2021. Deep learning based surrogate model for first-principles global simulations of fusion plasmas. *Nuclear Fusion*, 61, p. 126061–126071.

[7] Jones, D., et al., 2020. Characterizing the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29(A), pp. 36–52. https://doi.org/10.1016/j.cirpj.2020.02.002

[8] Niederer, S., et al., 2021. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science*, 1, pp. 313–320. https://doi.org/10.1038/s43588-021-00072-5

[9] He, Y., Guo, J., and Zheng, X., 2018. From surveillance to digital twin: Challenges and recent advances of signal processing for industrial Internet of Things. *IEEE Signal Processing Magazine*, 35(5), pp. 120–129. http://doi.org/10.1109/MSP.2018.2842228

[10] Moyne, J., et al., 2020. A requirements driven digital twin framework: Specification and opportunities. *IEEE Access*, 8, pp. 107781–107801. https://doi.org/10.1109/ACCESS.2020.3000437

[11] Womble, D., and Hembree, C., eds., 2022. *The Second Artificial Intelligence for Robust Engineering and Science Workshop Report*. Technical Report ORNL/LTR-2022/399. Oak Ridge, TN: Oak Ridge National Laboratory, March.

[12] Silva, S.H., and Najafirad, P., 2020. Opportunities and challenges in deep learning adversarial robustness: A survey. arXiv:2007.00753. https://doi.org/10.48550/arXiv.2007.00753

[13] Stracuzzi, D.J., et al, 2017. *Uncertainty Quantification for Machine Learning*. Sandia Report SAND2017-6776. Albuquerque, NM, and Livermore, CA: Sandia National Laboratories.

[14] Sutton, R.S., and Barto, A.G., 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

[15] Capra, M., et al., 2019. Edge computing: A survey on the hardware requirements in the Internet of Things world. *Future Internet*, 11(4), pp. 100–124. https://doi.org/10.3390/fi11040100

[16] Baracaldo, N., et al., 2017. November. Mitigating poisoning attacks on machine learning models: A data provenance based approach. In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 103–110.

[17] Kates-Harbeck, J., Svyatkovskiy, A., and Tang, W., 2019. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568, pp. 526–531.

05. AI AND ROBOTICS FOR AUTONOMOUS DISCOVERY

Evidence-based science demands a coupling between observation, analysis, experiment, and synthesis. This alignment represents a potential inflection point for the modern scientific process where automation and robotics, enabled with artificial intelligence (AI) and machine learning (ML) models, can accelerate experimental science in a similar fashion to the way that modern AI/ML have accelerated data analysis. Data has been acknowledged as the fourth paradigm [1], and the combined use of AI/ML and automation is positioned to become a fifth paradigm, enabling us to infer complex patterns from experimental and simulation datasets and to derive new knowledge that can be tested using subsequent experimental design(s). Advances in AI/ML approaches, including deep generative models, surrogate models, active learning, and reinforcement learning, are already enabling new discoveries across a variety of fields, including materials sciences, chemistry, physics, and biology. AI/ML methods applied to the design and optimization of high-throughput laboratory experiments, enabled by these advances in AI/ML techniques, offer new means to probe matter and understand complex phenomena in unprecedented detail. However, meaningful progress is impeded by the lack of connection between computing and high-throughput laboratory instruments and experiments. We posit that automation and robotics can accelerate the progress and increase the throughput of large-scale scientific experimentation, while driving novel means to investigate complex, emergent phenomena in scientific domains relevant to the U.S. Department of Energy (DOE).

For this report, the definition of autonomous discovery is borrowed from King, et al. (2009) [2], where it is envisaged as an independent robotic scientist that “automatically originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments by using laboratory robotics, interprets the results, and then repeats the cycle.” King et al. demonstrated this vision, not surprisingly in biology, by designing ADAM (Figure 5-1), a robot that could automatically generate functional genomic hypotheses about yeast. The vision of harnessing AI/ML capabilities (including those described earlier in this section of the report) to create autonomous robotic scientists (or even robotic facilities) has been seen as being of high benefit and high risk, with the latter impacting investment levels and limiting our ability to prototype, much less realize, such specialized autonomous discovery facilities.

New AI/ML approaches during the past several years, such as outlined in the previous two chapters, significantly reduce the perceived risk of autonomous laboratories, offering opportunity to reinvent each step in this discovery loop (data analysis→hypothesis→experiment/observe), as well as to speed the iteration of that loop, fundamentally accelerating the discovery process. We discuss these opportunities as well as the role that automation/robotics will play in tightening the integration of theoretical, experimental, and computational processes.

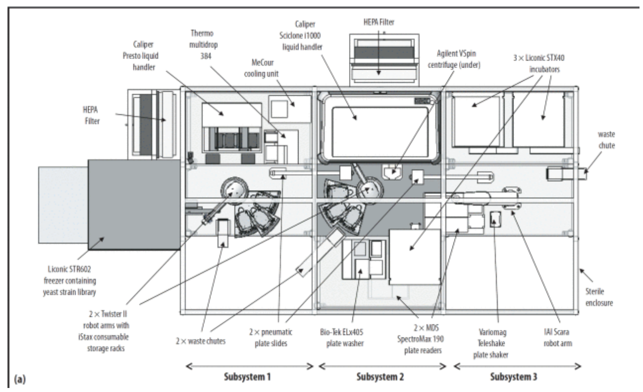

Figure 5-1. A schematic of the robot scientist ADAM that was used to generate novel hypotheses in examining yeast functional genomics [2].

In the observation stage, scientists use instruments (including microscopes, telescopes, sensor networks, etc.) to capture data representing diverse phenomena. Novel experimental instrumentation and the increasing resolution of these

> **PROJECT SPOTLIGHT**
>
> **Project Name:** Autonomous platform (Polybot) for electronic polymers discovery
>
> **PI:** Jie Xu
>
> **Organizations Involved:** Argonne National Laboratory
>
> **Goal**: Accelerate the development of a new class of polymer-based electronic materials that are flexible, durable, degradable/recyclable and easy-to-manufacture for our future electronics
>
> **Significant Accomplishment:** Developed a new experimental platform (Polybot, https://www.anl.gov/cnm/polybot) that combines the strengths of rapid and robust experiment acquisition from robotic technologies with fast analysis of complex datasets using ML, which enables autonomously electronic polymer engineering toward targeted structures for desired solid-state properties.
>
> **In the News:** Significant feature in *Newsweek* magazine: “America’s Greatest Disruptors: Budding Disruptors,” available at: https://www.newsweek.com/2021/12/24/americas-greatest-disruptors-budding-disruptors-1659089.html, accessed December 5, 2022.

instruments rapidly expand observation space and the resulting hypothesis space. Researchers can prove/disprove these hypotheses through the application of accepted research methods and experimental design. But these advanced instruments are expensive and in high demand, limiting access and consequently the number of hypotheses that can be tested and similarly confounding scientists' ability to reproduce, assess, and expand on research driven by these new advancements. The ability to automate scientific processes in the lab will also entail significant reduction in costs while enabling higher reproducibility and productivity for individual scientists as well as large team-science projects.

Increasingly, advances in AI and computing are necessary to enable exploration and access that are simply intractable today due to demands for large-scale laboratory instruments and/or computational demands. In biology, for example, new phenotypes within bacterial strains, such as to increase the production of threonine, are important subcomponents of bioproduction processes (including biofuels production). Yet engineering these strains involves a massive design space (Figure 5-2). Designs must optimize across at least *10 different genes, including several transcription factors, enzyme complexes, and other factors. Assuming even a single gene (which is translated to a protein product) with about 100 amino-acid positions, exhaustive mutagenesis and evaluation can exceed 20^{100} calculations. In addition, with various other components interacting, this problem can easily exceed 10^{160} considerations in the design space.* Indeed, the number of potential factors to explore expands with every new discovery of novel pathways, gene interactions, and even epigenetic factors influencing phenotypes. Concurrently, as our knowledge expands around alternate effects, the number of potential factors to explore also expands (e.g., the bacteria's ability to thrive in specific environmental conditions). Similar design space scales affect protein design, where human-guided design can at best explore extremely narrow subsets of the design space. AI/ML techniques offer the means to navigate the combinatorial complexity of these vast hypothesis spaces by, for instance by identifying novel patterns or potential designs based on inference models trained on the experimental and computational results from similar experiments. However, carrying out even 10^5 experiments exceeds the capacity of today's fully human-in-the-loop laboratory processes, even with modern instruments. Simply put, automation/robotics offers the only viable and practical means to address the combinatorial complexity of experimental design, and ultimately to accelerate scientific discovery.

Experimental design for observation and experimentation is also paced by human observations, decisions, and actions in the laboratory. Domains including biology, materials design, earth-systems modeling, ecological systems, and high-energy physics have these limitations in common. As with the biological systems design space discussed above, the combinatorics of computational or laboratory experiment design parameters are significant in even relatively simple physical, natural, or engineered systems.

Remote access and robot-assisted automation constitute the first step toward AI-driven laboratories (discussed below), as illustrated by many laboratory prototyping activities combating the COVID-19 pandemic. Scientists, forced to work remotely due to pandemic restrictions, used robotic instrumentation and automated instrument controls to perform critical research such as high-throughput screening of small

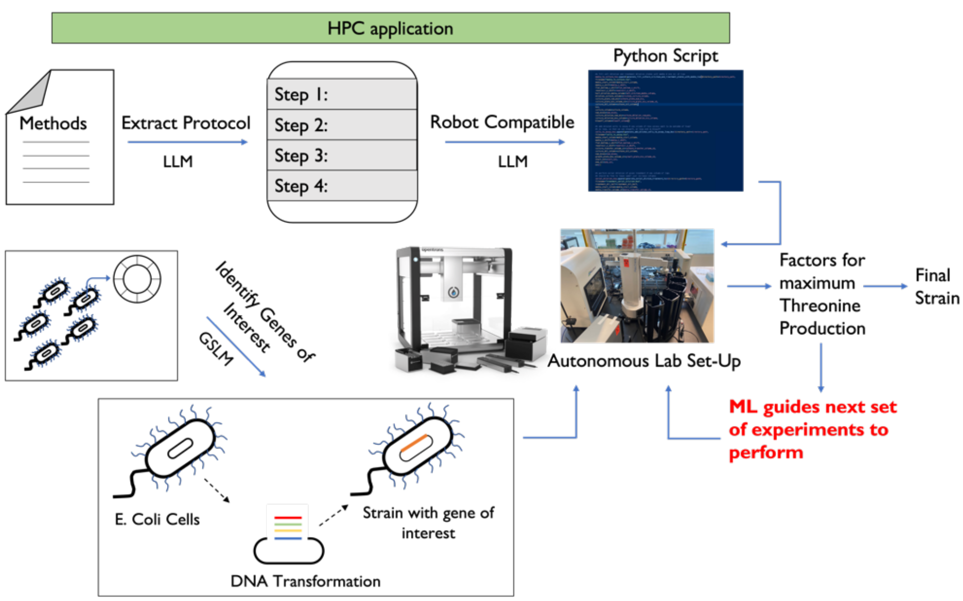

Figure 5-2. A schematic of how protocols can be automatically extracted from a methods section of scientific literature. This example is contextualized for engineering bacterial systems that can produce excess threonine (as the end phenotype). Various experimental steps such as DNA transformation, amplification, etc., and conditions are evaluated and automatically "coded" such that we can implement them on our robotic systems.

molecules/drugs and X-ray crystallography/cryo-electron microscopy or genetic sequencing of SARS-CoV-2 samples. In fact, the urgency of the pandemic combined with the use of robotics illustrated the democratization benefits with respect to accelerating science.

Automation, robotics, and AI/ML also offer a potential solution to a growing scientific crisis: reproducibility. More than 70% of researchers surveyed in 2016 failed to reproduce the results of another scientist, with 52% recognizing this as a significant crisis [3]. One estimate suggests that nearly one-third of published data may need to be re-evaluated because of poor reproducibility of these experiments; similar estimates are corroborative in other disciplines as well. Challenges in scientific reproducibility can stem from ambiguity in method and protocol, lack of specification of inputs and outputs, and faulty data analysis. Data generated via many experimental techniques are still difficult to reproduce, and human-implemented protocols are often tedious and error-prone, requiring specialized training (and technicians). Studies can require specialized equipment that can become obsolete or inaccessible, and protocols may have ambiguities or gaps, such as undocumented calibration parameters. Here again, the urgency of the COVID-19 pandemic led to increased sharing of protocols and, due to urgency, improvements in the precision and completeness of protocols. Addressing these challenges narrowed the gap that currently obstructs the development of AI-driven robotics that execute and eventually improve laboratory protocols.

Data management progress in recent years such as enabling access to AI/ML-related data and methods using the FAIR (findable, accessible, interoperable, reusable) principles has been promising but remains nascent. This is particularly challenging for laboratory data and will be even more daunting with the explosion in experimental data resulting from AI/ML and automated/robotic experiments.

The use of abstractions and programming languages has enabled scientific computation with codes that readily execute on different hardware platforms, workflow frameworks to combine resources across multiple platforms, and tools to migrate through multiple generations of hardware architectures. These abstractions, languages, and frameworks used with large-scale systems built today (including Cloud, high-performance computing [HPC] systems, and custom hardware such as accelerators) are a result of investing with a focus on the productivity, usability, verifiability, and validation of the written computer code (using compilers, model checking software, etc.). These tools grew organically across multiple hardware and software vendors and still maintained a level of interoperability and compatibility that allowed bespoke solutions to remain viable over many years of development and across multiple computing architectures. These abstractions, languages, compilers, and other tools are *lacking across the laboratory instrument (scientific) domains, locking the scientific community into bespoke, labor-intensive, and largely non-reproducible experiments. Absent a comprehensive approach, much of discovery science will continue to be dependent purely on human intuition and technical abilities—limiting experimental throughput and reproducibility, while continuing to be plagued with errors and quality control issues within the scientific process*.

These challenges span the thousands of individual laboratory experiments and their instruments but are equally consequential to the productivity and capacity of experiments using large-scale instruments at user facilities. Throughout the DOE complex, upgrades to various large-scale experimental and computing facilities are driving a notable increase in the volume of data collected and analyzed, straining the already limited capacity of fully “human in the loop” experiments. For example, the upgrade to the Advanced Photon Source at Argonne National Laboratory promises to improve the brightness of the X-ray beamlines by 500 times, implying that measurements that once took several days to weeks will produce at least as much data within only a few minutes to hours, dramatically accelerating the rates at which data accumulates. Combined with new capabilities in scalable workflow management, which directly enables access and analyses of data in situ through edge-enabled computing devices, the time-to-solution for analyzing generated datasets is being compressed such that the bottleneck is transferred to the “human-in-the-loop” decision-making. This outcome can potentially impede scientific progress if the data are not analyzed in a timely manner.

Automation and robotics within the scientific enterprise will democratize the scientific process, wherein participation from a variety of under-represented scientists and citizens can be evolved more organically and driven via engagement across disciplines. As noted earlier, increases in remote laboratory work necessitated by the COVID-19 pandemic—no longer limiting participation to those physically in the laboratory—opened the entire discovery process to more collaborators, increasing the diversity and inclusivity of many COVID-19 research projects [4–7].

5.1 State of the Art

Robotics has a long history in manufacturing, providing many examples for adaptation to scientific laboratory experiments. Until recently, AI capabilities limited the extent to which automation could be implemented. In the 1960s, one of the first deployments of robotics was in the General Motors production lines. The *Unimate* (from universal automation) robot automated the movement of high-temperature metal parts onto cooling water baths [8]. A decade later, the AI Center at the Stanford Research Institute (SRI) built *Shakey*, one of the first autonomous robots that was able to break down commands into a series of simple actions needed to achieve a particular goal with logic [9]. Beyond this, there have been several developments mostly focused on anthropomorphic systems that can interact and work with different instructions. The introduction of such automated systems drastically changed the face of modern manufacturing of cars (and other consumer products).

Today, in industry and defense we also see extremely advanced remote control and semi-autonomous robots, for example, from Boston Dynamics, which rely on AI capabilities for basic operational movement capabilities such as balance, reflex, adaptive locomotion, and fine movements. But for the most part, these robots, like their autonomous vehicle cousin AI systems, do not yet use AI models for higher-level, more complex activities such as problem-solving or adaptation beyond navigation. Similarly, the autonomy/robotics industry solutions are proprietary, closed systems, limiting the integration of multiple components to those from a single company. The potential, however, to combine the advances in AI model capabilities—including generalization and emergent properties—with such advanced robotic systems creates an unprecedented opportunity to transform scientific experimentation and discovery.

Within the scientific community, some of the first attempts at building a fully robotic scientist involved analyses of yeast genomes to characterize 13 orphan genes and their functions through a robot scientist called ADAM [2, 10], as well as to propose new small molecules (or drug candidates) for malaria using another automated system called EVE [11]. Since then, a number of studies have demonstrated robotic automation in laboratories to design new materials for energy storage [12, 13], for additive manufacturing to explore the toughness of a parametric family of structures [14], inorganic materials [15], two-dimensional (2D) crystal superlattices [16], novel biosystems [17], biocatalysts, de novo drugs [18], synthesis planning of small molecules [19], and many others. While an extensive list of articles covering autonomous discovery/self-driving labs is beyond the scope of this report, we refer the interested reader to [20].

Further, recent industry progress with cloud laboratories (remotely programmable and usable) such as Emerald Cloud, Strateos, and automated bioprocess/synthetic biology laboratories such as Ginkgo Bioworks and Zymergen, have demonstrated that automation/robotics can clearly accelerate industrial processes, from high-throughput screening (for biomedical applications and biomaterials design) to other allied areas. For DOE-specific applications, these industry-demonstrated approaches also bring opportunity to accelerate progress in other fields such as bio-catalysis, advanced manufacturing, climate sciences, high-energy physics, and beyond—if there is synergistic growth across integrated facilities (as we discuss in section 3.2, Grand Challenges). Recent investments in public-private partnerships focused on laboratory process automation in chemistry, materials, advanced manufacturing, and synthetic biology, along with DOE investments in Bioenergy Research Centers and the Agile BioFoundry, also imply that these AI/robotics approaches can catalyze broad-reaching benefits in improving productivity and reproducibility, managing and optimizing experimental resources, and ultimately driving and accelerating scientific innovation.

Despite their wide applicability and promise, AI-enabled automated labs face a consistent set of challenges that are common across multiple domains (biology, chemistry, physics, material science, etc.). Today, most high-throughput experiments are operated by a highly educated workforce (including PhD-level scientists) across DOE facilities—a need that has mostly emerged because of the robotic industry’s bespoke, complex solutions. For example, while several companies have developed proprietary robotics/control systems, there is a lack of open standards or community-based development, including scalable application programming interfaces (APIs) for ensuring easy integration across such robotics/control systems. In terms of the computational ecosystem discussed above, the automated laboratory mirrors computing in the 1960s and 1970s, where each computer had unique, proprietary programming languages, operating systems, and architectures whose fundamental storage and operating units might be 8, 16, or 24-bits. As in the early days of computing, this proprietary diversity in automated laboratory systems has led to a proliferation of ad hoc solutions even for the most common laboratory activities and procedures, requiring proprietary integration tools and software and adding to the cost and complexity of maintaining such systems. For scientific experiments, both the repetitive/common and bespoke solutions must co-exist and work seamlessly for automated execution of experimental steps to be achieved.

Advances in AI/ML techniques discussed in earlier chapters of this report are poised to revolutionize this landscape, and with tangible impacts that will also prove motivating for industry to move toward more open systems, accelerating the development of automated laboratories. For example, developments in large-language models (or foundation models, discussed in Chapter 02) are now enabling robotic systems to automatically understand and infer the “steps” involved within a particular task (e.g., inferring choice of “healthy” snacks after a workout or creating procedures for complex tasks), and similar extensions are enabling robots to often mimic human behaviors by just watching. Advances such as with surrogate models (Chapter 01) will enable the near-real-time operational decisions necessary for robotic laboratory work, while the inverse design capabilities discussed in Chapter 03 will further extend the capabilities of foundation models. Given the strides made in visual systems and language models, we posit that these technologies are ripe for advancing automated laboratories as well.

5.2 Grand Challenges

We highlight three grand challenges that exemplify the application of significant advances in AI and robotics to support autonomous discovery:

- Building a robot scientist to accelerate scientific discovery.
- Building a high-throughput automated facility for scientific discovery.
- Developing smart integration for connected scientific facilities.

5.2.1 BUILDING A ROBOT SCIENTIST TO ACCELERATE SCIENTIFIC DISCOVERY

A central goal for automated discovery is to improve the efficiency and speed of repetitive actions within scientific processes (e.g., DNA transformation). Each experiment is typically conceived based on prior knowledge; refined based on observations; and optimized through iterative executions, evaluations, and adjustments, to eventually be distilled into a discrete set of steps that are then executed in the laboratory by some combination of humans and instruments (a “protocol”). Near-term advances in AI and robotics will enable these protocols to be executed more efficiently and quickly, but the development of a robotic scientist that *can design and optimize the protocol itself* is a grand challenge. This experiment design process may involve accruing and/or assembling a set of instruments (or designing new ones), as well as other intermediate steps that are documented for further downstream use. Thus, any scientific experiment may be viewed as an iterative workflow involving multiple steps, captured as a protocol (analogous to a computer program, but executed by humans and instruments rather than a computer). While human intuition for designing such experiments is synthesized from existing knowledge and experience, automation (via robotic scientists) requires learning this experience and synthesizing knowledge from structured and unstructured data sources—which is quite different from how humans learn scientific experimental processes. Hence, the automation of scientific process design with robotic scientists requires advances in knowledge distillation and synthesis that go beyond current approaches (that mostly include structured data within datastores, or information represented via ontologies). With advances in foundation models, there is an opportunity to capture implicit representations of knowledge both in a domain-specific and domain-agnostic manner.

A robot scientist must be able to synthesize vast amounts of scientific knowledge and data, and then develop ways to incorporate prior(s) and generate new hypotheses based on current experimental observables. The aspect of constructing new hypotheses is dependent on an *inner* AI loop, which is dominated by fast analyses of existing data (e.g., exploiting surrogate models) and relating this analysis to scientific evidence drawn from distilled knowledge in prior literature and current observations. The *outer* AI loop will then provide a way to evaluate these hypotheses and select the “most promising” ones for experimentation. This outer AI loop will leverage active and reinforcement learning approaches.

Further, robot scientists will require extensive automated planning for designing experiments. This need will demand rigorous statistical techniques such as optimal experimental design or via robotic planning approaches that have been targeted for self-driving vehicles or automated design capabilities such as discussed in Chapter 03. Although limited prototypes foreshadowing robot scientists exist for specific domains such as in manufacturing, generalizing such robot scientists for domain-agnostic scientific experiment design is a grand challenge.

5.2.2 BUILDING A HIGH-THROUGHPUT AUTOMATED FACILITY FOR SCIENTIFIC DISCOVERY

A fully automated experimental facility will be essential to enabling human and/or robot scientists to connect and collaborate on multiple experiments simultaneously, or to facilitate the adaptation and translation of experiments and protocols from one domain to another. This concept of an autonomous scientific facility will also require capabilities for robot scientists to be modeled as an assembly of connected work cells (Figure 5-3a) consisting of closely related instruments; or for a robot scientist to be assembled on demand to implement an experimental protocol (Figure 5-3b). Eventually, such connected robotic platforms could be used to execute larger experiments (or even ensembles of experiments) based on a common operating environment. One of the key challenges for implementing such high-throughput automated facilities for scientific discovery is in **enabling integration of DOE infrastructure**, which comprises some of the nation’s large scientific instruments, with ad-hoc robotic instruments for specialized disciplines.

Advances in the use of AI for control and optimization of complex engineered systems (Chapter 04) will be critical for the necessary modeling of such large-scale facilities, including the use of advanced simulation toolkits that provide not only system schematics and visualizations but also simulate instrument self-assembly (to execute a scientific experiment) and overall operation for specific classes of experiments. This will require significant investment in the development of “digital twins” and associated virtual environments (e.g., with augmented/ virtual reality) to support the full range of design and operation, including interactive instrument design, scaling, and the modeling and prototyping of experiments at scale.

Similarly, different experimental techniques (e.g., neutron scattering, X-ray tomography, cryo-electron microscopy/electron tomography [EM/ET]) are often combined with computational simulations to characterize material properties at multiple temporal and spatial scales.

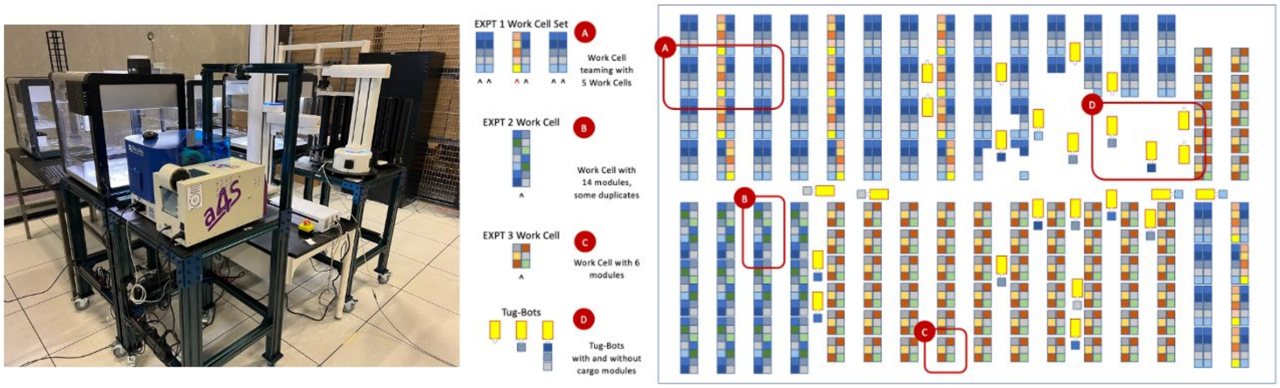

Figure 5-3. A conceptual overview of an autonomous scientific facility composed of flexible workcells. (a) A workcell is composed of a set of connected instruments/robots that can be operated as a single robotic scientist. (b) This assembly is facilitated by "on-demand" units that are brought together by tugboats and can be used to investigate a variety of scientific phenomena.

AI/ML techniques (specifically, surrogate models and active steering of experiments) can act as a "glue," providing a significant leap in how such complex/emergent phenomena are measured. In addition, by creating a digital twin of each experiment, which is used to constrain systematics to a much finer degree than currently possible and where the twin is better characterized by AI/ML than any person could replicate, we have the potential to self-calibrate simulations and digital twins for dynamically changing systematics.

A second challenge in this area is the development of ***self-calibrating facilities***, as illustrated in the processes being developed for future sky-mapping telescopes [21]. We posit that combining the execution and in situ analysis of multiple experiments and simulations via AI systems can improve upon instrument/experiment measurements by at least an order of magnitude, *without redesigning the instruments*. For example, in climate monitoring, edge AI and automation have enabled the controlling of scientific instruments (such as weather LIDAR) to detect features of interest in low-resolution scanning mode, then automatically adjust to high-resolution data collection with fixed aim at the detected feature.

5.2.3 DEVELOPING SMART INTEGRATION FOR CONNECTED SCIENTIFIC FACILITIES

DOE runs some of the nation's large scientific instruments, including particle accelerators, synchrotron light sources, high-energy laser systems, HPC environments, and other instruments that are integral to the scientific enterprise not only within DOE but also across universities and industry. However, these facilities run primarily as independent systems. This is in sharp contrast to our routine ability to combine multiple, distributed computing, data, and sensing resources into coherent experimental configurations through the use of workflows and high-performance networks including the Energy Sciences Network (ESnet).

DOE facilities and instruments are currently run at a scale that supports largely automated sample processing and loading conditions. Moreover, new instruments as well as upgrades to existing instruments introduce improved automation capabilities. These include partial (in situ) data processing and analyses that can provide operators with support for decision making (about the experiments to be conducted or even how they can troubleshoot some experiments). However, the automated sample processing utilizes robotics in a way that is largely driven by the motivation to increase the throughput and is less focused on creating smart interconnected experimentation. For instance, AI and interconnected experimentation can enable queued experiments to be combined with others to increase efficiency, reduce redundant work across facilities, and even include possible follow-up experiments in downtime. In these situations, AI/ML techniques can play an enabling role in not only improving the throughput but also in designing, optimizing, controlling, and executing experiments at unprecedented scale. With developments in active learning and reinforcement learning techniques and automation, these facilities could execute experiments autonomously with little human intervention needed for scheduling and operation.

The combination of autonomous laboratories with new robotic and sensing technologies and advances in HPC sets the groundwork to create ***connected scientific instruments of the future***: where federated experimental and computing facilities can "collaborate" on specific scientific tasks, providing significant acceleration (> at least an order-of-magnitude speedups) than currently enabled by automated laboratories. While present-day laboratory automation focuses on siloed throughput, there are synergies from considering many large-scale instruments collectively as a federated network alongside expanding DOE supercomputing infrastructure. In addition, edge computing technologies including sensor networks and novel hardware architectures can enable high-throughput data analysis, which can then be fed into HPC-enabled simulations. The complexity of such integrated instruments—with diverse computational, sensing, measurement, and other resources—will demand AI capabilities to guide the configuration, optimization, and operation of experiments.

5.3 Advances in the Next Decade

These three grand challenges motivate three suites of capabilities that must be developed within the DOE.

5.3.1 AUTOMATION-SPECIFIC CAPABILITIES

Open co-design of laboratory robotics. DOE investment in laboratory robotics is essential to reach the data volumes and quality needed to enable autonomous discovery and self-driving labs. While current approaches have produced working prototypes of self-driving labs, they do not exploit the full potential of robotics, as industry advances and academic research in robotics are often narrowly focused on replicating human actions and tasks. Unique micro- and nanofabrication expertise within DOE can be leveraged to operate at much smaller scales than conventional robotics, for example, for DNA assembly [22] or phenotypic screening [23], embedding molecular sensors on semiconductor chips [24], producing wireless and optically activated microscopic sensors [25] that enable a scientific internet of things (IoT), monitoring metabolism through quantum effects [26], or interacting with cells and their metabolism through light [27, 28]. Thus, DOE investment in laboratory robotics will provide unique abilities to study and manipulate matter at the appropriate data scale and cost, in domains where industry and academia cannot (or lacks incentive).

Apart from sensors and instrumentation development, DOE investments toward the open co-design of laboratory robotics must include open software ecosystems that can provide an interoperable environment for laboratory equipment. Related efforts in Europe have focused on adopting open standards such as SILA [29] for robotic instrumentation; however, such standardization and toolkit adaptation remains elusive given the diversity of vendors and custom (and often proprietary) software developed for operating such robots or a particular company's product line. This situation will require developing (1) open standards to enable interoperation, (2) protocols and frameworks to facilitate open exchange of information and metadata across experimental workflows, (3) self- and auto-calibration capabilities for robotic instruments, and (4) computer vision and modeling approaches for capturing how experiments can be run.

Robotics and automated laboratory in remote and/or harsh environments. Another important area of research is in the use of robotics and automated labs in settings that are dangerous or inaccessible (e.g., due to location or spatial scales) to humans or even extant electronics and robotics systems. There is a significant need for research that "lets" robotic labs handle the dynamic environments encountered in inhospitable environments, as well as for field research on scales or timescales that are impractical for human observers (e.g., longitudinal measurement campaigns). This research can potentially leverage work within the DOE National Nuclear Security Administration (NNSA) or other federal agencies such as the National Aeronautics and Space Administration (NASA); however, there are unique applications within the DOE scientific facilities where significant research investments are needed.

Additional items to address in this domain include interacting with unique (never-seen-before) settings or conditions, requiring prediction outside of trained data bounds; interacting intimately with unknown, multi-domain physics; creating physical interfaces between the robot and objects/environments; learning from mistakes in some scenarios and having zero tolerance for mistakes in others; intelligently placing sensors and understanding their performance in the environment; and reacting/adapting to changing conditions, limited access, and limited power. This underscores the need to infer both perception and action well beyond any available training set, with the need to perform complex and unpracticed *physical* tasks (e.g., manipulation).

Real-time autonomous agent learning for scientific facilities. Data generated by upgraded instruments and robotic instruments (including deployed field laboratories and sensor networks) pose an important challenge for managing autonomous scientific facilities. AI models will be required to support time- and/or resource-limited situations through the use of existing knowledgebases, while rapidly integrating new data in real time to feed forward into control and decision models will enable on-the-fly decision-making. Among other impacts, this capability would enable data collection by agents or robots whose distribution and data sampling could also be orchestrated by an AI system. Other potentially high-impact application areas beyond autonomous laboratories include experimental apparatuses, additive manufacturing, and robotics in the field and inhospitable environments. While some real-time data integration is being performed in narrow fields today, we envision a broad integration of disparate data that requires fast cleaning, processing, and blending for use in training AI models.

5.3.2 CROSSCUTTING TECHNOLOGY CAPABILITIES

Mathematics and fundamental research. The ambitions of autonomous discovery with AI and robotics all rest on (1) developing computational interfaces to manage scientific goals and methods; (2) correct and reliable robotic execution of the goals; and (3) navigation and optimization among the physical constraints of materials, supplies, and experimental conditions. If these three distinct challenges can be unified computationally with advances in AI, then the fundamental capability of autonomous discovery can be realized.

Abstractions and languages for encoding laboratory protocols for automated execution are also foundational to achieve the grand challenges discussed in Subsection 5.2 above. Specifically, scientific laboratory protocols in an automated setting will drive the generation of highly reproducible experiments and high-quality data. Scientific protocols are the step-by-step techniques used in research. A protocol in data science research, for example, can be shared via the code used to transform the input data to statistical conclusions drawn or charts generated. Similarly, a protocol in biology is typically shared through written step-by-step instructions with a sufficient level of detail such that a reader of the protocol could replicate the study. Furthermore, protocols have different levels of generality.

Protocols can be expressed solely in terms of reproducing the results of a single paper; however, protocols can also be used more generally for expanding on a particular result or applying it to different inputs. For example, a protocol to perform a type of gene "knockout" can simply outline the steps used to knock out a particular gene in a particular model given a particular study's downstream conditions; or a protocol can be abstracted to be applicable for different types of models, different genes, and even different conditions.

Protocols are today typically written for human colleagues who share a vast amount of background information, common sense, and practice. They involve a combination of human actions, instrument settings and procedures, and other functions in the context of a fully manual laboratory process. Advances are needed to translate this shared knowledge to computational infrastructure, along with standard APIs for digital control of instruments. The fluidity of abstraction as well as the challenge of translating traditional scientific protocols from open-natural language into closed-machine steps is a major challenge.

Model calibration is also essential for experimental design. Calibration refers to the ability of a model to correctly characterize its own uncertainty on a problem, and that its characterization nears exactness in practice. Calibration is achieved by linking experimental data to model inferences. In practice, calibration requires data infrastructures for collecting different conditions of data and for harmonizing across the data's organization, given that calibration is generally specific to a particular instrument and its own noise and uncertainty characteristics. Further advances are needed in areas such as automating the calibration of AI models within experimental settings against instruments and tasks, as well as understanding how to translate calibration settings across instruments and experimental conditions.

An allied area of AI research where there have been significant challenges is integrating neural network models with symbolic computation techniques (such as incorporating first-order logic). While emerging techniques are focused on developing neural-symbolic models, the best ways to incorporate them with emerging mathematical and formal logic constructs, such as probabilistic models, remain elusive. This situation is highlighted by the fact that language models developed for both general purpose and specialized (domain-specific) areas are poor in predicting out-of-distribution samples, as well as in logical/deductive/abductive reasoning. Significant advances in formal logic and reasoning will be necessary to enable autonomous discovery facilities to flourish.

Software and frameworks. Currently, scientists have no unified or accepted notion of a programmatic interface for performing experiments; collecting experimental results digitally; and using coding interfaces to control the instruments, the sample handling equipment, or other devices in the conduct of wet-lab research. Many unspoken assumptions are made in traditional laboratories—even with respect to important attributes such as ambient temperature, rinsing protocols, and handling of samples. These assumptions need either to be computationally inferred with the usage of AI planning models or explicitly outlined programmatically by the scientist. Furthermore, results from experiments need to be communicated back from the robotic platforms. While seemingly simple, it will be challenging to ensure that AI laboratory systems properly respond to cases such as experimental failure, noise, and even "serendipitous" anomalies—the root of much paradigmatic innovation in science. The process of transforming subjective experience into analytical and quantitative measures, emphasizing some aspects of experience over others, is also challenging. Finally, to fully leverage autonomous discovery especially in conjunction with AI techniques for optimal experimental planning, new techniques must be theorized for the convening of which experiments should be performed to achieve a specific goal.

Data management and AI workflows for autonomous scientific discovery. Current scientific data are generally balkanized, disorganized, and disaggregated, accessible mainly (if at all) through the supplemental information of an associated publication. In experimental laboratory settings, the harmonization problem across batches, instrument models, conditions, and even presentations of the data is an additional challenge. In experimental sciences such as biology, much of the data today is kept in complex Excel sheets, stipulating the range of assumptions and conditions used for generating the data. With the aim of leveraging AI to drive experimental planning and prediction and even to generate synergistic connections between experiments and prior data, it is essential that new data infrastructure and methods be designed to accommodate historically generated data, as well as the increases in data generation from automated and autonomous instruments and laboratories.

As facilities continue to expand and upgrade, the data volumes that will be generated, such as from various light sources, are projected to reach multiple exabytes per year. At these scales, data analyses processes and workflows need to be primarily autonomous: from identifying what raw datasets to store to the meaningful extraction of information from such datasets.

AI predictive models are built by learning from massive and often disparate sources of training data, and the crucial steps of assembling training datasets currently requires months of work for a human. The result is that opportunities for timely scientific discovery are missed. As the amount and complexity of available data across the science, energy, and security sectors continue to increase exponentially, the need for AI methods to augment, if not automate, the tasks related to curating and preparing data, managing heterogeneous data, and building training datasets will be indispensable for transformative change to occur in the efficiency and effectiveness of AI prediction in robotics, autonomous discovery, and beyond.

The enabling technologies that could transform future data management and infrastructure systems will rely on major advances that are enabled by new AI methods and capabilities, including data format standardization, optimal data sampling, and data transfer. Major data-intensive uses of these transformative AI technologies across science, energy, and security will also present researchers with additional challenges such as those related to data security and privacy. These and other challenges and opportunities are detailed in Chapter 14.

Beyond the challenges of building a data infrastructure to enable harmonization across user facilities, domains, and even computing and experimentation, such factors as explainability, trust, and rigorous system evaluation capabilities will pace the adoption of autonomous discovery within traditional scientific practice. For example, in developing a programmatic interface for user facility–based autonomous discovery, researchers will have to deal with the challenge of overcoming the experiential and education divide. This is an education and adoption challenge, and much progress is already underway with the increase in computational requirements across education in the scientific disciplines. Furthermore, such an interface needs to be co-designed with many disciplines, experiments, and future experiments in mind, likely requiring many workshops and outreach activities. These practical challenges are already being met with solutions, as many science education programs increase computational learning requirements.

AI-oriented hardware architecture. An additional area of research is needed in seamless integration of sensor networks with embedded AI/ML capabilities such that analyses of data can be performed in situ—where the data are generated. Edge analysis will be passed to subsequent steps of an experimental protocol such that downstream tasks can automatically “register” and “anticipate” failures as experiments are designed and executed. While advances in novel AI-oriented hardware continue to fuel the race toward exascale and zettascale computing, this race needs to be aligned with scientific use cases requiring compatible resources at the edge. This need extends the traditional co-design of individual computing platforms to also include their integration with (and the design of) scientific instruments.

5.4 Accelerating Development

We outline two pilot projects that can provide both near-term improvements while also demonstrating forward paths that will provide insight, including “early failures,” to additional pilot projects in different domains.

Regional, continental, and earth-scale monitoring systems initiative. With new sensor modalities, AI “at the edge” (within the sensors [30]) can analyze data in situ, detecting anomalous conditions and events, and ultimately provide (e.g., coupled with predictive models) real-time decision support for both natural and man-made events. Moreover, AI@Edge enables automation, for instance using an AI model to detect events or conditions of interest and reconfigure the instrument (e.g., sampling rate, focus, or direction of observation) to examine such phenomena in greater detail. Such intelligent sensing networks could then be used to monitor—and capture in detail—events such as earthquakes and extreme weather conditions. A pilot project is needed to integrate autonomous predictive and reactive modeling capabilities spanning the edge-to-HPC continuum in an Autonomous Discovery laboratory context. This effort will include training AI edge code for autonomy in detecting conditions or events of interest as well as pre-analyzing observational data. Additionally, these edge AI codes must include actuation capabilities, such as adjusting observational instrument settings (e.g., orientation, sampling rates, etc.). In turn, edge capabilities in this pilot must be integrated with HPC modeling systems to create a control and modeling loop that continually updates the HPC models (and ultimately will continually train AI models).

This pilot could leverage existing DOE facilities and resources, such as accelerating the adoption of edge AI capabilities in weather observation instruments operated by the Atmospheric Radiation Measurement (ARM) User Facility as well as the new Urban Integrated Field Laboratories. Similarly, edge AI systems are already being deployed for experiments supporting not only ARM and U-IFL sites but also NNSA’s NA-22 program (in situ radiation monitoring and characterization) and the Office of Energy Efficiency and Renewable Energy (EERE’s) Vehicle Technologies Office (vehicle mix and flow observations).

Pilot project on the design of (bio-)polymers for critical mineral extraction. Critical minerals are currently used in multiple clean energy technologies including electric motors and batteries. Although many of these elements are abundant in the Earth’s crust and are present in waste by-products like coal ash, acid mine drainage, or consumer electronics, they are often dilute or difficult to separate with existing technologies. For example, China controls 80% of the world’s supply of rare-earth elements, of which 920 lbs. are needed for each F-35 jet. Similarly, other critical elements, such as the lithium and cobalt needed in batteries, are primarily produced in Chile and the Congo, respectively.

Given the inhospitable regions where such critical elements are found, there is an immediate need to extract, concentrate, and recycle critical minerals in a more efficient manner, for instance, by using chemical sorbents that act as selective sponges, and which need to be designed from existing knowledge (and from scratch).

However, our knowledge in chemistry, materials, proteins, and organisms is siloed, and we need advances in both AI and robotics to enable the design of novel materials that can extract rare-earth minerals. Given the incompleteness of current knowledge, a pilot is needed to develop new, AI-enabled laboratory processes to inform our understanding of biological principles that can be used to capture and concentrate these minerals directly. For example, bacteria have already been engineered toward reducing certain types of phosphates complexed with certain minerals. Yet the ability to design, build, and test large-scale cycles of rare-earth extraction or processes and scaling them within reactors will require (1) automation and new biotechnology protocols to survey and design new synthetic organisms with the ability to process such materials, (2) new AI methods that go beyond interpolation to examine which pathways can be used in these applications, and (3) data collection at scale regarding rare-earth microbiomes including fungi and other organisms that can provide new ways to energize rare-earth extraction and clean energy technology.

The pilot would also need to develop AI approaches that can automatically identify datasets for developing general-purpose, multitask, and cross-discipline material property models for DOE-relevant domains, prioritizing data collection efforts for these materials/tasks where needed. The pilot initiative would entail the development of robotic standards using open-source standards such as the robotic operating system (ROS, [31]) and explore additional open standards to support the interoperability among instruments and scientific workflows and across facilities (for demonstrating a smart-interconnected facility).

In the ten-year timeframe, progress on two complementary areas would need to be achieved, including in (1) developing and promoting standards for modular hardware that support interoperability and discoverability with automatic data capture and storage; and (2) developing methods to automatically construct digital twins for laboratory equipment during operation. Intersection with other approaches, including property inference and inverse design, surrogates, foundation models, and prediction of complex engineered systems, will be needed.

5.5 Expected Outcomes

AI-enabled autonomous discovery presents a new, and urgently needed opportunity to increase the productivity and reliability of DOE's investments in scientific instruments and infrastructure. Through interconnected networking, automation, and integrated AI for experimental design, autonomous discovery will reduce bottlenecks due to human involvement and increase reliability through systematic handling of materials and smart tuning of instruments. The impact of increased throughput, analysis, and aggregation of experimental data has the potential to drive scientific discovery by accelerating the ability to (1) screen new materials or drugs experimentally; (2) increase the calibration and decrease the uncertainty of models through the leveraging of AI-driven exploration, typically occurring during instrument downtime; and (3) increase scientific productivity by offloading time spent on protocol design to computing methods.

There is potential for AI systems, as described here and in detail throughout Section 01 of this report, to revolutionize the nation's manufacturing, therapeutics, and sustainability industries through advances in biological and inorganic material design capabilities. For example, AI for drug design has been an accelerating and growing field, initially leveraging high-performance computing with traditional modeling and simulation. Surrogate (Chapter 01) and Foundation (Chapter 02) models offer the potential to create AI-driven computational systems that can screen billions of compounds a day for a target of interest. A similar system for inorganic material design would enable the design, testing, and manufacture of new materials, employing not only robotic laboratory and manufacturing systems but also inverse design approaches outlined in Chapter 03. However, without commensurate AI-enabled experimental throughput, these models will be limited due to insufficient training and calibration, in turn reducing our capacity to synthesize and test the resulting vast array of potential compounds. These challenges are also mirrored in other domains such as material design.

Autonomous discovery, integrating experimental science with AI-enabled computation, will also radically extend the reach of the enormous investments across the DOE complex in computing advances (e.g., the Exascale Computing Project [ECP]) and instruments ranging from genetic sequencers to entire user facilities. Illustrated in the life sciences domain, these advances offer the most promising path toward closing the massive gap between the identification of a disease or target of interest and an appropriate therapeutic—by accelerating the planning, design, and execution of experiments to identify targets and potential compounds of interest. The same outcomes will accrue not only to the life sciences but also to the material sciences and other domains (outlined in Section 02 of this report), lowering the overall cost of designing, engineering, and manufacturing novel materials.

5.6 References

[1] Tansley, S., and Tolle, K.M., 2009. *The fourth paradigm: Data-intensive scientific discovery*. Hey, A. J. G. (ed.). Vol. 1. Redmond, WA: Microsoft research.

[2] King, R.D., et al., 2009. The automation of science. *Science*, 324(5923), pp. 85–89. Available at https://www.science.org/doi/10.1126/science.1165620, accessed December 7, 2022.

[3] Baker, M., 2016. Reproducibility crisis. *Nature* 533(26): pp. 353–366.

[4] Al Saadi, A., Alfe, D., Babuji, Y., Bhati, A., Blaiszik, B., Brace, A., Brettin, T., Chard, K., Chard, R., Clyde, A., Coveney, P., Foster, I., Gibbs, T., Jha, S., Keipert, K., Kranzlmüller, D., Kurth, T., Lee, H., Li, Z., Ma, H., Mathias, G., Merzky, A., Partin, A., Ramanathan, A., Shah, A., Stern, A., Stevens, R., Tan, L., Titov, M., Trifan, A., Tsaris, A., Turilli, M., Van Dam, H., Wan, S., Wifling, D., and Yin, J., 2021. IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better Leads. In: *50th International Conference on Parallel Processing (ICPP 2021)*, Association for Computing Machinery, New York, NY, USA, Article 40, pp. 1–12. https://doi.org/10.1145/3472456.3473524.

[5] Clyde, A., Galanie, S., Kneller, D.W., Ma, H., Babuji, Y., Blaiszik, B., Brace, A., Brettin, T., Chard, K., Chard, R., Coates, L., Foster, I., Hauner, D., Kertesz, V., Kumar, N., Lee, H., Li, Z., Merzky, A., Schmidt, J. G., Tan, L., Titov, M., Trifan, A., Turilli, M., Van Dam, H., Chennubhotla, S.C., Jha, S., Kovalevsky, A., Ramanathan, A., Head, M.S., and Stevens, R., 2021. High-throughput virtual screening and validation of a SARS-CoV-2 main protease noncovalent inhibitor. *J. Chem. Inf. Model.*, 62(1), pp. 116–128. https://doi.org/10.1021/acs.jcim.1c00851

[6] Trifan, A., Gorgun, D., Salim, M., Li, Z., Brace, A., Zvyagin, M., Ma, H., Clyde, A., Clark, D., Hardy, D.J., Burnley, T., Huang, L., McCalpin, J., Emani, M., Yoo, H., Yin, J., Tsaris, A., Subbiah, V., Raza, T., Liu, J., Trebesch, N., Wells, G., Mysore, V., Gibbs, T., Phillips, J., Chennubhotla, S.C., Foster, I., Stevens, R., Anandkumar, A., Vishwanath, V., Stone, J.E., Tajkhorshid, E., Harris, S.A., and Ramanathan, A., 2022. Intelligent resolution: Integrating Cryo-EM with AI-driven multi-resolution simulations to observe the severe acute respiratory syndrome coronavirus-2 replication-transcription machinery in action. *The International Journal of High Performance Computing Applications,* 36(5-6). https://doi.org/10.1177/10943420221113513

[7] Bhati, A.P., Wan, S., Alfè, D., Clyde, A.R., Bode, M., Tan, L., Titov, M., Merzky, A., Turilli, M., Jha, S., Highfield, R.R., Rocchia, W., Scafuri, N., Succi, S., Kranzlmüller, D., Mathias, G., Wifling, D., Donon, Y., Di Meglio, A., Vallecorsa, S., Ma, H., Trifan, A., Ramanathan, A., Brettin, T., Partin, A., Xia, F., Duan, X., Stevens, R., and Coveney, P V., 2021. Pandemic drugs at pandemic speed: infrastructure for accelerating COVID-19 drug discovery with hybrid machine learning- and physics-based simulations on high-performance computers. *The Royal Society: Interface Focus* 11(6). https://doi.org/10.1098/rsfs.2021.0018

[8] Gasparetto, A., and Scalera, L., 2019. From the Unimate to the Delta robot: The early decades of industrial robotics. In: *Explorations in the History and Heritage of Machines and Mechanisms*, pp. 284–295, Springer, Cham.

[9] Kuipers, B., Feigenbaum, E.A., Hart, P.E., and Nilsson, N.J., 2017. Shakey: From conception to history. *Ai Magazine*, 38(1), pp. 88–103.

[10] King, R.D., et al., 2009. The robot scientist Adam. *Computer* 42(8). https://ieeexplore.ieee.org/document/5197424, accessed December 7, 2022.

[11] Williams, K., et al., 2015. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal Society: Interface*, 12(104). https://royalsocietypublishing.org/doi/10.1098/rsif.2014.1289, accessed December 7, 2022.

[12] Zhang, B., Merker, L., Sanin, A., and Stein, H.S., 2022. Robotic cell assembly to accelerate battery research. *Digital Discovery*, 1, pp. 755–762. https://doi.org/10.1039/D2DD00046F

[13] van der Westhuizen, C.J., du Toit, J., Neyt, N., Riley, D., and Panayides, J.-L., 2022. Use of open-source software platform to develop dashboards for control and automation of flow chemistry equipment. *Digital Discovery* 1, pp. 596–604. https://doi.org/10.1039/D2DD00036A

[14] Gongora, A.E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K.G., … and Brown, K.A., 2020. A Bayesian experimental autonomous researcher for mechanical design. *Science Advances*, *6*(15), eaaz1708.

[15] Szymanski, N.J., Zeng, Y., Huo, H., Bartel, C.J., Kim, H., and Ceder, G., 2021. Toward autonomous design and synthesis of novel inorganic materials. *Materials Horizons* 8(8), pp. 2169–2198.

[16] Masubuchi, S., Watanabe, E., Seo, Y., Okazaki, S., Sasagawa, T., Watanabe, K., ... and Machida, T., 2020. Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials. *npj 2D Materials and Applications*, 4(1), pp. 1–9.

[17] Harnden, K.A., Wang, Y., Vo, L., Zhao, H., Lu, Y., 2021. Engineering artificial metalloenzymes. In: *Protein*

Engineering: Tools and Applications, First Edition. Zhao, H., et al. (eds.), Wiley-VCH GmbH. https://doi.org/10.1002/9783527815128.ch8

[18] Grisoni, F., and Schneider, G., 2022. De novo molecular design with chemical language models. In: *Artificial Intelligence in Drug Design,* pp. 207–232, Humana, New York, NY.

[19] Thakkar, A., Johansson, S., Jorner, K., Buttar, D., Reymond, J.L., and Engkvist, O., 2021. Artificial intelligence and automation in computer aided synthesis planning. *Reaction chemistry & engineering* 6(1), pp. 27–51.

[20] Häse, F., Roch, L.M., Aspuru-Guzik, A., 2019. Next-generation experimentation with self-driving laboratories. *Trends in Chemistry* 1(3), pp. 282–291. https://doi.org/10.1016/j.trechm.2019.02.007

[21] Janssen, M., Falcke, H., Kadler, M., Ros, E., Wielgus, M., Akiyama, K., Baloković, M., Blackburn, L., Bouman, K.L., Chael, A., and Chan, C.K., 2021. Event horizon telescope observations of the jet launching and collimation in Centaurus A. *Nature Astronomy* 5(10), pp. 1017–1028.

[22] Gach, P.C., et al., 2016. A droplet microfluidic platform for automating genetic engineering. *ACS Synthetic Biology* 5(5), pp. 426–433. https://doi.org/10.1021/acssynbio.6b00011

[23] Iwai, K., et al., 2018. Automated flow-based/digital microfluidic platform integrated with onsite electroporation process for multiplex genetic engineering applications. *IEEE Micro Electro Mechanical Systems (MEMS)*, pp. 1229–1232. doi: 10.1109/MEMSYS.2018.8346785

[24] Fuller, C.W., et al., 2022. Molecular electronics sensors on a scalable semiconductor chip: A platform for single-molecule measurement of binding kinetics and enzyme activity. *PNAS,* 119(5). https://doi.org/10.1073/pnas.2112812119

[25] Cortese, A.J., et al., 2020. Microscopic sensors using optical wireless integrated circuits. *PNAS*, 117(17), pp. 9173–9179. https://doi.org/10.1073/pnas.1919677117

[26] Nie., L., et al., 2021. Quantum monitoring of cellular metabolic activities in single mitochondria. *Science Advances* 7(21). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8133708/, accessed December 7, 2022.

[27] Rienzo, M., et al., 2021. High-throughput optofluidic screening for improved microbial cell factories via real-time micron-scale productivity monitoring. *Lab on a Chip* 15. https://pubs.rsc.org/en/content/articlelanding/2021/LC/D1LC00389E, accessed December 7, 2022.

[28] Wegner, S.A., et al., 2022. The bright frontiers of microbial metabolic optogenetics. *Current Opinion in Chemical Biology* 17: 102207. https://www.sciencedirect.com/science/article/pii/S1367593122000928?via%3Dihub, accessed December 7, 2022.

[29] Wikipedia, “Standardization in Lab Automation,” last edited November 21, 2022. https://en.wikipedia.org/wiki/Standardization_in_Lab_Automation, accessed December 8, 2022.

[30] Beckman, P., Sankaran, R., Catlett, C., Ferrier, N., Jacob, R., and Papka, M., 2016. Waggle: An open sensor platform for edge computing. In: *2016 IEEE SENSORS*, pp. 1–3, Oct.

[31] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., and Ng, A.Y., 2009. ROS: An open-source Robot Operating System. In: *ICRA Workshop on Open Source Software*, 3(3.2), p. 5, May.

06. AI FOR PROGRAMMING AND SOFTWARE ENGINEERING

Software is ubiquitous throughout the scientific and energy infrastructure, whether it is controlling large-scale, complex instruments or monitoring and optimizing electricity generation and distribution. As these systems become increasingly complex, ensuring their robustness, reliability, security, and rapid recovery are not only more challenging but also more critical given the central role that these infrastructure assets play in every facet of society. Increasingly interconnected, today's software-rich infrastructure is also vulnerable to both cybersecurity and natural disruptions that can propagate through systems—demanding verifiably correct and robust code. Consequently, major productivity, quality, and verifiability improvements are needed in programming and software engineering for applications ranging from **complex systems controls** (see Chapter 04) to **scientific applications** (Section 02 of this report) to fully exploiting **new HPC architectures**. Recently, advances have been made demonstrating the feasibility of large-scale language models (e.g., GPT-3 [1], Codex [2], etc.) to write code and to translate code from one language to another. This chapter discusses the opportunity for using artificial intelligence (AI) to address the software needs of the control systems (including scientific instruments and energy infrastructure), scientific computing, and high-performance computing (HPC) communities. The importance of this topic is not only highlighted in the 2020 AI4Science report [3] but also in a follow-on report, "Program Synthesis for Scientific Computing" [4]. More recently, the DOD community also created a report on "The Science of Software Development and Use," [5], further laying the groundwork for pursing the opportunities discussed in this chapter.

A central strategy is to develop AI assistants for code development and maintenance, software optimization and tuning, and software architecture and design to improve the productivity of human programmers by 10x or more, to improve the reliability of software systems, and with a focus on the needs of the U.S. Department of Energy (DOE) in science and engineering codes and for leading-edge architectures. Beyond the critical importance of addressing these and other current challenges, a key long-term objective is to develop AI systems whose algorithms exceed the best human-known algorithms and that ultimately include novel algorithms unanticipated by experts.

6.1 State of the Art

Early-stage commercial tools, such as GitHub Copilot [6] and Amazon CodeWhisperer [7], act as AI-assisted co-programmers, generating code recommendations based on prior code and user comments. At the system software level, tools such as Ithemal [8] use AI techniques to make data-driven choices with system software (e.g., compilers and runtime systems). Concurrently, there is extensive research in this area, focusing on using AI to perform a broad range of critical tasks such as *text-to-code generation* (to generate code from natural language description) [1], *code completion* (to predict following tokens based on code context) [9], *code translation* (from one programming language to another) [10], *defect detection* (to identify resource leaks, code vulnerabilities) [11, 12], *clone detection* (to measure the semantic similarities between codes) [13], *cloze test* (to predict the obscured section of a code) [14], *code search* (e.g., for a natural language query, to find the most relevant code in a collection of codes) [15], *code repair* (to fix bugs automatically), *code summarization* (to generate natural language comments for code), and *documentation translation* (from one natural language to another).

The scale and complexity of DOE science and energy systems differentiates them from typical commercial systems and their associated applications (e.g., desktop computers, mobile devices, or cloud or web services) targeted by today's commercial code assistance tools, including CoPilot and CodeWhisperer. While there are some extremely large industry data centers with tremendous computing capability, the types of applications and workloads they support often differ greatly from the demands of large-scale science. The

PROJECT SPOTLIGHT

Project Name: FourCastNet

PI: Anima Anandkumar

Organizations Involved: Lawrence Berkeley National Laboratory, NVIDIA, Caltech

Goal: Scale deep learning models to forecast global atmospheric dynamics at high resolution to accelerate expensive numerical models in weather and climate.

Significant Accomplishment: Developed the first deep learning model capable of forecasting global weather patterns with accuracy and resolution comparable to operational numerical weather models, which features orders-of-magnitude reduction in computational cost per forecast.

In the News: Perlmutter-Powered Deep-Learning Model Speeds Extreme Weather Predictions. Available at https://www.nersc.gov/news-publications/nersc-news/science-news/2021/perlmutter-powered-deep-learning-model-speeds-extreme-weather-predictions/, accessed December 2, 2022.

unique needs, complexity of architectures, sparsity of code examples, and rapid evolution in systems all make the use of AI for programming and software engineering particularly challenging and limit the extent to which commercial systems can play a significant role. For example, the state-of-the-art approach to simplify code development for the diversity of HPC architectures is to invest heavily in performance-portability programming abstractions such as Kokkos [16] and Raja [17]. While there have been recent examples of successful use and deployment of AI for HPC [18], the use of AI in support of HPC programming for scientific computing does not yet exist.

Throughout this chapter, the terms "software" or "code" refer not only to specific source code for individual executable programs but also to more complex software *systems*, such as workflows, or to systems comprising various AI model components, along with their associated configurations (hyperparameters, weights, etc.).

6.2 Grand Challenges

A number of AI building blocks such as foundation models (Chapter 02) and inverse design (Chapter 03) will be critically important to achieving the cost, agility, and quality improvements necessary for current and future software systems. Here we outline three grand challenges that will themselves form the underpinnings for applying AI to the programming and software engineering needs across DOE's scientific, energy, and security mission areas. The first addresses the software necessary for the control and automation of complex, interconnected systems—discussed in Chapters 04 and 05. The second grand challenge focuses on codes embedded throughout the DOE enterprise that support science and engineering across desktops, clusters, laboratory instruments, and other experimental, computational, and data infrastructure. The third grand challenge in this area specifically targets DOE's leadership HPC infrastructure.

6.2.1 AI FOR PROGRAMMING CONTROL SYSTEMS

Control software for complex engineered systems, such as those discussed in Chapter 04 for the electricity grid, high-performance computing facilities, nuclear power generation systems, and others, is critically important to support national security, economic competitiveness, and the quality of life in the U.S. These systems face ever-increasing threats from evolving and growing demand patterns, the changing climate (driving both changing demand and weather disruption), aging infrastructure, reliance on international supplies, and adversarial attacks. Similarly, control systems are at the heart of automated or AI-driven laboratories operated throughout the DOE complex as detailed in Chapter 05. A real-time control system will thus comprise many scales, many components, and many subsystems—*each controlled by software systems*—that must not only be internally correct and robust but must interoperate and adapt to both short-timescale disruptions and long-timescale evolution.

A grand challenge to use AI to generate control-system software could significantly improve our ability to ensure reliability and resilience for control systems that adapt to rapidly changing conditions. In the envisioned system, ***AI*** capabilities comprise ***the control software***. Utilizing systems such as inverse design (Chapter 03), the AI system for control will be generated automatically based on domain-specific design objectives, operational and simulation-generated data, and assurance mechanisms quantifying the trustworthiness and correctness of the control system with respect to its design objectives. The system must also be robust, reliable, and resilient to faults from natural and adversarial causes. For example, a mission-critical infrastructure operating at 99.999% ("five nines") of reliability will experience 5 minutes of outage annually. If this level, or higher, is required for the overall system, then the control software can be no less, and would ideally be much more, reliable.

6.2.2 AI FOR SOFTWARE ENGINEERING OF SCIENCE AND ENGINEERING CODES

Science and engineering codes designed for DOE are distinguished from the broader software community by algorithmic complexity and rigorous validation and verification requirements. In addition, scientific codes are very specialized and are not likely to exist in large repositories, confounding approaches such as those used by Co-Pilot and CodeWhisperer, which learn from vast landscapes of common methods and classes of applications. Finally, scientific codes are typically a composition of codes and libraries that require multi-physics, complex numerical methods, and a range of multi-fidelity and multi-scale solutions. The national importance of DOE's science, engineering, energy, and security missions requires the laboratories to make significant investments in scientific code development. The potential for using AI to aid in that development is a grand challenge that could lead to massive improvements in productivity, software quality, and application sustainability—all critical challenges for the DOE complex. We briefly discuss each of these improvements next.

AI-generated software could *significantly improve productivity*. The engineering and science codes that represent the bulwark of DOE's science, energy, and security mission areas each take years, sometimes decades, to develop to the quality level required by those missions. These codes are developed with a wide range of requirements for fidelity, uncertainty quantification (UQ), verification and validation (V&V), and time to solution. The traditional approach—independent development by thousands of teams across the complex—is time consuming and expensive. AI-generated scientific codes would enable the generation of custom codes that incorporate new algorithms (ultimately including those created by AI models), are constrained by user requirements, and implicitly certified for production use. For example, integrating new components such as a new preconditioner or eigensolver today is gated by the ability of many software teams incorporating the components into libraries or other codes. An effective AI-driven software maintenance system would accelerate the adoption of such improvements. Success would mean order-of-magnitude productivity improvements and significant savings in expenses historically used for software design, development, evaluation, and production hardening. From a DOE mission perspective, our simulation and modeling capabilities would exhibit the agility, quality, and responsiveness increasingly demanded by mission needs, along with equally important reductions in costs associated with human-in-the-loop factors such as long development cycles and identifying and addressing software flaws.

Software quality is a significant and growing challenge for scientific codes [19]. AI-generated software constrained by strict guidelines for software engineering [20] could lead to consistent quality of software and explainable, enabling human verification, and simplifying debugging. High-quality AI-generated and AI-verified software could also be more secure than our existing code base—ensuring code that is without known code vulnerabilities and is responsive to emerging cybersecurity threats.

Finally, AI-developed software systems and the associated quality improvements will also dramatically improve software *sustainability.* Today's codes, developed over many years by a succession of programmers, can involve hundreds of thousands of lines of code—a daunting challenge to maintain, much less to extend or port to new computing architectures or laboratory instruments. For the AI model, however, modifying, extending, and porting code are innate capabilities that make these tasks no more challenging, and potentially even easier, than code generation.

6.2.3 AI FOR PROGRAMMING HIGH-PERFORMANCE COMPUTERS AND ADVANCED ARCHITECTURES

The past decade has seen extreme growth in heterogeneous architectures for high-performance computing. A recent DOE report stated that heterogenous accelerators are used in more than 100 of the TOP500 systems and in the majority of the TOP10 [21]. Each of the DOE's leadership-class computing systems at the DOE's science laboratories and the National Nuclear Security Administration (NNSA) implement different overall system architectures that each support a different organization of heterogeneous central processing unit (CPU) and graphical processing unit (GPU) node architectures, as well as multiple levels of complexity in the memory and storage systems. Future trends point to even greater complexity, with potential accelerators for dataflow [22, 23], neuromorphic [24], and quantum [25, 26] computing that could soon make their way into our HPC platforms.

Designing codes that are portable and performant for the diversity of HPC systems and architectures that exist consumes a large number of staff and computing resources at these national laboratories and their industry and academic partner institutions. Given a well-defined scientific problem with user constraints, reducing the time and resource costs of these activities will require AI models that can generate the algorithms and software system design that would support a range of HPC systems, effectively exploiting their unique hardware features. These codes must not only support large-scale parallelism in the system but also node parallelism and diverse internal architectures, while adapting computing and communication algorithms based on the network and storage topology and capabilities of the underlying platforms. For systems with configurable hardware, the AI models must further understand how to adapt the hardware to meet the primary objectives of the code, which could be optimized for energy efficiency, scalability, or time-to-solution.

AI-generated software that makes effective use of our HPC systems will dramatically reduce the time it takes to transition those systems into a production state and also could reduce or even eliminate the need for years of development on early-access systems. Moreover, these AI models have the potential to address emerging challenges associated with the extreme scale, complexity, and energy demands of exascale systems and beyond, notably in energy efficiency, scalability, and performance. Such improvements would enable better utilization of our platforms and accessibility to a much broader community of HPC users.

6.3 Advances in the Next Decade

To achieve the AI-enabled end states identified above, DOE and the broader research community must solve many intermediate and foundational challenges. Some of these challenges are themselves grand in their ambition and potential impacts on scientific and engineering generally. We highlight two technical advances that are of highest priority.

6.3.1 ADVANCES NEEDED FOR AI-ASSISTED SOFTWARE DEVELOPMENT AND CODE GENERATION

The grand challenges in this chapter describe an AI-assisted software development environment that is fundamentally different from the process involving human effort that exists today. In this new environment, a DOE scientist or control-system engineer will act as an architect or orchestrator, providing high-level requirements and directives to an AI system tasked with creating the software. That AI system will generate performant, portable, scalable, and correct code for a variety of different architectures ranging from large instruments and HPC systems to edge devices including intelligent sensors and scientific instruments. It will also generate the test suite, documentation, and codes necessary for V&V, as well as UQ. Equipped with these software systems, the scientist will iterate with the AI system as necessary to refine and finalize requirements and verify the results.

All three grand challenges described above assume a “language” for expressing requirements, and this language does not yet exist. An important step toward creating such a language or set of languages is to define methods (e.g., natural language, programming models, symbolic algebra) to express requirements, constraints, and design objectives in a way that minimizes ambiguity for the AI system and that maximizes the system’s ability to generate verifiable and correct code. Equally important, the languages must be ***accessible to engineers, analysts, or scientists*** in order to enable precise articulation of the design criteria specific to that domain. For example, a control-systems engineer may want to articulate the design objectives and constraints of individual control subsystems, including the power, memory, and speed constraints in the design of an energy-distribution network. A physicist interested in using AI and HPC to model turbulent flow as part of a re-entry code may provide constraints for uncertainty, fidelity of the result, and deadlines for completion. While the end state of these languages is different from anything currently in existence, initial work can build on domain-specific programming models and some of the excellent early results from the broader AI community, including those discussed above in the context of the state of the art.

While the end goal expressed earlier in the description of grand challenges is for fully automated AI code generation, tremendous progress could be made even in the immediate term using AI for “recommender” systems that provide guidance to software engineers, computational scientists, and control-system engineers. A focus on AI for software quality, productivity, and system portability is a natural evolution toward fully automated, AI-generated programs.

To satisfy the level of rigor required for V&V and UQ, we will need a focused research effort toward AI tools that generate test suites from standards/specifications written in natural language (e.g., automatically generate tests from ingesting the message passing interface [MPI] or OpenMP standards). Both correctness and performance measures are needed. Realizing AI-generated test suites and the generation of ensemble workflows for V&V seem plausible in the next 5 to 15 years.

6.3.2 ADVANCES IN THE AI-ASSISTED HPC SOFTWARE STACK

To create AI systems capable of generating codes to effectively utilize DOE’s assets such as HPC systems and research facilities, significant research is also needed on the HPC software stacks themselves. For example, achieving the objective of AI-generated code that can fully exploit unique architecture features will require a reconfigurable HPC software stack operating on the target system. This reconfigurable stack would, in turn, both enable a more flexible hardware design and simultaneously relax constraints on interfaces among runtime systems, programming models, and system software components. The resulting real-time adaptation of the software stack would also reduce or eliminate many trade-offs that currently end up being “baked into” the low-level software or even hardware during the design phase. Rather, with AI-generated code, these settings would be exposed to the AI-enabled software stack for resolution at execution time, factoring in the actual workloads being run on the system at that moment. Achieving this “on-the-fly” adaptation will require significant advances in composability, reconfigurability, and observability of the numerous components comprising the HPC software stack—in effect, not only using AI models to generate user codes but also to generate lower-level components of the software stack. This capability would increase the achievable performance of the system, as well as its ability to accommodate a much broader set of workloads, resulting in an increased “democratization” of these systems in terms of supported programming models and runtime systems.

6.4 Accelerating Development

The goal of AI-generated codes for advanced architectures could be significantly accelerated through a co-design approach with HPC and AI-hardware vendors. We anticipate a rapidly evolving commercial market for AI tools and hardware. With strategic DOE investments in co-design, as demonstrated by the Exascale Computing Project (ECP), we expect the vendors to be responsive to ideas that enable DOE scientists and engineers to make effective use of their hardware. In turn, communicating the specific needs of our science and engineering missions will lead to hardware designs more appropriate for our missions.

Workforce development—detailed in Chapter 16—is also a critical issue, where AI-enabled programming and software engineering capabilities will have a pronounced impact. By loosening the current entanglement between computational science and computer programming skills, the envisioned AI systems will remove barriers to entry for a much broader audience. We need to recruit top-tier researchers as well as educate/train DOE scientists in the fast-moving world of AI and this new software-development paradigm. The challenges associated with developing the AI methods, particularly those around composition of complex scientific codes, will require not only a mix of computer scientists, mathematicians, and software engineers but also new ideas and novel approaches that often come from those with expertise outside of these disciplines. The somewhat unique challenges for DOE (e.g., multi-modal, multi-fidelity, multi-scale) will require us to work closely with our university partners to evolve curricula and develop talent pipelines that understand and embrace the unique requirements underpinning DOE's scientific, energy, and security mission areas.

Finally, the use of AI-generated codes introduces new challenges for intellectual property (IP)—in particular, for licensing/copyright issues (e.g., what is the license for code generated from a training set that also contains GPL-licensed code? Who owns copyright to AI-generated code?). While we expect some of these challenges to be resolved in the broader research community, these types of issues can often create roadblocks for innovation. DOE should have a plan for how to deal effectively with the IP, legal, and cybersecurity concerns associated with AI-generated code.

6.5 Expected Outcomes

Achieving the grand challenges, through advances outlined above, can be accomplished through the development of a series of increasingly sophisticated methods, components, and similar stepwise increases in the autonomy and level of controls afforded the AI systems. The target destination for this path is the creation and use of AI systems that generate codes from high-level requirements, including designing innovative algorithms for a vast array of scientific, energy, and security problems. The objective is to develop AI systems whose algorithms exceed the best human-known algorithms and that ultimately include novel algorithms unanticipated by experts. To achieve full impact, the AI systems will not only generate operational codes but will also provide accompanying products including test suites, documentation, and verification.

By enabling scientists and control-system engineers to focus on domain science through expressions of requirements, this ecosystem will significantly reduce the human, time, and financial costs associated with the development, maintenance, and performance tuning characterizing today's methods. Beyond individual codes, AI-created workflows that generate, deploy, and optimize code operating across the full spectrum of HPC, networks, and edge devices will significantly increase the effectiveness and efficiency of our systems and lead to innovative designs that interoperate at an unprecedented scale, ultimately increasing efficiency and accelerating scientific discovery.

6.6 References

[1] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I., et al, 2018. Improving language understanding by generative pre-training. *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, accessed November 8, 2022.

[2] Chen, M., Tworek, J., Jun, H., Yuan, Q., Ponde de Oliveira Pinto, H., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Petroski Such, F., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Hebgen Guss, W., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W., 2021. Evaluating large language models trained on code. https://arxiv.org/abs/2107.03374, accessed on November 8, 2022.

[3] Stevens, R., Taylor, V., Nichols, J., Maccabe, A.B., Yelick, K., and Brown, D., 2020. *AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science*. https://doi.org/10.2172/1604756, accessed January 10, 2023.

[4] Finkel, H., and Laguna, I., 2020. *Report of the Workshop on Program Synthesis for Scientific Computing*, August.: https://www.anl.gov/cels/program-synthesis-for-scientific-computing-report, accessed January 10, 2023.

[5] Bernholdt, D.E., Cary, J., Heroux, M., and McInnes, L.C., 2022. *The Science of Scientific-Software Development and Use*, United States. https://www.osti.gov/servlets/purl/1846008 and https://doi.org/10.2172/1846008, accessed January 10, 2023.

[6] Github, 2022. Github Copilot: Your AI pair programmer, October. https://github.com/features/copilot, accessed November 9, 2022.

[7] AWS, 2022. Amazon CodeWhisperer, October. https://aws.amazon.com/codewhisperer/, accessed November 9, 2022.

[8] Mendis, C., Renda, A., Amarasinghe, S., and Carbin, M., 2019. Ithemal: Accurate, portable and fast basic block throughput estimation using deep neural networks. In: *International Conference on Machine Learning*, pp. 4505–4515, PMLR.

[9] Bruch, M., Monperrus, M., and Mezini, M., 2009. Learning from examples to improve code completion systems. In: *Proceedings of the 7th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering*, ESEC/FSE '09, pp. 213–222,

New York, New York, Association for Computing Machinery.

[10] Roziere, B., Lachaux, M.A., Chanussot, L., and Lample, G., 2020. Unsupervised translation of programming languages. In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (editors), *Advances in Neural Information Processing Systems* 33, pp. 20601–20611, Curran Associates, Inc.

[11] Chernis, B., and Verma, R., 2018. Machine learning methods for software vulnerability detection. In: *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, IWSPA ’18, pp. 31–39, New York, New York, Association for Computing Machinery.

[12] Chakraborty, S., Krishna, R., Ding, Y., and Ray, B., 2022. Deep learning based vulnerability detection: Are we there yet? *IEEE Transactions on Software Engineering*, 48 (9), pp. 3280–3296.

[13] Ghofrani, J., Mohseni, M., and Bozorgmehr, A., 2017. A conceptual framework for clone detection using machine learning. In: *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, pp. 0810–0817.

[14] Allamanis, M., Brockschmidt, M., and Khademi, M., 2018. Learning to represent programs with graphs. In: *International Conference on Learning Representations*.

[15] Sachdev, S., Li, H., Luan, S., Kim, S., Sen, K., and Chandra, S., 2018. Retrieval on source code: A neural code search. In: *Proceedings of the 2nd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, MAPL 2018, pp. 31–41, New York, New York, Association for Computing Machinery.

[16] Edwards, H.C., Trott, C.R., and Sunderland, D., 2014. Kokkos: Enabling manycore performance portability through polymorphic memory access patterns. *Journal of Parallel and Distributed Computing*, *74* (12), pp. 3202–3216.

[17] Hornung, R.D., and Keasler, J.A., 2014. The RAJA portability layer: Overview and status. Lawrence Livermore National Laboratory, LLNL-TR-661403, September 24.

[18] Huerta, E.A., Khan, A., Davis, E., Bushell, C., Gropp, W.D., Katz, D.A., Kindratenko, V., Koric, S., Kramer, W.T.C., McGinty, B., McHenry, K., and Saxton, A., 2020. Convergence of artificial intelligence and high performance computing on NSF-supported cyberinfrastructure. *Journal of Big Data*, 7 (1), p. 88. https://doi.org/10.1186/s40537-020-00361-2.

[19] Koteska, B., Mishev, A., and Pejov, L., 2018. Quantitative measurement of scientific software quality: Definition of a novel quality model. *International Journal of Software Engineering and Knowledge Engineering*, 28 (03), pp. 407–425.

[20] Storer, T., 2017. Bridging the chasm: A survey of software engineering practice in scientific programming. *ACM Comput. Surv.* 50 (4), August.

[21] Vetter, J.S., Brightwell, R., Gokhale, M., McCormick, P., Ross, R., Shalf, J., Antypas, K., Donofrio, D., Humble, T., and Schuman, C., et al., 2018. Extreme heterogeneity 2018-productive computational science in the era of extreme heterogeneity: Report for DOE ASCR Workshop on Extreme Heterogeneity, technical report, U.S. DOE Office of Science (SC), Washington, DC, (United States). https://www.osti.gov/biblio/1473756, accessed November 8, 2022.

[22] Delgado-Frias, J., Ahmed, A., and Payne, R., 1991. A dataflow architecture for AI. In: *VLSI for Artificial Intelligence and Neural Networks*, pp. 23–32, Springer.

[23] Emani, M., Vishwanath, V., Adams, C., Papka, M.E., Stevens, R., Florescu, L., Jairath, S., Liu, W., Nama, T., and Sujeeth, A., 2021. Accelerating scientific applications with sambanova reconfigurable dataflow architecture. *Computing in Science & Engineering*, 23 (2), pp. 114–119.

[24] Schuman, C.D., Kulkarni, S.R., Parsa, M., Parker Mitchell, J., and Kay, B., et al., 2022. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2 (1), pp. 10–19.

[25] Britt, K.A., Mohiyaddin, F.A., and Humble, T.S., 2017. Quantum accelerators for high-performance computing systems. In: *2017 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1–7.

[26] Möller, M., and Vuik, C., 2017. On the impact of quantum computing technology on future developments in high-performance scientific computing. *Ethics and information technology*, 19 (4), pp. 253–269.

SECTION 02: SCIENTIFIC DOMAINS

Ultimately, the value of developing and implementing the new AI approaches outlined in Section 01 is to address the unique needs of DOE's specific application and program areas. Throughout the report, we emphasize co-design approaches to this end, integrating expertise from program and application areas, mathematics, AI/ML foundations, computer science, instruments and data sources, new software and hardware architectures, frameworks, and platforms. This approach will also *change the nature of computational workloads and significantly increase the scale of resources needed from DOE's exascale systems* as the emphasis shifts more toward model training rather than solely for executing models. Transforming these science, energy, and security endeavors will entail rethinking fundamental concepts and approaches including the traditional simulation, modeling, and data analysis approaches, and addressing new and rapidly evolving demands placed upon underlying physical and software infrastructure. For these programmatic domain areas—each comprising multiple programs and program offices—we highlight the open opportunities for harnessing new AI approaches and capabilities, the challenges that must be overcome to do so, and what investments are needed to seize those opportunities.

07. OFFICE OF SCIENCE (SC: ASCR, BER, BES, HEP, NP, FES, AND SCIENTIFIC USER FACILITIES)

The U.S. Department of Energy (DOE) Office of Science (SC) programs underpin the basic and applied research of DOE mission areas and domains across the complex. The Advanced Scientific Computing Research (ASCR) program supports the SC and broader DOE science, energy, and security mission areas through basic research in computer science, applied mathematics, distributed systems, and computational sciences. SC's Biological and Environmental Research (BER) program supports transformative science and scientific user facilities to achieve a predictive understanding of complex biological, earth, and environmental systems necessary to ensure U.S. leadership in energy, infrastructure, science, and security. SC's Basic Energy Sciences (BES) pursues scientific research to lay the foundations for new energy technologies through discovery and to uncover new physics and phenomena spanning a wide range of materials and chemical processes that will drive innovation in areas such as energy resources, production, conversion, transmission, storage, efficiency, waste mitigation, quantum science, and microelectronics. Its High Energy Physics (HEP) programs aim to discover the ultimate constituents of matter and uncover the nature of space and time. The underlying theory and associated experiments in the three HEP frontiers—cosmic, energy, and intensity—cover science at all scales, from the smallest to the very largest [1]. The SC Nuclear Physics (NP) program aims to discover, explore, and understand all forms of nuclear matter. Nuclear physicists create, detect, and describe the different forms and complexities of nuclear matter that can exist in the universe, thereby better understanding the building blocks of the smallest nanostructures to the largest stars. SC's Fusion Energy Science (FES) program focuses on the scientific and technological innovations necessary to enable a unique U.S. vision for economically attractive fusion energy, with the goal of a fusion pilot plant by the 2040s [2]. Magnetic confinement fusion (tokamak) reactors are a major focus area within this effort.

DOE SC also supports a portfolio of 28 scientific user facilities supporting an international community of tens of thousands of researchers from DOE laboratories, universities, and industry across all of the SC scientific programs. These facilities range from light sources and accelerators to field laboratories, from high-performance computing (HPC) centers to DOE's national Energy Sciences Network (ESnet). Experimental scientific user facilities enable exquisite characterization, synthesis, and simulations (theory) of a very wide variety of materials and devices, allowing new understandings of underpinning mechanisms and spawning new advances in biology, materials science, physics, and chemistry. Fundamentally, major open opportunities exist both in increasing the efficiency of the synthesis-characterization-understanding workflow (via autonomous control and design of experiments) at user facilities, as well as in developing new algorithms and methods to improve solving of inverse problems relating structure to functionality. A fundamental "grand challenge" for the DOE Scientific User Facilities lies in how to best utilize these theory, computation, synthesis, and characterization facilities to solve specific problems in the most efficient and comprehensive manner possible. Today, only individual researchers address this challenge by making such judgements based on combinations of experience, cost of each experiment, and perceived utility. AI-based methods to optimize this workflow could transform the process with respect to critical measures including time-to-solution and reductions in cost.

Finally, we note that the pervasive nature of ASCR research in the context of artificial intelligence (AI) is reflected throughout this report, notably in Section 01: AI Approaches and Section 03: Technological Crosscuts, and thus is covered only briefly in this chapter.

7.1 Open Opportunities

Each of the SC programs described above has active research applying and advancing AI/machine learning (ML) methods while developing strategies to harness the emerging capabilities outlined throughout Section 01 of this report. This work and planning reveal opportunities across the SC programs, relying heavily on ASCR.

PROJECT SPOTLIGHT

Project Name: Reinventing coherent imaging data inversion

PI: Mathew Cherukara

Organizations Involved: Argonne National Laboratory, Advanced Photon Source

Goal: Use AI@Edge to enable real-time ptychography.

Significant Accomplishment: An AI model (PtychoNN) allowed us to realize speeds that were 100x faster and required 25x less data than used in classical approaches.

In the News: Cherukara, M. J., Zhou, T., Nashed, Y., Enfedaque, P., Hexemer, A., Harder, R. J., and Holt, M. V., 2020. "AI-enabled high-resolution scanning coherent diffraction imaging," *Applied Physics*

Two opportunities illustrate the potential for AI approaches to make advances. The first is to enable understanding of new materials, and the second is to transform scientific user facilities.

- **The development of general-purpose, AI-powered simulation tools could boost our capability to simulate materials and processes with high fidelity and spanning multiple orders of magnitude in spatial and temporal scales**. These tools could greatly expand our fundamental understanding of the behavior and dynamics of materials and complex biological systems over larger timescales, something that is critical in a wide range of domains, from the development of better energy storage materials to the exploration of complex quantum materials, to the development of the next generation of microelectronic devices. AI can play three critical roles in the development of a next generation of simulation tools through: (1) the acceleration of computations using surrogate models; (2) the ability to generalize to new systems not encountered before; and (3) the development of intelligent systems capable of adopting the best set of conditions, parameters, and components for simulations, digital twins, and experiments. Two examples are the development of *digital twins* (discussed in detail in Chapter 04) that capture the full life cycle of a material, and the development of AI-powered *universal atomic potentials* for atomistic simulations that are 1,000 times faster than non-AI methods while retaining first-principle molecular dynamics precision.
- **Significant opportunity exists for AI to transform facilities such as the Facility for Rare Isotope Beams (FRIB), Jefferson Laboratory (JLab), the Deep Underground Neutrino Experiment (DUNE), and ton-scale detectors for neutrinoless double beta decay**. This ranges from operations—AI-based control of accelerators and detectors—through experimental design, to enabling more autonomous discovery. Accelerator science and engineering provide the foundation for these facilities and underlie discovery in other sciences, including medicine and technology. Development of digital twins of emergent accelerator technology and capabilities is occurring at increasingly high-fidelity levels. This advance provides an opportunity to pursue inverse design strategies (see Chapter 03) for enhancing today's facilities and optimizing the building blocks of the facilities of the future.

The opportunities we outline in the following section are organized in terms of the six AI approaches detailed in Section 01 of this report.

7.1.1 AI SURROGATE AND FOUNDATION MODELS FOR SCIENTIFIC COMPUTING

The development of hybrid models that use a combination of traditional numerical prediction approaches and data-driven architectures will enable new capabilities in nearly every domain, including observationally informed components directly coupled into modeling frameworks. Furthermore, hybrid models will be well suited to new leadership class computing with mixes of central processing units (CPUs), graphical processing units (GPUs), and new architectures.

One example is DOE's flagship Earth system model (E3SM), which is designed to answer questions regarding climate impacts on food, water, and energy security at global scales, E3SM typically requires significant computational resources and time that sponsors are often unwilling or unable to support. Replacing sub-models in E3SM with much faster AI surrogates could reduce computational requirements and help enable E3SM's adoption as the primary tool for answering questions of climate impacts.

A second example is nuclear physics. At all energy levels, nuclear theory has increased its use of high-performance computing [3]. AI-based surrogates provide opportunities both to accelerate the numerical routines underlying these advanced calculations as well as to provide training data for downstream uses.

AI-based models represent a significant new opportunity to harness the large volumes of complex data that are costly to create, process, and manage, e.g., using foundation models (See Chapter 02) to create open and connected knowledge graphs. Trained on HPC systems, these models could also result in inference capabilities capable of running on embedded processors to enhance data collection through AI "at the edge." BER, for instance, has invested significantly in measurement facilities (e.g., Atmospheric Radiation Measurement [ARM] facility and Environmental Molecular Sciences Laboratory [EMSL]) and field data collection (e.g., Next Generation Ecosystem Experiments). Building AI into sensors (edge computing) enables targeted data collection and preprocessing that reduces data volumes while simultaneously targeting the ideal measurements for specific science questions.

These opportunities apply not only to existing observation systems but also to the design of new facilities and instruments. BER recently initiated the design and creation of three Urban Integrated Field Laboratories (U-IFLs). These U-IFLs in Chicago, Baltimore, and Texas represent opportunities in the application of AI to urban science, for instance, employing surrogate models to build rapid-running regional climate models that will enable urban planners to evaluate many potential interventions in addressing climate change impact on urban communities. Such a surrogate model could potentially lead to the use of an urban climate-specific foundation model trained on the diverse and extensive volumes of data spanning regional weather and climate models through remote-sensed land surface temperature to traffic movement and socioeconomic and demographic data.

The tight coupling of experimentation with AI/ML models could also provide effective guidance in the bioengineering process to produce renewable bioproducts. Next steps include the development of integrated and explainable AI-driven models of complex biological systems that encompass all omics, structural, phenomic, and environmental layers of information. Finally, combining mechanistic and machine learning models will increase the accuracy of both approaches.

There is also an opportunity for AI, and Surrogate models in particular, to revolutionize the nuclear data pipeline, wherein data are compiled, evaluated, processed, and validated for end-user applications [9]. The existing pipeline is the result of human-intensive efforts; AI can be used to automate this process as well as to improve the fidelity of the resulting data.

7.1.2 AI FOUNDATION MODELS FOR SCIENTIFIC KNOWLEDGE DISCOVERY, INTEGRATION, AND SYNTHESIS

There is an opportunity to build general large-scale AI models that can be applied to a wide range of downstream programs and priorities in the materials and chemistry domains targeting structure and property predictions. In areas such as natural language and image processing, the development of large-scale models has revolutionized the way AI is applied, shifting from many bespoke and task-specific models to the use of one large-scale model that can be refined with a small amount of additional data to carry out many specific tasks. Implementing this approach in the biology, materials, and chemistry domains would be transformational for BES and BER priorities, creating core capabilities that could be reused across programs and improved over time. Using the chemistry domain as an example, potential downstream tasks enabled by a single large-scale model would include predicting properties of complex organometallic molecules that could lead to the discovery of more efficient catalysts and better separation technologies for rare earths, greatly improving our understanding of actinide chemistry or developing better electrolytes for electrochemical and energy storage systems.

7.1.3 AI FOR ADVANCED PROPERTY INFERENCE AND INVERSE DESIGN

Beyond accelerating the discovery of novel materials and molecules, the use of foundation models for property inference and inverse design would deepen our fundamental understanding of the connection between composition, structure, and properties, leading to new insights that would otherwise be extremely difficult to “tease out” of our traditional research and development (R&D) approaches and modeling and simulation.

Problem domains involving anomaly detection [4], fast surrogates [5], interpretability, uncertainty quantification (UQ) [6], searches and inverse problems in high-dimensional spaces, and AI-based control and optimization of complex systems (e.g., accelerators, detectors) are lively research areas and constitute typical open opportunities for AI in the near future.

7.1.4 AI-BASED DESIGN, PREDICTION, AND CONTROL OF COMPLEX ENGINEERED SYSTEMS

HEP science has a major focus on UQ [6], and the use of high-fidelity digital twins is already widespread and growing, providing examples and insight across SC programs.

For magnetic confinement fusion (tokamak) reactor R&D, the use of AI to predict and control plasma states in magnetic fusion energy (MFE) and inertial confinement fusion (ICF) systems has the potential to significantly improve our ability to optimize fusion performance. In turn, high-fidelity plasma predictions could be used to design improved facilities and operations. With the use of AI-enhanced modeling, HPC systems could be leveraged as real-time assets.

AI-driven modeling, design optimization, and diagnosis could also fundamentally advance capabilities for control and optimization of high-repetition-rate inertial fusion energy (IFE) facilities. Success in this area will entail the integration of AI capabilities across compute scales—from edge/diagnosis, through orchestration, to HPC, with the goal of executing at increasing scale, eventually up to a fusion pilot plant.

7.1.5 AI AND ROBOTICS FOR AUTONOMOUS DISCOVERY

The criticality of these new approaches for BES and BER applications is outlined within Chapter 05. HEP is another example where AI-based autonomous discovery and robotics capabilities are critical in that research is highly data-driven, with deep theoretical roots and some of the most complex engineered systems in the DOE complex. For instance, data rates and volumes in all major current and future HEP experiments already require heavy use of automation and are ripe for the exploitation of transformative AI techniques in the coming decade and beyond; indeed, a substantial community has recognized the opportunity and is working actively in this direction [7, 8], laying the groundwork for the necessary advances and new approaches outlined in Section 01 of this report, including the use of autonomous discovery and control capabilities.

7.1.6 AI FOR PROGRAMMING AND SOFTWARE ENGINEERING

Nearly every SC science program will benefit from AI-accelerated software engineering, particularly to mitigate disruptions from computer architecture changes as well as to integrate new, AI-enabled laboratory instruments and facilities with computational and data infrastructure.

7.2 Challenges to Overcome

The AI opportunities listed above in Subsection 7.1 highlight many specific challenges to be overcome; these center around model development (including explainable AI, faithfulness, validation, composability, and multi-scale), datasets (collecting, curating, storing, and making them usable for the community), and integration (with existing scientific facilities, instruments, and software, including issues of access, instrumentation, steering, interoperability, and adaptability). We discuss each of these below.

7.2.1 MODEL DEVELOPMENT

In order to advance scientific understanding, AI models must be grounded in the rules of nature. Multiple techniques for creating AI constrained by known biological and/or physical principles have been proposed and are discussed in Chapter 01; however, the field is young and needs significant attention. Concurrently, the "black box" nature of AI models confounds our ability to validate the results, hindering adoption. This challenge is also outlined in Chapter 12 (Mathematics and Foundations). Nascent methods for interrogating internal AI states for physical relevance have shown promise. Investments in efforts to maintain physical relevance and translate what AI has learned into physical understanding are essential to fully unlock the potential of new AI models.

Additional advances are needed in digital twins, discussed in Chapter 04. Increasing their faithfulness to actual systems will help ensure that digital twins can reliably advise and eventually control system operations. The wider application and usability of digital twins will also need to be expanded, ideally to the point where users and operators can virtually predict the operation of a planned experiment. Such expansion of digital twins will help enable AI-based autonomous discovery.

Building and optimizing what is effectively a digital twin of the lifecycle of a component/process will require modeling from the molecular to the fully operational system. This multi-scale task will require significant effort in developing the models and coupling across the scales. Developing AI methods that achieve this outcome will require new approaches; and for these approaches, we must also create uncertainty quantification that works across the necessary scales and maturity levels. Moving from synthesis to manufacturing will be facilitated if the original design and synthesis are informed by the subsequent manufacturing processes.

A longer-term challenge is the development of a reference library of production-quality AI models that can be composed in turn to build large foundation models. Although initial success has been documented applying foundation models to material science [10], considerable research remains to understand and evaluate strategies to effectively apply the concept of large-scale/master/foundation models to the materials and chemistry domains at scale. This research is needed in areas including the models themselves and the definition of the right input space, as well as the self-supervised learning methods required to maximize the usefulness of unlabeled data.

There are fundamental challenges to the adoption of any of the AI approaches from a validation standpoint. For many tasks, there is limited data (e.g., design of materials for controlling degradation, or predicting material behavior in extreme environments), and we will need to understand how we can maximize our ability to transfer broader knowledge contained within the representation of the large-scale model into these downstream tasks.

7.2.2 DATASETS

The widespread adoption of new AI methods in SC research and program areas will require high-quality, curated, AI-ready datasets; however, today in nearly every area of applied AI science, there is a dearth of available datasets for training and verification. Additionally, benchmark datasets are vital across SC science areas, as summarized in [11] and [12].

The FAIR (findable, accessible, interoperable, and reusable) data principles are directly relevant to these data needs that span every AI application area. These must be applied not only to training data but also to the training process and the models themselves. Moreover, large amounts of data generated by experimental facilities will need careful curation, provenance tracking, and storage before being used to train the AI. Despite DOE's leadership in traditional modeling and simulation resources and expertise, the necessary infrastructure for developing and adopting AI methods—both in terms of infrastructure and humans trained in data science—is currently lacking in the DOE complex.

The size and complexity of HEP's, BER's, and other SC programs' datasets are considerable; however, the migration from traditional modeling and simulation to training and expanding AI models adds new dimensions. For instance, a significant challenge is the availability of the data management and computational infrastructure needed to support training/inference applications at large scales. Open, curated experimental and observational datasets will need to be processed (e.g., tokenized) for use in model training and to be available for meaningful collaboration with the broader applied mathematics and data and computer science communities in academia, industry, and national laboratories. The data infrastructure requires not only the generation of large datasets for quick visibility, but also that the datasets encode sufficient meta-data to enable labeling, data integration, and provenance tracking for reproducibility. Here, open data from surrogates based on high-fidelity simulations will be essential for training and validating AI techniques: examples include simulations of detailed future detector designs and synthetic sky maps based on large-scale cosmological simulations.

7.2.3 INTEGRATION

To enable autonomous workflows that can incorporate many levels of AI at scientific user facilities, scientific instruments will require abstraction layers for functions such as operational control, experiment configuration, and data routing. The resulting data flows from instruments, as well as their inputs (i.e., beams), need to be accessible, with analysis carried out in some cases with very low latency (edge detection), as well as for use asynchronously, such as to train models.

Nuclear theory, environmental models, and other SC programs' software have advanced significantly under programs such as ASCR's Scientific Discovery through Advanced Computing (SciDAC). However, these novel software stacks have been developed within the context of traditional simulation and modeling, and consequently are typically not AI-ready. For example, additional work will be needed to ensure that codes are endowed with the automatic differentiation and uncertainty quantification capabilities necessary to accelerate AI-based development. Many of the advanced computing code bases have been developed in relative isolation. A focus on software interoperability throughout the SC community would significantly benefit AI efforts for the creation of surrogates, inverse design, etc.

Facility operation and control systems, whether for nuclear physics, environmental observations, or infrastructure such as ESnet, are highly complex. Current optimization and control efforts have focused on individual components, typically tuning a small number of parameters based on fast diagnostics. Coupling multiple components, and tuning their interplay, will require greater interoperability throughout a facility's systems. Furthermore, these systems will evolve over time, and hence training data will be equally dynamic. Advances are needed to provide additional machine-ready hooks for AI methods to diagnose changes and react accordingly.

From the theory and computation perspective, many forms of computation for autonomous steering of user facilities will require developing and deploying rapid decision-making algorithms, as well as addressing issues related to task scheduling under resource constraints as detailed above (in this "Integration" subsection), as well as in Chapter 13.

In fusion energy, the U.S. is rather "experimental facilities poor"; and the facility time and access necessary for AI innovation and exploration will exacerbate the need for such infrastructure, as well as for AI-ready instrumentation that can interface with the broader AI ecosystem. The integration of automation with AI/ML computational techniques will also require deeper collaborative efforts across domains and scales, such as in BER to bring together biological research, data science, computer science, and engineering.

7.2.4 GENERAL

Traditionally, scientists are trained either in AI-related disciplines (math, computer science, etc.) or domain-specific disciplines (e.g., biology, chemistry, physics, earth science, etc.). SC programs have increasingly encountered the need for expertise in both AI-related and domain-specific fields to fully leverage AI. Communication and education between AI and other domain experts are thus of utmost importance and can be encouraged through funding calls requiring co-participation and workshops aimed at bridging this gap. Here, the DOE national laboratories have a long history of collaboration among domain-specific disciplines and mathematics, computer science, and computational science. This unique DOE strength will be instrumental in fostering collaborations related to AI and domain-specific disciplines and facilities.

Another significant challenge becomes apparent when reviewing the many science programs, where traditionally independent vertical approaches come at the expense of fragmentation. The magnitude of intellectual and resource investment needed to move from traditional modeling and simulation to the use of AI models and methods will demand new approaches to collaboration, with much larger scientific teams spanning SC domain and ASCR programs and user facilities.

7.3 Investment Needed for Achievement

In order to realize the many exciting scientific opportunities outlined herein, investments are essential to address the challenges described above—effectively representing a roadmap for ASCR co-design with other SC offices. We organize investment needs around (1) AI methods and datasets, (2) self-driving laboratories, and (3) critical partnerships. Each subsection contains bulleted descriptions of the programs needed to fully harness the potential for AI across SC.

7.3.1 AI METHODS AND DATASETS

Necessary AI methods and datasets include the following examples:

- Better data acquisition, curation, and utilization. Designing and training AI capabilities in FES will require massive amounts of data. From large ensemble simulations to AI-ready instrumentation of facilities, data acquisition and its curation require an immediate effort to understand the requirements unique to FES, as well as a sustained investment in preparing AI-ready instrumentation and simulation workflows to acquire, curate, and distribute the needed data to the communities most capable of driving AI innovation.

- Increasing the number and fidelity of digital twins for nuclear physics instruments, experiments, and facilities.
- Advancing current computational software to be AI-ready through differentiable and probabilistic programming.
- Expanding natural language processing efforts to extract semantics from documents pertaining to nuclear data.
- An Environmental AI Data Library: Creating curated and easily accessible (application programming interface [API]) datasets for training, etc., such as on global storms to bacteria and beyond.
- Coupled with datasets, creating data proximate compute and notebook-based workflows that incorporate ways to increase the FAIR-ness of AI analyses.

7.3.2 SELF-DRIVING LABORATORIES

Necessary components for self-driving laboratories include the following:

- Environmental AI testbeds at the edge. Edge computing nodes connected with simple and advanced sensors at DOE labs and facilities—such as ARM and EMSL—allowing AI research with active and configurable sensors to test new ideas alongside baseline measurements. Such testbeds would be a BER version of self-assembling laboratories. Next-level investments could be made in mobile autonomous data collection, including an unmanned aerial vehicle (UAV) facility for adaptive sensing of the atmosphere, Earth system, and biosphere.
- Self-driving labs that couple robotics for automated experiments and data collection [13], with AI systems that use these data to recommend follow-up experiments.
- Digital infrastructure as well as edge computing fabric to enable integration, such as to drive interactive, AI-driven experiments at facilities and remote locations.
- Abstractions to enable experimental theory-coupled workflows to be fully defined in a coding language. These abstractions should enable the automation of specific tasks in synthesis and characterization instruments.
- The need to invest in cross-disciplinary research, including the need for facilities (such as autonomous laboratories) to explore, validate, and test approaches.

7.3.3 CRITICAL PARTNERSHIPS

Necessary critical partnerships include the following:

- Increased engagement in facilities (U.S. and international). As AI methods improve in providing sophisticated control and fault prevention, integration of these new methods and their validation will require a vibrant ecosystem of pilot facilities, and the continued involvement of the FES community with international collaborations (e.g., ITER). Co-design both with material science efforts and public-private partners will also be required to establish an AI-driven U.S. pilot plant and energy dominance in 10 years.
- Centers for co-design. Institutes where domain scientists are partnered with AI experts to attack a distinct and well-defined science question. Chosen by application (such as a user facility), principal investigators (PIs) would begin with an intensive three-month (nominally) engagement in person and an extended, less intensive engagement for over a year. The outcomes should be tools, datasets, and publications. In addition, materials studies should focus on the *how*, thus aiding in the reproducibility and reusability of techniques.
- Continued close partnership between SC domain programs and ASCR**.** Investments in theory and computation are vital to the continued development of complex validated models, from data acquisition and curation to improvements to modeling capabilities. A continued close partnership is needed between FES and the advanced computing community to ensure that new methods in real-time control, UQ, and AI surrogates are used to improve the material, design, and control system of FES facilities (including pilot plants).
- An essential requirement is a data and compute infrastructure that has the flexibility to support both large individual projects and many exploratory forays. Substantial investment will be needed to establish a number of joint programs (across ASCR and HEP) to build up and maintain curated datasets. These datasets will include supporting software that allows for data interpretation and reduction and thus ingestion by an AI model. An organized investment plan for software development and sustainability (Exascale Computing Project [ECP]/SciDAC-like focused programs) targeted to specific opportunities and challenges mentioned above will need to be developed. At least some fraction of this investment will need to be made at the facilities to manage specific issues for the HEP community, such as the diversity of AI platforms.
- The benefits of diversity, equity, and inclusion (DEI) across SC programs are clear in terms of the quality and breadth of data, ideas, and strategies. Concurrently, a focus on environmental justice recognizes that the brunt of impacts from challenges such as climate change and energy security is disproportionately borne by these communities. The importance of these programs to DOE's continued scientific leadership and service to the nation is detailed in Chapter 16.

7.4 References

[1] Rosner, J., et al., 2013. *Planning the Future of U.S. Particle Physics: Chapter 1: Summary*, arXiv:1401.6075.

[2] FESAC (Fusion Energy Sciences Advisory Committee), 2022. *Powering the Future: Fusion & Plasmas*. https://science.osti.gov/-/media/fes/fesac/pdf/2020/202012/FESAC_Report_2020_Powering_the_Future.pdf, accessed December 2, 2022.

[3] Office of Science, 2016, *Exascale Requirements Review: Nuclear Physics*. https://exascaleage.org/wp-content/uploads/sites/67/2017/05/DOE-ExascaleReport_NP_R27.pdf, accessed December 2, 2022.

[4] Nachman, B., 2020. Anomaly detection for physics analysis and less than supervised learning. arXiv:2010.14554.

[5] Butter, A., and Plehn, T., 2020. Generative networks for LHC events. arXiv:2008.08558.

[6] Chen, T., et al., 2022. Interpretable uncertainty quantification in AI for HEP. arXiv: 2208.03284.

[7] Shanahan, P., et al., 2021. CompF3: Machine learning, Snowmass 2022 Report. arXiv:2209.07559.

[8] HEPML-Living Review, undated. *A living review of machine learning for particle physics*. https://iml-wg.github.io/HEPML-LivingReview/, accessed December 2, 2022.

[9] Boehnlein, A., et al., 2022. Colloquium: Machine learning in nuclear physics. *Reviews of Modern Physics* 94 (3). 10.1103/revmodphys.94.031003

[10] Hatakeyama-Sato, K., and Oyaizu, K. 2020. Integrating multiple materials science projects in a single neural network, *Communications Materials* 1(1): pp. 1–10.

[11] Dueben, P.D., et al., 2022. Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook. *Artificial Intelligence for the Earth Systems* 1(3), e210002. https://journals.ametsoc.org/view/journals/aies/1/3/AIES-D-21-0002.1.xml, accessed December 2, 2022.

[12] Carbonell, et al., 2019.

[13] Martin, H.G., et al., 2022. Perspectives for self-driving labs in synthetic biology, arXiv:2210.09085.

08. ENERGY (EERE, OE, FECM, NE)

To function, modern society is critically dependent on large, networked, engineered, complex energy systems—some of which were outlined in Chapter 04. Such systems have scales ranging from individual buildings and facilities (e.g., power plants) to districts and metropolitan areas, to regional and continental (and combinations of these). They are designed to support society—for the environments where people live and work; for transport of commodities such as electric power, natural gas, oil, hydrogen, water, etc.; and for transport of goods and people using highways, public transit, rail, etc.

The importance, scale, and complexity of these challenges are reflected through the work of multiple U.S. Department of Energy (DOE) offices and programs. The Office of Energy Efficiency and Renewable Energy (EERE) is working to build a clean energy economy that benefits all Americans, with programs including energy efficiency, renewable energy, and sustainable transportation. The Office of Electricity (OE) works with industry and other stakeholders to ensure that the Nation's electricity delivery system is secure and resilient to disruptions. The Office of Fossil Energy and Carbon Management (FECM) focuses on minimizing the environmental impact of fossil fuels while working towards net-zero emissions, with programs encompassing carbon capture, management, transport, and storage as well a critical minerals carbon dioxide removal, carbon conversion, and methane mitigation. The Office of Nuclear Energy (NE) advances nuclear energy science and technology through innovation in continued operation of existing U.S. nuclear reactors, deployment of advanced nuclear reactors, development of advanced nuclear fuel cycles, and maintaining U.S. leadership in nuclear energy technology.

Unfortunately, disruptions to energy supply are becoming more frequent and serious, driven by factors such as: (1) an energy system that is becoming more complex, interdependent, and less stable with the addition of renewable and co-generation sources; (2) more intense and more frequent extreme weather events; and (3) inadequacies in tools (extensions, monitoring, and control) for managing these systems. The status quo has led to poor and costly decision making, wasted resources, slow recovery from interruptions, suboptimal planning decisions, and susceptibility to catastrophic disturbances and cascading failures. Indeed, each year yields new cycles of *reactive* reports highlighting challenges and lessons learned—underscoring the fact that better planning, improved predictions, and enhanced response could have significantly improved the outcomes that were experienced. Two recent and notable examples are Hurricane Sandy in November 2012 and the February 2021 "arctic blast" that disrupted power throughout Texas [1]. Regarding the former, the North American Electric Reliability Corporation's (NERC's) *Hurricane Sandy Event Analysis Report* remarked that many entities had challenges with system control, both during the storm and during restoration, balancing loss of load with loss of generation, all of which may have contributed to the sizes and lengths of power outages that affected populations experienced. At the storm's peak, 8.35 million customers were without power, some of whom were without power for a month [2].

Such situations are exacerbated by the increasing interconnectivity within our energy infrastructure (e.g., natural gas and electricity systems) as well as with other infrastructure systems, such as communication and

PROJECT SPOTLIGHT

Project Name: Automated and scalable active ensemble machine learning frameworks for rapid optimization of product design and manufacturing processes

PI: Pinaki Pal

Organizations Involved: Argonne National Laboratory; Parallel Works, Inc.; Convergent Science, Inc.; Aramco Americas

Goal: Develop automated and end-to-end workflows coupling active machine learning (ML) and simulations for rapid optimization of product design and manufacturing processes.

Significant Accomplishment: Argonne National Laboratory developed, demonstrated, and commercialized (through adoption by industry partner Parallel Works, Inc.) a unique ML-genetic algorithm (ML-GA) software technology that integrates ML-based ensemble surrogate models and active learning within an adaptive, automated, portable, and scalable framework to accelerate virtual design optimization campaigns by an order of magnitude (from months to days over current industrial state-of-the-art approaches).

In the News: Awards include the 2021 R&D 100 Award (Software/Services category) and 2021 HPCwire Readers' Choice Award for Best Use of High Performance Data Analytics & AI. Also: O. Owoyele, P. Pal, A. V. Torreira, D. Probst, M. Shaxted, M. Wilde, and P. K. Senecal, 2022. "Application of an automated machine learning-genetic algorithm (AutoML-GA) to engine design optimization based on computational fluid dynamics simulations," *International Journal of Engine Research*, Vol. 23 (9), pp. 1586–1601.

transportation, and the expansion of new infrastructure systems such as those that support electric vehicles and the emerging hydrogen economy to improve the nation's energy independence. These factors point to the need for artificial intelligence (AI) systems that *proactively* predict, mitigate, and prevent extreme scenarios that are experienced today, and future scenarios that will emerge as the nation's future energy system evolves. For example, as the nation's transportation electrifies, how can AI for the grid plan and respond to an increased need for charging during extreme events that require evacuation?

Moving beyond the structure and complexity of integrated networked systems, the individual technologies and materials that are required to produce, store, and deliver energy each present unique challenges as they must meet simultaneous requirements for reliability, cost, resilience, and sustainability. Examples include new materials to increase efficiencies in solar photovoltaics, sensors for monitoring the health of energy components, power electronics, new materials for energy storage, new fuels, and materials for harsh environments such as those inside nuclear reactors. Developing new materials and technologies is currently costly and time consuming, with limited guarantees that investments will yield the desired payoffs.

Here, advances in biotechnologies would provide alternative, sustainable fuels for transportation requirements that are difficult to achieve with electricity (aviation, heavy freight, etc.) [3]. Advances in storage technologies, such as battery materials, would reduce the cost of utility-scale storage to a level where these technologies would become an attractive alternative to fast-ramping fossil fuel generators required to manage variability in renewable energy resources. And finally, new advances in manufacturing processes and supply chains would support rapid and efficient deployment of technologies as they become available. In all these examples, the application of new AI methods will enable researchers to examine extremely large, complex, and multivariate problems in ways not possible today, catalyzing new discoveries in materials and manufacturing that are necessary for transformational energy technologies.

Within the U.S. Department of Energy (DOE) applied energy offices—its Office of Energy Efficiency and Renewable Energy (EERE), Office of Electricity (OE), Office of Fossil Energy and Carbon Management (FECM), and Office of Nuclear Energy (NE)—there are significant programs that are seeking to address each of these factors and dimensions—both individually and in combination—by leveraging AI and related technologies.

- **EERE** programs have long sought to utilize AI systems to improve predictive models for energy output from variable and uncertain renewable energy sources, such as wind and solar, to support reliable, resilient, and extensive adoption of clean energy solutions. AI approaches have also been used to assist in efficient and grid-responsive operation of buildings. EERE/Advanced Manufacturing Office (AMO) is championing next-generation ("beyond CMOS" – that is, complementary metal oxide semiconductor) microelectronics to support energy-efficient processing and control of energy generation and transport systems by exploring AI-enhanced co-design of new electronic devices, components, and computing systems. Likewise, there is significant emphasis within the DOE-Vehicle Technologies Office (VTO) and DOE-AMO programs to leverage AI/machine learning (ML)-based surrogate models and algorithms (see Chapter 01) for rapid, high-dimensional design optimization of novel fuel-engine systems and manufacturing processes, respectively. Finally, EERE/Bioenergy Technologies Office (BETO) has funded the pioneering use of AI and ML to enable biodesign of cells for renewable biomanufacturing in the form of the Agile BioFoundry (ABF) [4].
- **OE** initiatives, such as the Smart Grid, Microgrid R&D, Advanced Grid Modeling, Transmission Reliability, and Energy Storage programs, have sought to leverage AI to construct predictive tools that anticipate when extreme weather will induce grid disruptions, with the objective of utilizing such predictions to improve operator response and thus limit the impact of such disruptions. This effort involves enhancing grid resilience to enable decarbonization while simultaneously enabling resilience to extreme events. Here, AI is used to analyze data from multi-domain (e.g., gas, electric, and wind) infrastructure to understand interdependencies across infrastructure assets and to minimize the impact of extreme events on the grid. AI is also being used to help improve the observability of the electric grid, particularly in the context of limited data sources or missing data. OE's initiatives also depend on predictive, high-reliability electronic hardware to improve the resiliency of the grid, where AI-enhanced co-design has driven developments in next-generation grid hardware infrastructure [5].
- **FECM** programs—such as the Science-informed Machine Learning for Accelerating Real-Time Decisions in Carbon Storage Applications (SMART-CS) project—seek to dramatically reduce the climate impact of fossil-fuels-based generation by harnessing AI to enable efficient, stable, and effective management of subsurface reservoirs for secure carbon storage. Within FECM's Advanced Turbines Program, efforts are underway to enable 100% hydrogen-fueled gas turbine engines for decarbonization of the stationary power generation sector. However, these energy systems are prone to catastrophic failure from rare combustion events (such as flashback, thermoacoustic instabilities, etc.). Consequently, the development and deployment of AI systems, such as surrogate models described in Chapter 01, are sought for automated discovery/assessment of causalities behind these rare

events and for developing predictive control strategies to prevent their occurrence.

- **NE** is seeking to develop new and advanced reactor designs; design, selection, and manufacturing of materials for nuclear systems; and flexible controls to manage the overall lifecycle of nuclear power technologies. AI-based capabilities throughout these activities have the potential to lower capital costs, reduce ongoing operations and maintenance costs, allow nuclear energy to meet emergency needs for energy (e.g., electric power after extreme events), and balance the requirements of clean energy policies. Of particular importance is the development of "digital twins" (virtual models of operating nuclear power systems, structures, and components), as detailed in Chapter 04, that reflect the real-time system state and may be applied toward developing solutions for the challenges, ranging from real-time controls to long-term planning, as discussed elsewhere in this chapter.

While not exhaustive, these examples spanning DOE's applied energy office programs highlight the complex interplay between the nation's interconnected and interdependent energy systems. Figure 8-1 illustrates the inherent complexity that crosscuts the applied energy offices. The figure highlights how decisions and disruptions within any one of these systems have the potential to influence, and in the case of disruptions, cascade through other systems, causing catastrophic events where recovery can take days or weeks and at significant economic cost [1].

AI capabilities such as those detailed in Section 01 of this report are becoming an increasingly attractive solution for managing the complexity in modeling, predicting, operating, controlling, and planning these systems [6], both in isolation and from a system-of-systems perspective characterized by the dynamics of their interconnections and interdependencies. Below we identify some of the major open challenges in energy that are central to DOE's applied energy missions in the offices of EERE, OE, FE, and NE. We focus on those challenges where expected advances in foundational and crosscutting AI capabilities—beyond what a single office can support—will play a critical role in providing solutions to these challenges.

8.1 Open Opportunities

One of the core opportunities for AI systems in the energy domain is to support the modernization of the nation's integrated energy delivery system to simultaneously achieve

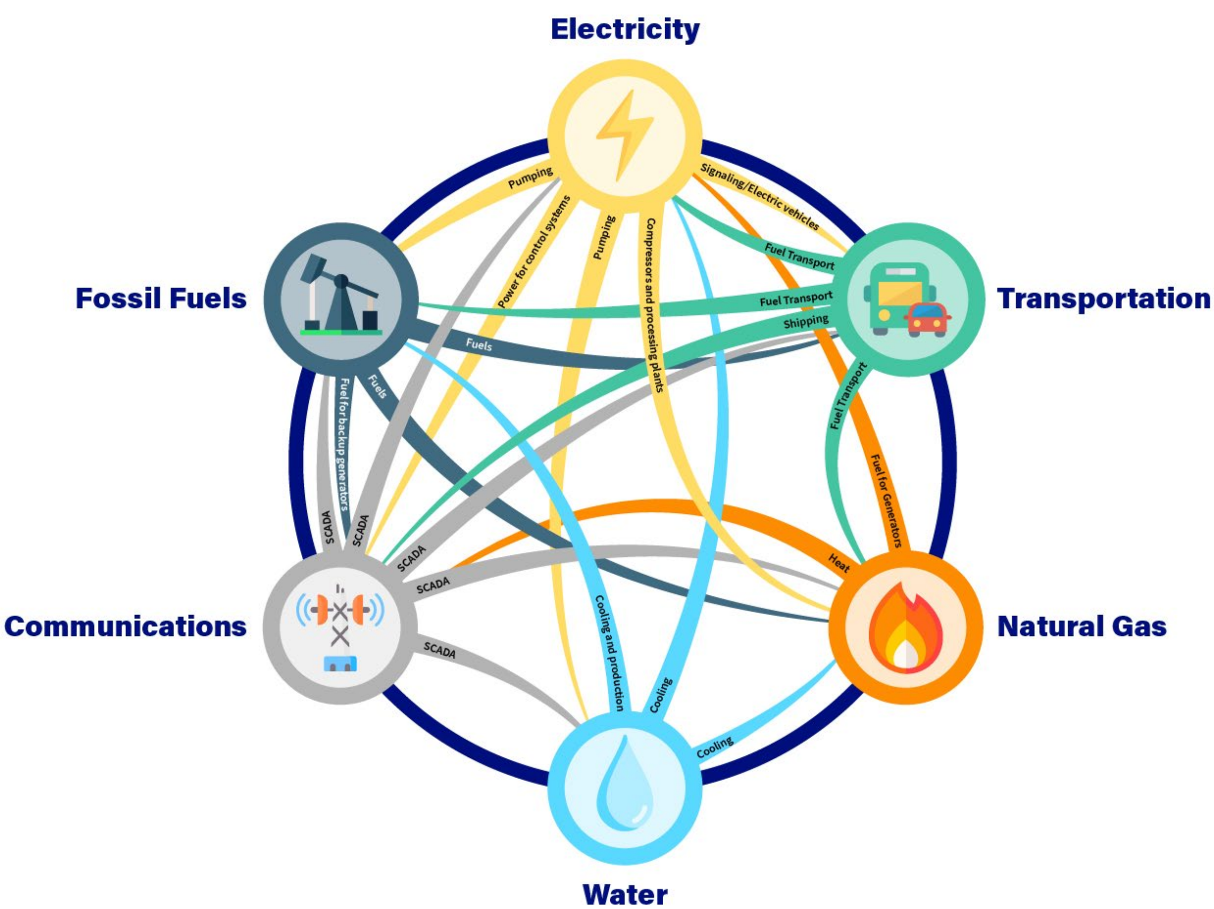

Figure 8-1. The complex interdependencies of the electric power grid, gas, oil, transportation, and communication, and emerging energy sources such as hydrogen, that complicate modeling, predicting, and controlling energy systems.

affordability, carbon neutrality, reliability, and resilience to extreme (both natural and intentional) events beyond what today's system can achieve. A second opportunity is to not only maintain but to exceed today's expectations of energy reliability and low energy costs [7]. Moreover, AI can play an important role in ensuring energy equity and environmental justice through the optimization of new technologies and operations. To achieve these opportunities several key AI-enabled capabilities are required, including:

- **AI-Enabled Design of Control Systems.** Control theory has a long history of contributing to and impacting the operations of energy systems. AI presents new opportunities to form the backbone of next-generation control for seamlessly integrating heterogenous sensor platforms. These platforms would operate at varying timescales to ultimately yield self-composing and self-healing control that adaptively incorporates new devices, reconfigures itself during adverse conditions, and can recommend what devices and sensors are needed to improve performance. Such an AI control system approach, detailed in Chapter 04, also holds the potential to improve the reliability and resilience of modern energy systems through self-healing, distributed, and potentially multiscale control that leverages compute capabilities at the edge.
- **Trustworthy Decision-Making under Uncertainty.** Energy systems represent a high-consequence environment where the impacts of failure or inaction are potentially significant in terms of economics, loss-of-life, etc. AI systems as described in Section 01 have the potential to improve our ability to provide fundamentally robust and theoretically sound decisions for operating, planning, and maintaining energy systems, accounting for inherent uncertainties and being resilient to bad, missing, and adversarial data. Such AI capabilities, among the common requirements for systems described throughout Section 01, would allow energy systems to robustly handle high penetrations of variable and uncertain renewable energy and to secure energy systems from malicious actors.
- **Materials to Components Co-design.** Achieving decarbonization goals will require innovations that scale from components to integrated systems. Often, innovations in next-generation materials do not translate into functional components due to limitations in environmental, operational, and other requirements. When exploring the space of material designs, AI systems such as property inference and inverse design, discussed in Chapter 03, will support the discovery and evaluation of novel materials through co-design methods that account for system-level requirements (such as grid integration, operational reliability, lifecycle durability, etc.). Such frameworks will accelerate the development of technologies and materials for higher-efficiency solar photovoltaics with higher power density; component-level, in-situ sensors for monitoring operational health and observability; high-efficiency power electronics for converters and inverters; hybrid manufacturing of conventional and additive approaches for components; and harsh-environment electronics [8].
- **Load Forecasting and State Estimation.** An important objective of equitable energy infrastructure is its openness: the ability for the end user to have significant autonomy in how and when they use it. The entities responsible for load balancing and stability must be able to forecast the load mix and estimate the state of the system at places with low visibility. While the steady-state response is quite accurately forecasted, the dynamical one is far behind. This challenge will be exacerbated by increased fluctuations in voltage, amplitude, and frequency associated with the growing adoption of renewable generation, and by increasing privacy and security concerns. AI models—such as foundation models discussed in Chapter 02—that are trained using multimodal data, including anonymized smart infrastructure data, public infrastructure deployment records, existing infrastructure signatures, and new sources such as social media data, have the potential to provide unprecedented fidelity in load estimation. This capability will reduce average interruption times, improve situational awareness, and significantly improve reliability.
- **Federation and Privacy.** In the operation of energy infrastructure, data access remains a major concern, driven by the multi-stakeholder nature of energy infrastructure and data and concerns about security, privacy, and market integrity. One promising approach is to develop distributed, federated AI-based mechanisms that guarantee a high level of privacy and that approach or, ideally, exceed the performance of centralized data analysis systems.

Meeting these opportunities will allow the nation to reduce, if not eliminate, climate impacts induced by energy production, transport, and consumption, while potentially saving billions of dollars in outage costs.[1]

8.2 Challenges to Overcome

The adoption of AI capabilities in energy systems to harness these opportunities will require addressing the following challenges:

- **Scalable Computation.** The combinatorial control and design space of energy systems is impossible to explore with current techniques, and the interconnected systems

1 For example, it is estimated that the 2003 power blackout that originated in Ohio and spread across much of the Northeast cost $10 billion [9]. More recently, the South-Central United States cold weather outage in 2021, which had impacts spreading between the natural gas and electric power systems, had economic impacts estimated to be as high as $130 billion in Texas [10].

result in large-scale coupled systems that are computationally intractable and too complex to fit into existing combinatorial optimization modeling and solution tools. Thus, AI systems will need to overcome these and other fundamental scaling challenges for energy control and prediction as outlined in Chapter 01 regarding surrogate models.

- **Validation and Verification of AI Methods.** Because of the high consequences of energy systems failure, new AI approaches, models, and tools will require formal validation and verification (V&V) of correctness throughout the life cycle of data and associated model development. These challenges are discussed in detail in each of the chapters in Section 01 as well as in Chapter 12, Mathematics and Foundations.
- **Uncertainty-Aware Robust AI Systems.** For AI to provide solutions in the energy domain, an AI system is required to make provably robust inferences and recommendations locally (e.g., at the edge of or within a subsystem of an energy system) and globally (e.g., centralized operations), with human-understandable explanations for why the AI makes the decisions it does. Moreover, the AI must account for and characterize the uncertainties in measurement data and forecasts when making decisions and to certify that it is resilient to interference (natural or adversarial). This is a requirement across decision applications in energy systems, ranging from control systems for power grids with mixes of centrally dispatched generators, locally controlled distributed energy resources (DERs), as well as control systems for operating pipeline systems (natural gas, petroleum, carbon dioxide [CO_2], hydrogen, etc.), handing uncertainty in renewable generation sources, and optimizing the operation of sophisticated reactors. As with V&V, uncertainty, explainability, robustness, and related requirements are discussed throughout Section 01 regarding capabilities and in Chapter 12: Mathematics and Foundations.
- **Adaptative and Self-Configuring AI Systems.** Integrated energy systems are evolving systems with increasingly large numbers of sensors and devices being added over time. Sensors and other devices have controllable phenomenology that occur at the multiple timescales of decision-making in energy systems—ranging from sub-second frequency control to decadal capital investments. Thus, for AI systems to provide planning, optimization, and control solutions to energy, they must respond quickly enough to match the scales of the phenomena, have an implicit understanding of the domain (e.g., physics-informed constraints), and ultimately become self-composing optimization and control systems that adapt to the changing conditions, environment, and configurations of an energy system over appropriate timescales. Chapter 04 discusses these factors at length.
- **Data Sensitivity and Curation.** AI methods require large amounts of labeled, curated data to be effective. Although energy system sensor arrays generate large volumes of data, there is misalignment between the input data required by typical AI models and the data that energy systems can provide. First, energy data are not typically well labeled nor centrally collected, requiring that AI methods work with partially structured data collected and stored in an accessible platform. Research in foundation models (Chapter 02) suggests the potential for emergent capabilities in large-scale models with respect to self-supervised learning from large, unstructured, and multi-model data sources. Second, energy data often have protection requirements that limit how widely they can be disseminated, presenting AI challenges including anonymizing data and developing federated or shareable AI models that are non-invertible (e.g., they cannot be used to recover the information that was used to develop the AI model). Third, energy data includes measurements with high degrees of uncertainty and incompleteness, with missing data for significant periods. Similarly, these data may or may not include low-frequency, high-consequence events, increasing the potential for misinterpretation of phenomena that are absent in the training data.

8.3 Investment Needed for Achievement

Investments in AI capabilities applied to energy challenges provide long-term as well as immediate benefits. For example, AI capabilities such as those in surrogate (Chapter 01) and foundation (Chapter 02) models provide the opportunity for a paradigm shift away from traditional optimization solutions that, while trusted, are inadequate for today's (and certainly future) energy systems. The interpretation of simulations on the scale of the electric grid is difficult, and even more challenging for integrated energy systems with sector coupling, but AI systems such as inverse design models (Chapter 03) and digital twins (Chapter 04) can provide system designs and operational capabilities that improve the ability of decision-makers, policy-makers, and stakeholders to identify relationships that are non-intuitive, opaque to human observation, or beyond the view of traditional solutions such as correlation techniques.

At present, humans build system models that they can intuitively understand, which are thus limited in size, scope, and complexity and in the questions that they can address. The development of AI surrogates (Chapter 02), in turn enabling digital twins (Chapter 04), is one immediate path toward accelerating and scaling the modeling and simulation development of energy systems, concurrently reducing reliance on specialized subject matter experts. Finally, investments for AI capabilities to support real-time decision and control (Chapter 06) could target replacing more complex simulation models. Here, the discussion of surrogates in

Chapter 02 is an appealing approach, as it is targeted on surrogates for the complex physics and dynamic phenomena of energy transport over networks. Further, investments are needed to create explainable and interpretable methods and surrogates for interconnected energy system that integrate streaming, multi-modal, and multi-fidelity data. These AI systems can act as closure models that integrate and transform the inherently multi-modal data of energy systems into new models.

Over the long term, several investments will be essential. First, there is the need for investing in AI for intelligent and composable control systems (Chapter 06). Key investment areas include in the ability to continuously adapt to changing and increasing numbers of sensors that generate information at different time intervals as well as in AI to support decision-making that occurs quickly enough for the temporal scales of the phenomenology of the energy system under control. Moreover, such control is inherently constrained by legacy systems that must be combined with more modern technologies. The investments need to leverage the compute power emerging at the edge of energy systems (e.g., with intelligent sensors that can both process data and actuate controls without the hundreds of millisecond data propagation and processing delays involved in centralized control systems), with the result being AI-enabled, distributed monitoring and operations. The long-term goal of this investment is self-composing AI control systems.

Second, there is a need for AI investments in trustworthy decision-making under uncertainty. This is inherently critical for provably robust decision making, providing both intuitive, human-interpretable, investment-grade explanations and resilience to adversarial attacks. Associated development of metrics for quantifying trust in an AI model, including AI explainability, are also centrally important. Together, these developments are necessary to provide quantitative and qualitative means to certify AI model trustworthiness, as necessary for operational adoption. The investments noted elsewhere in this report (including Chapters 01, 02, 06, and 12) for trustworthy and interpretable AI are directly connected to this recommendation.

Third, there is a need for investments in AI for harnessing the vast and fragmented landscape of energy-systems-data. Efforts of identifying, acquiring, securing, curating, and contextualizing (encoding, compressing, and representing) the massive, multi-modal, heterogeneous, and rapidly growing data from energy systems spanning orders of timescales together constitute a computational science challenge that requires significant advances in the state of the practice.

Finally, the diversity in terms of timescales for the design, implementation, and operation of energy systems confounds the development of holistic, integrated design capabilities. For example, investments in energy systems are made at the scale of decades, whereas geothermal storage needs to be charged seasonally and daily, solar and wind energy need to account for days of scarce energy harvesting, and grid-responsive buildings—the consumers and prosumers in such integrated energy systems—need to manage loads at timescales of hours to minutes and seconds. Simply put, investments in developing AI capabilities across multiple scales of time and space is a computational and scientific challenge that requires focused research investment and demands new approaches and capabilities, particularly as described throughout Sections 01 and 03 of this report. Figure 8-2 illustrates how operational timescales (sub-second) interact with decadal decisions.

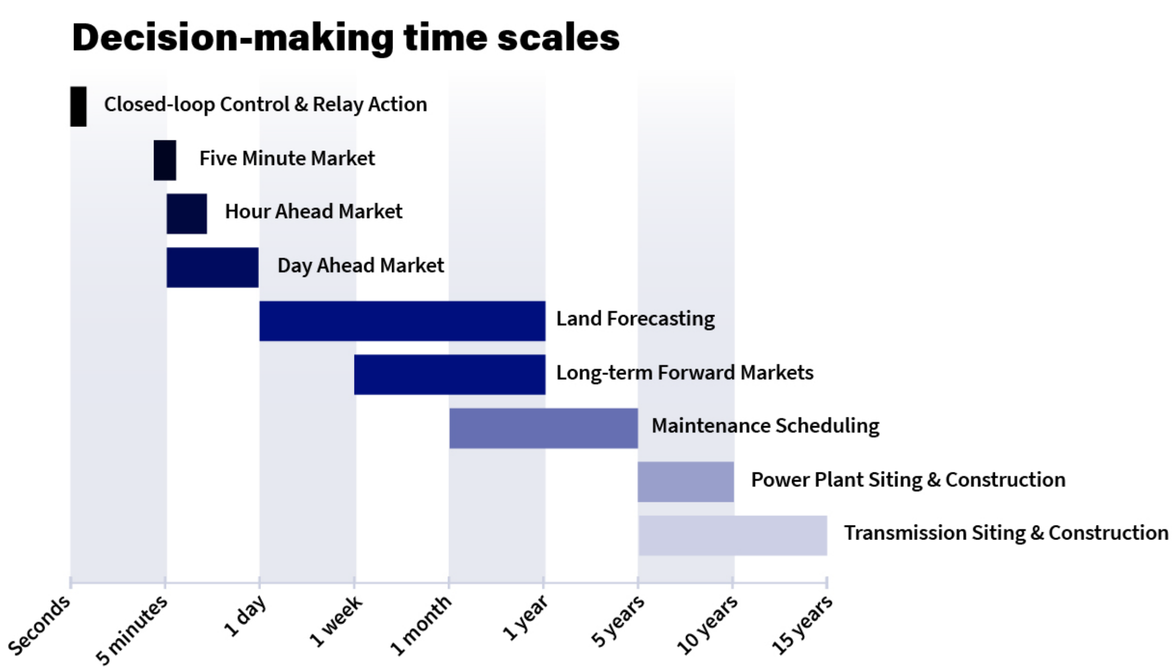

Figure 8-2. Examples of decision-making timescales for electric power systems that is inspired by the report [11].

8.4 References

[1] Bhattacharyya, A., and Hastak, M., 2022. Indirect cost estimation of winter storm–induced power outage in Texas. *Journal of Management in Engineering* 38(6), 04022057.

[2] NERC (North American Electric Reliability Corporation), 2014. Hurricane Sandy Event Analysis Report, January. https://www.nerc.com/pa/rrm/ea/Oct2012HurricanSandyEvntAnlyssRprtDL/Hurricane_Sandy_EAR_20140312_Final.pdf, accessed November 9, 2022.

[3] Argonne National Laboratory, Berkeley Lab, National Renewable Energy Laboratory, Oak Ridge National Laboratory, and Pacific Northwest National Laboratory, 2021. *Designing for Deep Decarbonization: Accelerating the U.S. Bioeconomy*. https://biosciences.lbl.gov/wp-content/uploads/2021/12/21-BAO-3054-Designing-the-Bioeconomy-for-Deep-Decarbonization-Report_v5.pdf, accessed November 9, 2022.

[4] Carbonell, P., Radivojevic, T., and García Martín, H., 2019. Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synth. Biol*. 8(7), pp. 1474–1477. https://pubs.acs.org/doi/full/10.1021/acssynbio.8b00540, accessed November 9, 2022.

[5] DOE-SC (U.S. Department of Energy Office of Science), 2018. *Basic Research Needs for Microelectronics: Report of the Office of Science Workshop on Basic Research Needs for Microelectronics*, October 23–25. https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN_Microelectronics_rpt.pdf, accessed November 9, 2022.

[6] EPRI (Electric Power Research Institute), 2019. *An Introduction to AI, its Use Cases, and Requirements for the Electric Power Industry*, August. https://www.epri.com/research/products/000000003002017143, accessed November 9, 2022.

[7] DOE-OE, 2022. *NOTICE of INTENT: Building a Better Grid Initiative to Upgrade and Expand the Nation's Electric Transmission Grid to Support Resilience, Reliability, and Decarbonization*, 6450-01-P. https://www.energy.gov/sites/default/files/2022-01/Transmission%20NOI%20final%20for%20web_1.pdf, accessed November 9, 2022.

[8] Descour, M., Tsao, J., Stracuzzi, D., Wakeland, A., Schultz, D., Smith, W., and Weeks, J., 2019. *Workshop Report: AI-Enhanced Co-Design for Next-Generation Microelectronics: Innovating Innovation*, Sandia National Laboratories, SAND2021-16012R. https://www.osti.gov/servlets/purl/1845383, accessed November 9, 2022.

[9] Natural Resources Canada and DOE, 2006, *Final Report on the Implementation of the Task Force Recommendations*, U.S.-Canada Power System Outage Task Force, September. https://www.energy.gov/sites/default/files/oeprod/DocumentsandMedia/BlackoutFinalImplementationReport%282%29.pdf, accessed November 9, 2022.

[10] FERC (Federal Energy Regulatory Commission), NERC, and Regional Entities, 2021. *The February 2021 Cold Weather Outages in Texas and the South Central United States*, November. https://www.ferc.gov/media/february-2021-cold-weather-outages-texas-and-south-central-united-states-ferc-nerc-and, accessed November 9, 2022.

[11] Tang, L., and Ferris, M., 2015. *Collection of power flow models: Mathematical formulations*, University of Wisconsin: Madison, WI, USA.

09. EARTHSHOTS

The U.S. Department of Energy (DOE) has created Energy Earthshots™ initiatives to drive research activities needed to achieve its 2050 net-zero carbon goal [1]. As of September 2022, there are six Energy Earthshots: Hydrogen, Long Duration Storage, Carbon Negative, Enhanced Geothermal, Floating Offshore Wind, and Industrial Heat (Table 9-1). A common thread throughout the Energy Earthshots is that they require the development of novel complex engineering systems, comprising complex components ranging from electrolyzers to flow batteries to gas turbine engines to floating wind turbines. General capabilities to design and develop complex engineered systems across different domain application areas are therefore critical for success of each of DOE's Energy Earthshots.

New technologies present new challenges for established system engineering practices and design tools. When the complexity of a new system design exceeds capabilities of existing tools, developers need to fall back to excessive hardware testing, which leads to massive cost overruns and missed deadlines. Perhaps the best-known example is from the defense domain, where complex engineered systems are also common. The F-35 Joint Strike Fighter, a complex mobile weapons system, was delivered three years behind schedule and roughly $200 billion (nearly 100%) over budget [2]. Similar experiences (albeit at smaller scale) have occurred with virtually every new transformative technology development. The success of Energy Earthshots will depend on the availability of design and rapid prototyping tools that can handle designs of such complexity.

As global competition increases, other nations are improving system design capabilities. China, for example, commissioned its advanced Shadong aircraft carrier only six years after its construction began [3]. In comparison, the newest U.S. aircraft carrier, USS Gerald Ford, was commissioned 8 years after the start of construction [4]. This demonstrates that the Chinese military-industrial complex has made significant strides in closing the competitive gap with our nation in terms of building capability to develop extremely complex systems over the last decades. Capability to design complex engineering systems rapidly and at a low cost will be critical for meeting DOE carbon targets, as well as for the United States to maintain its leadership in new technology development.

Each Energy Earthshot has a well-defined cost target and deadline (Table 9-1), and each involves developing new or scaling up existing technologies. In order to meet Energy Earthshot objectives:

- new energy systems need to be designed and built within the specified Earthshot timelines;

PROJECT SPOTLIGHT

Project Name: ExaSGD: Stochastic grid dynamics at exascale

PI: Christopher Oehmen

Organizations Involved: Pacific Northwest National Laboratory, Oak Ridge National Laboratory, Lawrence Livermore National Laboratory, Argonne National Laboratory, National Renewable Energy Laboratory

Goal: Deliver capability to optimize transmission grid economic dispatch with respect to a large number of possible contingencies and different stochastic weather scenarios to enable grid planning with large number of renewable resources as a critical analysis capability needed for grid decarbonization.

Significant Accomplishment: Developed mathematical methods and implemented them in a software stack that performs economic dispatch analyses for transmission grid planning and operation at unprecedented scales (100,000s scenarios for a U.S. size grid), with our software stack also serving as a platform for deployment of different AI methods to further aid grid planning and operation.

In the News: Maintaining the Nations Power Grid by Exascale Computing, by Lawrence Bernard, 25 August 2022, https://www.exascaleproject.org/maintaining-the-

Table 9-1 Summary of DOE Energy Earthshots and their targets. Source: Energy Earthshots Initiative [1].

Earthshot	Cost Target	Target Subject	Timeframe
Hydrogen	$1	Production of 1kg of H_2	10 years
Long Duration Storage	Reduce cost by 90%	10+ hours energy storage	10 years
Carbon Negative	$100	Sequestration of 1 ton of CO_2	10 years
Enhanced Geothermal	$45	Production of 1 MWh of electricity	By 2035
Floating Offshore Wind	$45	Production of 1 MWh of electricity	By 2035
Industrial Heat	N/A	85% reduction of CO_2 emissions	By 2035

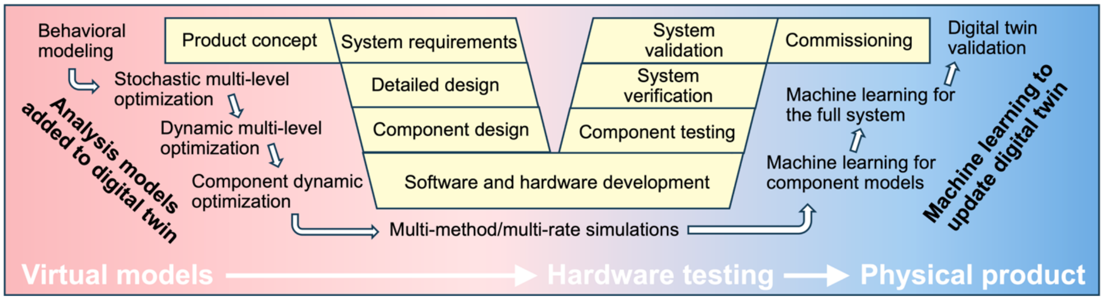

Figure 9-1 Typical system design workflow following the V-Model.

- new energy systems need to perform efficiently to help meet the Earthshot cost targets; and
- development costs of new energy systems and technologies need to be amortized during the system's lifecycle without exceeding the Earthshot cost targets.

Effective system design is therefore essential for success of all Energy Earthshots and artificial intelligence (AI) capabilities such as those detailed in Section 01 of this report that represent key enabling technologies.

The design process for complex energy systems, such as electrolyzers or flow batteries, typically follows a V-Model [5] (Figure 9-1). This approach begins with a concept design from which system requirements are derived, followed by design details of specific system controls and components. At each design stage, models of increasing fidelities and for different types of analyses must be produced. Once all system and component requirements are specified and verified in simulations, the system prototype is built up through several stages, starting from individual components and working all the way up to overall system integration. At each prototype building stage, physical parts of the system are tested through hardware-in-the-loop (HIL) simulations. In HIL simulations, physics-based models are connected with real-time data streams from the actual hardware. Massive amounts of high-quality data are generated during hardware testing stages. However, today, that data is typically used only for rudimentary model calibration and validation. There is a great opportunity to use this data for more advanced learning methods to develop, train, and deploy AI models to improve the system design workflow.

Because the vast majority of product development costs are in the hardware testing stages, reducing their number and duration is critical for meeting Energy Earthshots cost targets. This is especially true in later hardware testing stages (subsystem- and system-level), which are significantly more expensive than the early component-level testing.

The "digital twin" paradigm (Chapter 04) is closely related to the V-Model for system design. Originally the digital twin was meant to be a construct that captured/represented the performance and degradation of a component or system over its service life. Over time, the concept has been extended into the design realm where there is not yet a physical system. Although the term digital twin is not firmly defined [6], the most common use is to describe a hierarchical set of models that provides desired system representation at each stage of the product design as well as during the product development and operational lifecycle.

The digital twin is designed and constructed concurrently with the physical system prototype and is used at each design stage of the V-model (Table 9-1). In many instances, a digital twin is integrated within the final product (e.g., for automated controls, health monitoring, and fault prediction) [7]. A digital twin often implies a certain level of automation. It is not merely a collection of models but rather a virtual object that seamlessly provides the system representation at any desired fidelity level and for any analysis. Typically, a digital twin also includes learning and adaptation capabilities, updating its overall and component models based on test data during the HIL simulations or from the system sensor data collected as the system operates.

With increased accessibility and advances in AI capabilities such as surrogate (Chapter 01) and foundation (Chapter 02), and inverse design (Chapter 03) models, there are more opportunities to equip digital twins with advanced AI.

9.1 Open Opportunities

9.1.1 FIVE- TO 10-YEAR TIME FRAME

Several opportunities to take advantage of AI in complex system engineering can be leveraged in the short to medium term (five to 10 years). These opportunities present themselves at different levels, from model and system design to AI-human partnership, including the design of control systems embedded in the complex systems and their operation. Acting upon these opportunities will enable DOE to meet Energy Earthshots cost targets.

Models and Systems Design. Advances in AI, notably surrogate, foundation, and inverse design models (Chapters 01, 02, and 03), open significant new opportunities to fundamentally change how complex engineering systems are designed and to take advantage of the massive amounts of data generated in the design and operation process. Design data is of high quality because it is generated in a strictly controlled lab environment and is typically obtained in tests far from the operating point (stress tests) where a number of different modes are excited and nonlinear effects are dominant. With such a wealth of data, machine learning (ML) techniques can be devised to verify accuracy of physics-based component models within the digital twin and to automatically find corrections to those models when needed. AI models can be further used to automatically rerun a sequence of design computations with the updated digital twin to obtain corrected component and system requirements feeding back to improvements in hardware tests. This level of automation would represent a massive improvement over the current state of the art. Presently, re-running virtual system design stages requires significant manual intervention and is typically expensive and error prone.

Control Systems Software Design and Robustness. Embedded software systems must be developed, tested, and optimized alongside the physical systems to control their behavior. The same system design opportunities described above for the physical components and subsystems apply to control systems, along with capabilities such as AI-enabled software engineering (Chapter 06). Combined, these techniques have the potential to revolutionize the reliability and resilience of the complex systems central to the Energy Earthshots. Developing, testing, and continuously optimizing embedded software in control devices are as critical to these complex systems as the hardware and subsystem design processes described above. Engineered systems today have millions of lines of code embedded in their control devices. AI techniques including autonomous discovery (Chapter 05) and inverse design (Chapter 03), can be used to generate stress tests for software-in-the-loop simulations during the virtual testing stages, identify software bugs, and suggest fixes to programmers. During hardware/software system testing stages, AI models can be used to learn control response and communication latencies within and among subsystems. These are factors not captured in today's physics-based models.

Operation and Optimization of Complex Systems. AI techniques such as those enabling digital twins can also make operational systems more robust and resilient to disruptions. The same digital twin used for system design and test, deployed on a commissioned product, will provide on-board health diagnostic, prognostic, and supervisory control. Here, the digital twin can be a reduced-order model, e.g., obtained using AI/ML techniques guided by physical insights into the system. AI models deployed within these systems, such as on-board sensors and controllers that include "edge AI" (Chapter 15) hardware processing, can adapt to different usage patterns or operating environment conditions and modify control strategies. This will also allow for prototype digital twins that can be used for demonstrations and feasibility/cost studies, training of operator and maintenance crews, enabling one generation of complex system to "jump start" the next. Indeed, the AI capabilities described throughout this report, from surrogate models to digital twins to edge AI sensors and controllers, represent underlying technologies, methods, and systems that can be deployed in multiple Energy Earthshots. This will be important to realize economies of scale, and will be accelerated through the development of a set of tools that can be used by AI/ML engineers, ideally with domain expertise.

Data. Data collected from a fleet of commissioned products can be used for a variety of purposes. They can model uncertainties in the product's operating environment and update the control logic over the entire fleet accordingly. In turn, learning from this data can improve predictions in design computations and reveal modifications to make in future products. With years of accumulated fleet data, aging effects on the product performance could be modeled. Moreover, this growing, multi-modal data corpus holds promise for the development of foundation models (Chapter 02) that improve quality and reduce the development time and costs for new complex systems. To that end, AI methods in workflows (Chapter 13) and data management (Chapter 14) are required to support the capabilities described above and the collection, curation, and evaluation of data used for model training and optimization.

Human-AI Partnership. As designers, scientists, and system operators interact with AI design and control systems, digital twins, and similar capabilities, each human interaction provides data representing the opportunity for the AI systems to learn the interests and objectives behind human interventions in the system. Whether these are operational controls or design changes, AI systems can provide computational support (e.g., decision support, including suggested actions) to aid the human cognition during the process. For this cooperative learning loop to function optimally, both the human and the AI system need to "understand" each other. This will require advances in natural language processing, which is already a rapidly improving capability (Chapter 02), as well as in explainability. Many decisions relating to critical infrastructure require explainability and if the AI "box" remains closed to the human, the human may not reasonably trust the AI design suggestions that they receive. AI and robotics capabilities are already showing promise of automating laboratory workflows, including those integrating computational models (Chapter 05). These emerging capabilities will provide both insights and basic building blocks for AI and robotic capabilities interacting with system designers and operators.

9.1.2 10- TO 20-YEAR TIME FRAME

Automated Design. The long-term objective of AI for Energy Earthshots is to have fully automated system design processes, which will allow for rapid prototyping of new energy technologies and dramatically reduce product development costs across different domain areas. This will demand AI capabilities and advances, such as those outlined in Sections 01 and 03, to deliver a level of automation where domain experts provide a concept design and objectives, with all subsequent design stages (including requirements propagation) created by the AI system. Such an AI system should provide implementation options and associated evaluations and recommendations from which domain experts could choose. Once the best implementation candidate is selected, the AI system would orchestrate the hardware tests, collect data, and make design adjustments as needed. This process could be integrated with additive manufacturing so that component prototypes are 3D-printed on site (Chapter 05). This would also enable rapid hardware prototype adjustments based on the hardware testing results. Such fully automated design and prototyping process will strengthen and extend the nation's global leadership in new technology development.

9.2 Challenges to Overcome

Models and Systems Design. Data-driven and physics-based approaches typically have been investigated and applied in different contexts, but there has been relatively little crosscutting research across the two areas. That has hindered the adoption of data-driven methods in system design applications. Physics-based approaches have been and will likely remain in the foundation of system design, as they give predictions that can be interpreted in terms of domain science. Furthermore, physics-based methods provide a way to verify operational constraints, thereby minimizing security and safety design constraints with high levels of certainty. On the other hand, data-driven approaches are more effective in quantifying design variations, such as epistemic uncertainties or stochastic processes in the operating environment. Ideal modeling and analysis strategy for system design lies at the intersection of data-driven and physics-based approaches. Deployment of AI at the scale where it will deliver transformational changes to system design requires significant new research of novel "hybrid" methods that learn and make decisions based on acquired data. At the same time, these methods will need to strictly enforce laws of physics, security, and safety constraints. Advances in the development of AI-based surrogate models (Chapter 01), such as physics-informed and reduced-order models, will be essential to closing this gap.

Control Systems Software Design and Robustness. The state of the art for control system design is limited by numerical analysis methods used to model and simulate complex engineering systems, which in turn affects the ability to deploy more advanced methods, including AI, at scale. A typical complex system model consists of three parts: (i) differential-algebraic equations (DAEs) describing physical components, (ii) a finite state machine (FSM) describing control logic (which is implemented in the embedded software), and (iii) a Petri net model of communication between system components. There are significant numerical and computational challenges for scaling up each of these computations to the size required for the deployment of advanced AI models. Instead of scaling up the computation, a typical approach in industry today is to reduce the fidelity of systems' physical components models. By doing so, one is able to simulate a more complex system without significantly increasing the complexity of the model. This means that the simulation does not exceed capabilities of the existing tools. However, such models are often too coarse to take advantage of and incorporate fine resolution effects that can be captured by machine learning.

Operation and Optimization of Complex Systems. Complex systems relevant to Energy Earthshots have strict security and safety operational requirements. These requirements are challenging to enforce when using deep neural network (DNN) surrogate models, which appear as black boxes and whose behavior cannot always be interpreted in terms of physics. Using DNN surrogate models for components typically leads to high-dimensional system models with strong nonlinearities [8]. This makes system analyses (e.g., uncertainty quantification, adjoint sensitivity analysis, and constrained optimization) computationally challenging and beyond capabilities of standard system design tools and embedded devices used today. In model predictive control, for example, computing uncertainty propagation through DNN component models may be extremely challenging to perform within real time operation requirements.

Lack of physical intuition and computational complexity of DNN surrogates also raises questions about how to optimize systems, validate controls, and ensure that security and safety constraints are enforced. It is particularly challenging to understand limits of applicability for DNN models in system optimization. Are the models learned for one system configuration still valid after system parameters are optimized? Significant new research in physics-informed ML methods is needed to answer these questions.

Data. While hardware testing generates large amounts of data, many industries have limited capability to take full advantage of it due to lack of scalable data acquisition and management infrastructure. At present, most of the hardware testing data is analyzed directly by engineers. Therefore, the amount of data used is limited by how much data a human engineering team can process. The supporting infrastructure is built accordingly. To be able to deploy AI analysis, one needs to build an entire supporting ecosystem (also discussed throughout Sections 01 and 03 of this report).

This will bring new challenges, such as how to label acquired data for subsequent processing. There is a large number of configuration and environment parameters that specify a single hardware test. Furthermore, data entries with different labels are connected through laws of physics that need to be preserved throughout the analysis. There are also inherent aleatoric (irreducible) uncertainties in hardware testing processes that need to be quantified and factored in the system design properly.

Since system design involves multiple stakeholders, including multiple suppliers, there are proprietary and intellectual property issues associated with data and that need to be considered as well. For example, most component suppliers explicitly prohibit reverse engineering of their products. There is a risk that some learning methods deployed at hardware testing stages may be interpreted as a reverse engineering of system components provided by suppliers. There needs to be an organizational framework for complex systems design specifying how the intellectual property of each stakeholder will be protected and who has ownership and access to which data.

Human-AI Partnership. While large amounts of data are generated during hardware testing, a relatively small fraction of that data is collected and used today. The current bottleneck is the ability of system designers to process large quantities of data in a timely and cost-effective way. How to process large amounts of data to give engineers actionable information and help them navigate complex design spaces is still an open challenge. Here, the interaction among humans and AI systems is also critical, requiring research in human factors and in AI interaction mechanisms to interpret human input with consideration to context and intent.

Automated Design. Realizing the goal of automated designs requires that most, if not all, of the challenges presented (as they relate to shorter term opportunities) be addressed. There are also further challenges specific to automated designs. Currently, the system design process is fragmented, utilizing different and often incompatible design tools at different stages. The majority of these tools are proprietary, closed source, and have limited ability to interface with other tools. The lack of interoperability and limited data exchange capability with these existing tools poses serious challenges when deploying new methods and automating an established system design process. Developing AI approaches to improve and fully automate system design workflows will also require access to suppliers' databases and the ability to process and learn from historical data from various sources. Finally, the concept of fully automated complex system design can itself be posed as an AI problem with a massive number of parameters. What makes this problem particularly challenging is that couplings between components are extremely complex, with changes to one component potentially cascading through the entire system.

9.3 Investment Needed for Achievement

The main investment needed is in AI methods, frameworks, and models that can learn from hardware tests, interpret results in terms of physics, and update system design to meet (or exceed) Energy Earthshots targets. New methods also need to strictly enforce security and safety constraints.

To support Energy Earthshots, DOE also needs to invest in the development of generic components for system design relevant to each Energy Earthshot. Such investments will help standardize modeling practices across different modeling areas, especially since some components (e.g., power conversion or thermal management devices) are part of almost all energy systems relevant for the Energy Earthshots. Below, we describe some key investments that will create needed capabilities for multiple Energy Earthshots.

Models and Systems Design. Development of component models for system design computations typically makes up most of the modeling and software development costs during the system design. Component models are often tailored to specific numerical simulation schemes used in the design computations. While these models capture correct physics, their scope of application in terms of different analyses is narrow. In the context of digital twins, system components are not modeled by a single model but by a hierarchical set of models that capture the same physics, but which are adapted for different stages of product design. Modeling data for systems such as electrolyzers or flow batteries are neither easily accessible nor provided in a form suitable for mathematical analysis due to various proprietary and/or practical issues. DOE needs to invest in creating libraries of hierarchical generic component models for complex energy systems (with complete sets of their mathematical equations and modeling parameters available) to support and incentivize research related to Energy Earthshots. Investment by DOE's Advanced Research Projects Agency–Energy (ARPA-E) in creating generic transmission grid models [9] has spurred a flurry of research activities related to power grids. This success should be replicated for other energy systems as well in order for Energy Earthshots to be successful.

Furthermore, component models for digital twins need to support updates and modifications from different learning techniques. This poses nontrivial mathematical problems that have not been addressed completely thus far. There are also many challenges with data for creating surrogate models required by digital twins.

Control Systems Software Design and Robustness. To harness the power of AI for controls and embedded software design, there needs to be scalable modeling and simulation infrastructure, which can support multiscale hybrid models comprising both continuous and discrete dynamics components. Such a framework must allow dynamic analysis, obtain analytic derivatives for simulation and optimization, and enable code generation for real-time application with guaranteed solvability, execution time and memory footprint. This class of framework is needed to incorporate surrogate models for control systems obtained through ML from hardware testing data. Multimethod numerical integration frameworks [10] have been proven effective for multiscale problems cast in terms of ordinary and partial differential equations. However, the theory is not fully developed for DAEs, which are typically used to model complex engineering systems. Having hybrid simulations that capture continuous dynamics and discrete events and scale to large systems is still an open research topic and requires significant new investment.

Operation and Optimization of Complex Systems. Using AI to learn from observation data to optimize system performance and, at the same time, strictly enforce system security and safety operational constraints is of critical importance for the success of Energy Earthshots. Furthermore, in order for engineers to make sound design decisions, it is of utmost importance that AI analysis results are explainable. Delivering this capability requires significant new investment in physics-informed AI methods. Early results in this area combining physics-based modeling with data-driven learning are very encouraging [11, 12]. Preliminary numerical investigation shows that adding physical constraints can dramatically increase the data-driven model's accuracy in turbulent flows [13].

To address well-known limitations in data-driven modeling (e.g., sensitivity to noise in input data or lack of explainability), a targeted investment is needed to develop a symbiotic physics-data-driven modeling framework in which data is parsimoniously used to model only missing information in well-tested mathematical methodologies and improve their physical fidelity and numerical accuracy. This paradigm shift from "data-driven modeling" to "data-driven correction" is essential for efficient system design. It will allow for both reducing epistemic uncertainties in the digital twin by leveraging hardware testing data and for modeling aleatoric (irreducible) uncertainties accurately based on field operation data. This will provide basis for efficient uncertainty propagation models in the digital twin that can be used in real time for model predictive control during system operation. More importantly, accurately modeling deterministic and stochastic processes in the system enables engineers to strictly enforce security and safety constraints within the context of stochastic optimization.

Data. The precondition for deployment of AI at scale to system design processes is the development of appropriate data acquisition, management, and storage infrastructure (Chapter 14). This work requires additional, new research into optimal approaches and mechanisms to label data samples from inherently multi-modal and multi-scale sources, ranging from sensors to AI models to operational settings and outputs. Moreover, the development of AI models that can evaluate and analyze these data streams is critical for establishing (and discovering) proper correlations between them, given that each sample is associated with a large number of configuration parameters and environment sensor readings. In addition to data science research, significant new investment is needed to create an adequate software ecosystem, develop open-source middleware, standardize application programming interfaces, and specify data transfer protocols (Chapters 11 and 13).

Human-AI Partnership. Integrating AI into existing human-centric processes requires the development of new AI methods that will enable deeper interactions than those associated with a purely assistive role. The AI methods must understand the goals as well as the processes. Therefore, investment is needed in AI algorithms that embrace human incremental formalism. This will enable the human to absorb the AI into their cognitive process and allow for gradual construction of the product design stages, beginning with the initial concept design supplied by the human. The AI models should consult the human for expert feedback and the human should consult the AI for suggestions or assistance. Eventually, AI methods should be able to design simple processes with input from humans; then, humans would score the performance of the AI. Such an AI system would demonstrate the co-learning relationship between the human and the AI and gradually build more robust human-robot interactions. These human-AI partnership methods can ultimately help to achieve the concept of self-driving facilities.

Automated Design. Following the concept design, the subsequent design stages involve a lot of routine work, such as requirements propagation or model updates, that can be automated. Each subsequent design stage also generates large amounts of new data that need to be processed and fed back to prior design stages for design reevaluation. The decision-making process when moving from one design stage to the next is often influenced by the ability of human actors to process newly generated and often quite heterogeneous data, as well as ability to re-run prior design stages with the new data fed back in. Investment is needed in AI methods that automate this iterative system design workflow, starting from a concept design as the input and then automatically generating subsequent design stages while giving engineers several options to choose from at each stage. The automated design should be integrated with databases of different materials, components, and system designs to automatically identify best matches for the concept at hand and optimize its implementation for cost and performance. Further investment should be made in algorithms that can make more aggressive departures from previous designs in order to explore broader segments of the design space. The AI models also need to incorporate learning from hardware tests, interpret results in terms of physics, and update system design to meet (or exceed) Energy Earthshots targets. New methods must also ensure that security and safety constraints in each proposed design are satisfied and verifiable. Finally, investment is needed to support an effort to create AI models capable of orchestrating co-dependent activities in the design process and interfacing with additive manufacturing facilities to create component model prototypes on site.

9.4 References

[1] U.S. Department of Energy, 2021. *Energy Earthshots Initiative*, June. https://www.energy.gov/policy/energy-earthshots-initiative, accessed September 21, 2022.

[2] Choen, Z., 2015. The F-35: Is the world's most expensive weapons program worth it?, *CNN*, July 16.

[3] China Power Team. 2017. What do we know (so far) about China's second aircraft carrier?, China Power, Center for Strategic and International Studies, April 22, updated June 15, 2021. https://chinapower.csis.org/china-aircraft-carrier-type-001a/, accessed September 21, 2022.

[4] Kaufman, E. and Liebermann, O., 2022. US Navy's latest and most advanced aircraft carrier deploys for the first time. *CNN*, October 4.

[5] Clark, J.O., 2009. System of systems engineering and family of systems engineering from a standards, V-model, and dual-V model perspective. In: *2009 3rd Annual IEEE Systems Conference*, pp. 381–387.

[6] Barricelli, B.R., Casiraghi, E., and Fogli, D., 2019. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE Access*, 7, pp. 167653–167671.

[7] Nguyen, T., Ponciroli, R., Bruck, P., Esselman, T.C., Rigatti, J., and Vilim, R., 2022. A digital twin approach to system-level fault detection and diagnosis for improved equipment health monitoring. *Annals of Nuclear Energy*, 170, June.

[8] Zhang, X., Xie, F., Ji, T., Zhu, Z. and Zheng, Y., 2021. Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization. *Computer Methods in Applied Mechanics and Engineering*, 373, p.113485.

[9] Birchfield, A.B., Xu, T., Gegner, K.M., Shetye, K.S., and Overbye, T.J., 2017. Grid structural characteristics as validation criteria for synthetic networks. *IEEE Transactions on Power Systems*, 32(4), pp. 3258–3265.

[10] Sandu, A., and Gunther, M., 2015. A generalized-structure approach to additive Runge-Kutta methods. *SIAM Journal on Numerical Analysis*, 53(1): pp. 17–42.

[11] Mou, C., Koc, B., San, O., Rebholz, L.G., and Iliescu, T., 2021. Data-driven variational multiscale reduced order models. *Computer Methods in Applied Mechanics and Engineering*, 373:113470.

[12] Xie, X., Mohebujjaman, M., Rebholz, L.G., and Iliescu, T., 2018. Data-driven filtered reduced order modeling of fluid flows. *SIAM J. Sci. Comput.*, 40(3): B834–B857.

[13] Mou, C., Merzari, E., San, O., and Iliescu. T., 2022. A numerical investigation of the lengthscale in the mixing-length reduced order model of the turbulent channel flow. In: *Proceedings of 19th International Topical Meeting on Nuclear Reactor Thermal Hydraulics (NURETH-19)*, Brussels, Belgium, March 6–11.

10. NATIONAL NUCLEAR SECURITY ADMINISTRATION (NNSA)

The goal of bringing artificial intelligence (AI) systems into the National Nuclear Security Administration (NNSA) mission space is to dramatically reduce the time to execute across multiple mission programs, including stockpile stewardship, production and modernization, and nuclear nonproliferation.

Prior to this report and the 2022 DOE AI for Science, Energy, and Security workshops, there have been strategic planning meetings within the NNSA Office of Defense Programs (DP)'s Advanced Simulation and Computing (ASC) program, which provided a significant baseline informing this report. With the potential for AI and machine learning (ML) to create efficiencies in the nuclear deterrence (ND) lifecycle, the ASC program launched its tri-lab[1] Advanced Machine Learning (AML) initiative in FY 2019 with the objective of accelerating the ND design cycle and improving stockpile surveillance through advanced data analytics and by using AI/ML techniques.

This chapter presents five exemplar problems that are drawn from stockpile stewardship and nonproliferation mission spaces, with connections to some NNSA experimental facilities as well. These exemplars align with the goals outlined in the forthcoming ASC AI for Nuclear Deterrence (AI4ND) Strategy Plan, which will address AI technology needs for full weapon lifecycle - spanning discovery, design optimization, manufacturing and certification, and deployment and surveillance (DDMD) lifecycle phases, as well the detection, location, and characterization of proliferation activities.

Within this chapter, sections 10.2.1 through 10.2.4 provide grand challenges and goals for the role of AI within the ASC AI4ND strategy, and section 10.2.5 illustrates a grand challenge related to the nonproliferation work. Within the stockpile stewardship examples, there are multiple thrusts of the AI4ND strategy that seek to accelerate the time to deliver on lifecycle management. As *a most aggressive goal, AI systems could potentially help to support reducing the time to manufacture a first production unit (FPU) from more than a decade to a much shorter timeframe.*

10.1 Open Opportunities

Building upon AML, the ASC program aims to advance high-performance simulation capabilities with AI/ML-enabled tools to solve current and emerging national security challenges. Integration of AI/ML techniques offers the promise of: (1) bringing simulations in line with experimental reality; (2) gleaning insight from the vast troves of multimodal data across the NNSA mission space; (3) identifying rare or anomalous events; and (4) helping to identify, model, and characterize systematic uncertainty. Because of these opportunities and current NNSA investments in AML and exascale computing, NNSA DP is formulating a strategy that makes use of AI/ML across the entire nuclear weapons lifecycle. The ASC AI4ND strategy is an opportunity for NNSA to enhance scientific and technology leadership globally and execute dramatic and sweeping changes in the stewardship mission that aim to drastically reduce the time-to-solution across the full DDMD weapon lifecycle.

Enabled through the new AI approaches detailed in Section 01 of this report by the ASC AML initiative and by exascale computing, the strategy consists of new capabilities in the following lifecycle areas (Figure 10-1):

- **Discovery**: Discover new materials that are vital to national security priorities such as stockpile modernization. This effort would involve, for example, development of new

PROJECT SPOTLIGHT

Project Name: Machine learning of interatomic potentials with applications to materials aging

PI: Kipton Barros and Benjamin Nebgen

Organizations Involved: Los Alamos National Laboratory

Goal: Use an ensemble of neural networks to learn interatomic potentials from fine-scale simulations to accelerate larger simulations of shock and aging in mission-relevant materials, where the AI system continuously improves itself by testing its ability to make predictions in order to learn which new training simulations to run.

Significant Accomplishment: We have developed several interatomic potentials for bulk metals and have made significant progress on modeling, where our large-scale active learning framework runs effectively on the Sierra HPC system and uses GPU resources to perform DFT-based quantum calculations, perform ML-driven molecular dynamics simulations, and retrain the neural network potentials.

In the News: Our flagship publication that describes our Sierra workflow (an active learning framework that couples machine learning, quantum calculations, and molecular dynamics) appeared in *Nature*.

[1] The three participating laboratories are Lawrence Livermore National Laboratory, Los Alamos National Laboratory, and Sandia National Laboratories.

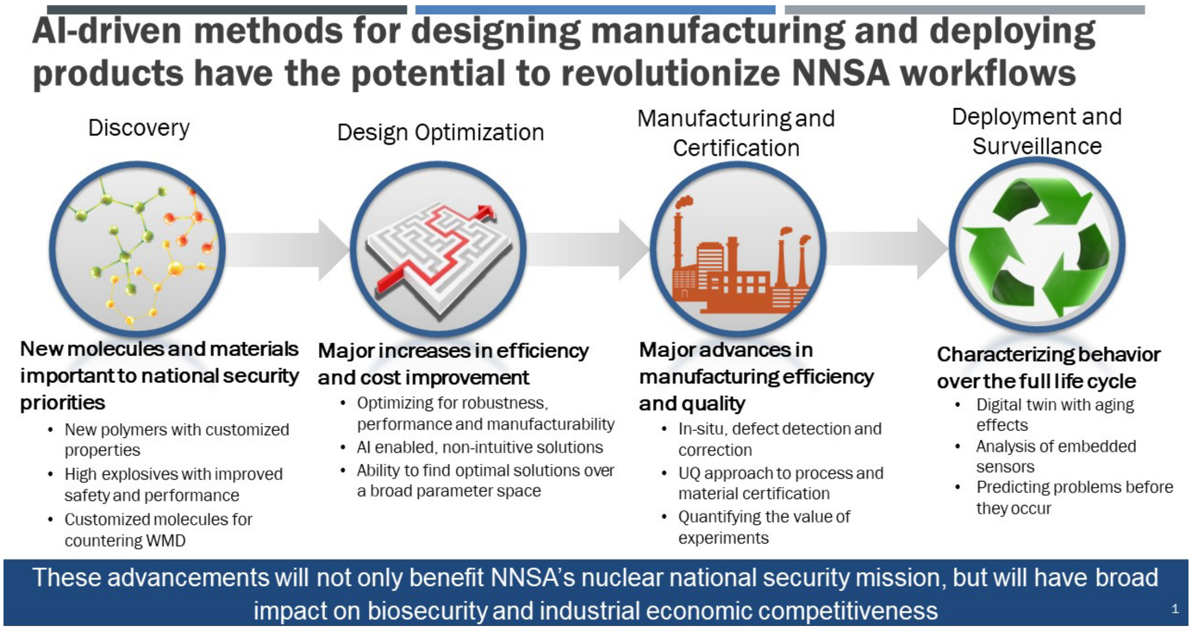

Figure 10-1. Future investments in research, development, test, and evaluation of AI/ML within the NNSA Advanced Simulation and Computing (ASC) program will enable significant improvements and enhancements of discovery, design exploration, manufacturing, and deployment (DDMD) processes.

polymers with designed physical properties, or high explosives with improved safety performance.

- **Design Exploration and Optimization**: Explore major efficiencies in a complex design parameter space and optimize weapons parts and system designs for requirements such as manufacturability, reliability, or cost efficiency.
- **Manufacturing and Certification**: Advance manufacturing efficiency and quality, comprising AI-enabled adaptive manufacturing controls, inspection, and qualification optimized in a tight loop with design and production.
- **Deployment and Surveillance**: Characterize behavior over the full weapons system lifecycle, including the use of digital twins (Chapter 04) with aging effects, analysis of data from embedded sensors, and awareness of potential problems before they occur.

10.2 Challenges to Overcome

This section describes five foundational research and proposed grand challenge problems whose solution will be required in the next 10 years to successfully harness the advantages of AI/ML to transform and accelerate the pace of discovery and development in high-consequence NNSA missions. The first four, in order, map to the DDMD lifecycle, and the fifth maps to the non-proliferation mission:

- 10.2.1 Scientific Discovery for Areas such as: Fission, Fusion, and High-Energy Physics
- 10.2.2 Design Exploration and Optimization using Multiscale and Multiphysics Simulations
- 10.2.3 Manufacturing and Certification of Parts and System Parts
- 10.2.4 Deployment and Surveillance for Stewardship Management and Global Security
- 10.2.5 Non-Proliferation

We note that Infrastructure grand challenges crosscut these grand challenges and are particularly prominent in the third grand challenge.

10.2.1 SCIENTIFIC DISCOVERY: FISSION, FUSION, AND HIGH-ENERGY PHYSICS

Grand Challenge: Develop an AI system that can identify new materials synthesis that couple unique NNSA requirements and enhance both performance and safety in extreme environments.

Introduction. High-energy density physics (HEDP) and fusion physics calculations are based on various multiphysics codes that include, but are not limited to, radiation-magnetohydrodynamics (radMHD) density functional theory (DFT) and molecular dynamics calculations. These are computationally expensive calculations that display low-dimensional emergent behavior. HEDP research is also associated with costly experimental modalities utilizing multiple diagnostic measurements that are designed to test and calibrate existing and novel physical models. To test and calibrate the model, researchers have a critical need for methods that can construct high-fidelity, efficient surrogate models (Chapter 01) of the physics, identify the low-dimensional sub-manifold structure of the modeled physics and the data, and finally assimilate the data with the model to refine and extend the estimate of the sub-manifold structure.

While good-quality surrogates for DFT calculations and ML-informed interatomic potentials are starting to emerge [4, 5], combining these with diagnostic data in near real time is a beyond-exascale challenge.

Opportunities. Solving this grand challenge problem will have a major impact on our understanding of uncertainty quantification (UQ) as well as validation and verification of HEDP, inertial confinement fusion, magneto-inertial fusion, magnetic confined fusion, and the factors affecting stockpile safety and readiness. This approach could also be applied to a broad range of other physical problems such as climate physics, geophysics, and astrophysics. In particular, the use of AI/ML methods for magneto-inertial fusion would enable new designs and reduce the risk of any proposed design not performing, both at current scale and at future scales. Such methods would also enable significant improvement in experimental design leading to greater understanding of HEDP physics (hypothesis test) and reduced risk of experiment failure. Success in this area could lead to commercial fusion energy and a more reliable stockpile.

There are numerous multiphysics codes of different fidelities that have been optimized to run on the exascale computational platforms, so that ensembles of many simulations (100s to 100,000s, depending on the fidelity) can be produced, generating sufficient training data to create AI surrogates (Chapter 01), inverse design (Chapter 03), and control system (Chapter 04) models necessary to support new experiment design and optimization opportunities. Additionally, large databases of experiments are available from experimental facilities with multiple high-quality diagnostic measurements for each experiment, spanning a broad range of physical regimes. Physics-informed ML methods that can be trained on different materials with generalization capabilities for different temperatures also show promise [1]. Finally, this grand challenge could leverage tools and capabilities developed for the domain areas within the Office of Science (Chapter 07).

Risks. Without realizing the improvements from AI-based methods, the fidelity that is required for HEDP use cases remains beyond the reach even of exascale-class HPC systems. This gap means that the pace of our science discovery will not match that of other actors, affecting our national security and scientific competitiveness. Moreover, the continued aging of the stockpile will increase the demand for modeling and simulation, which is challenging even with exascale systems. The number of different materials, as well as the different physics and scenarios that need to be studied, represent a grand challenge that is beyond the reach of current conventional methods due to limitations in compute and the scaling of some first-principles methods. Without AI-enabled approaches, we run the risk of critical gaps in our understanding all of the physical properties, at all scales of interest, for all materials.

Advances in this area would enable us to use current and future exascale systems to solve hundreds of problems by harnessing the speedups of surrogates, property inference, etc.

Challenges. An AI grand challenge problem is to focus on near-real-time workflows that enable discovery of new materials vital to national security priorities (Section 03: Technological Crosscuts discusses workflows, software frameworks, data infrastructure, and other factors). The AI-assisted workflow will use data generated from sensors, images from cameras, and other diagnostic sources to enable edge analytics near the accelerators/experimental facilities or in the field (e.g., detecting radiological sources in urban areas or major ports of entry). This workflow will also enable inference using surrogate models at device scale or online learning approaches deployed in computational resources physically near accelerators or other experimental facilities. These new AI/ML-enabled workflows would not only improve an individual experiment but would evaluate results for the purposes of designing the next set of experiments and for retraining the surrogate models on capability-class computational systems using carefully chosen diagnostic data and generated configurations from the experimental data. Ensembles of 100s to 100,000s of multiphysics simulation runs would be performed based on AI-specified configurations to generate new training data, which can be combined with experimental data, such as from the Z-machine, the National Ignition Facility, DIII-D, and the Tokamak Fusion Test Reactor, to train surrogate models offline.

10.2.2 DESIGN EXPLORATION AND OPTIMIZATION USING MULTISCALE AND MULTIPHYSICS SIMULATIONS

Grand Challenge: Develop a master model—i.e., a foundation model (Chapter 02) specifically trained for a range of related downstream tasks—for material design or multiscale physics to enable weapons designs that are optimized for performance, ease of manufacturing, short qualification times, and/or specialized mission needs.

Introduction. Enabled by increases in available computing power and driven by rapid developments in applied mathematics and computer science, the ASC program has demonstrated positive impact across many areas of the NNSA. Many aspects of the NNSA mission, including weapons design, production modernization, and qualification and certification, rely heavily on our ability to simulate everything from fundamental physics and material response under a wide range of physics regimes to full-system performance calculations for complex engineered systems. Simulation can significantly accelerate the design cycle, limit the need for costly or prohibited experiments, and are key to stockpile assessment. However, because the computational requirements of first-principles modeling approaches exceed available HPC resources—even in the exascale regime—for all but the most fundamental sciences, computational models grapple with a trade-off between accuracy and performance. That is, scientists must choose the least approximate solution that is feasible to compute with the available resources. Fully resolving many important problems remains out of reach due to either lack of computational resources, lack of physical models, or lack of sufficient data to parameterize more accurate models. Many of the core phenomena of interest in weapons science span many scales in space and time, and often entire subfields are dedicated to understanding and approximating just a single scale. The corresponding simulations represent all smaller scales in the aggregate while all larger ones are effectively ignored. Consequently, the utility of such models is limited to exploring very specific questions and always carries the risk that some unresolved effects at scales below the resolution of the relevant simulation may lead to significant errors in the answers.

One common approach is multiscale models that couple simulations at different scales; fine-scale, expensive models are restricted to the most important parts of a problem and other models cover larger scales with more approximate solutions. This type of coupling also extends to different types of physics, that is, connecting hydrodynamics with radiation transport, for example. Many of the most impactful simulations are assembled as a collection of different physics models at different scales and are carefully chosen to provide the most accurate overall solution given the available computing resources. Nevertheless, even the most sophisticated multiscale, multiphysics simulations remain many decades away from explicitly resolving all known physics effects, even assuming an unabated increase in computational power. AI-based techniques such as AI surrogates, foundation models, and property inference (as outlined in Section 01 of this report) have the potential to fundamentally alter this trajectory, leading to unprecedented capabilities in the next five years and a radical restructuring of computational science in general within the next decade. Moreover, this initiative could leverage models and capabilities developed in the broader context of energy science (Chapter 08) or as part of the effort to address the U.S. Department of Energy's (DOE) Energy Earthshots initiative (Chapter 09).

Opportunities. As discussed in Section 01, recent advances in AI/ML have given rise to scalable and efficient AI-based surrogate models that—once calibrated from sufficient training data—can replace a broad range of physics modules with surrogates that accelerate the computation by factors of 1000s and beyond. Consequently, given an existing assembly of multiscale and/or multiphysics components, one can iteratively replace the most computationally expensive parts with AI surrogates, leading to unprecedented speedups. A master model could be developed that addresses the needs for multiple material design needs under different conditions. This capability would allow for the composition of a hybrid system based on AI using a master model (see Chapter 02) and multiphysics calculations, creating a simulation that is truly greater than the sum of its individual parts. This hybrid system will enable the design and development of true scale-bridging simulations in which even the largest scales are informed not only by bulk physics at the respective scale but via trained models that directly incorporate information from all finer scales.

An AI-empowered multiscale/multiphysics framework in its fully developed form will enable an autonomous approach to accelerate any existing simulation capability in a transparent and easy-to-adopt manner. Given an existing modeling system, the new framework will target the most expensive components of the system and replace them with AI surrogates. Subsequently, a new decision point is introduced that, at each invocation of the submodule in question, uses UQ techniques to determine the trustworthiness of the trained model. If the inference requirements are deemed to be within acceptable limits based on the uncertainty calculations, an accelerated AI-surrogate for that model is used in place of the traditional (and more computationally expensive) component. Whenever challenging data ("out of domain" or high uncertainty data) are encountered, the system reverts to the original physics module instead of the AI surrogate. This challenging data augments the training dataset to iteratively improve the model. This approach could be applied continuously and recursively at all scales. Ultimately, DOE will be able to assemble a master model for key constituent physics modules that collectively enable simulation at unprecedented speeds with ultrafast, trusted AI models and thereby replace traditional strategic computing components. Consequently, AI-driven multiscale simulations will enable design, exploration, and optimization using massive simulation ensembles at exceptional fidelities with the potential to drastically accelerate the entire DDMD lifecycle.

Risks. Much of the nation's success in, for example, stockpile stewardship and the corresponding technological advantages, has relied on superior simulation capabilities that both substitute for extensive nuclear tests and enable rapid design. AI-based surrogates have already been demonstrated in key application areas such as radiation-hardened microelectronics design and fabrication, HEDP, additive manufacturing, and high-energy materials. Consequently, it is virtually certain that capabilities like those described above are being developed by other actors, eroding the advantages of superior simulation capabilities. Realizing this grand challenge will cause a disruptive advance in simulation capabilities for whoever achieves operational status first. In contrast to the current state of the art, in which high-fidelity simulations can require months to complete and yet disagree with experiments in critical details, these new AI/ML approaches will support massive parameter sweeps of highly predictive simulations with enormous design potential (Chapter 03). Such capabilities have potential to leap-frog decades of prior advantages, creating tremendous security and industrial benefits. It is, therefore, imperative for DOE to secure its leadership in this field, both to boost the nation's competitiveness and to adequately judge the capabilities of other actors. Simply maintaining the status quo is a significant risk.

A critical risk associated with surrogate-based frameworks is rooted in the fact that they represent a fundamentally different technology than existing simulations, and thus past performance may not predict future success. That is, significant advances toward any of the remaining technical challenges discussed below might come from otherwise unrelated research such as in computer vision, natural language processing, or any number of other application areas employing AI/ML. This orthogonal nature of AI advances in one area allowing profound impact in entirely different science domains opens the possibility for adversaries to potentially assemble a working system without a large lead time and, with the exponential increase in predictive capability provided by AI, quickly erase prior deficits in physics capabilities and could quickly gain an advantage.

Challenges. While AI-based surrogates for some critical applications have been demonstrated, a full master model and an AI system working within or in a composable fashion with multiphysics simulations as described above will require several fundamental advances. Here, we focus on the specific technology needs for AI-enabled multiscale/multiphysics modeling identified during the workshops organized according to the technology crosscuts in Section 03 of this report.

The first set of needs relates to the ***underlying theory of machine learning***. To achieve a master model for material design implies the ability for reliable UQ to answer such questions as, "Is the current model trustworthy or does it require retraining?" Additionally, in order to build confidence, verify outputs, and explain unexpected results, the full master model will need to meet the requirement that researchers can interpret any of the models being deployed as well as their complex interactions. Finally, the efficiency and effectiveness of the approach can be significantly improved by integrating active learning algorithms that proactively improve models instead of waiting for answers to be deemed unreliable.

The second set of needs address the ***changing nature of the overall software and system design*** and combines challenges in software, workflows, and data management. The AI system as outlined above implies a shift from the complex, modular applications used today to a more flexible, dynamic, and unpredictable mixture of simulations, model inferences, and training. We will need new software frameworks that can seamlessly integrate into the current computational ecosystem. Additionally, the composable use of traditional components and AI models recast otherwise monolithic applications as complex workflows that manage a variety of different components.

Another important consequence of deeply integrating AI-based surrogates is the need to manage the training data, models, and their provenance as necessary to ensure accountability and repeatability. There will exist a set of persistent and ever-evolving master models that represent significant investments and capabilities akin to current experimental databases. Maintaining a detailed record of what data was used to build such models, which fidelity was used, and which algorithms were used for training will require **sophisticated data management across DOE sites and programs**. This critical concept is detailed in Chapters 14 and 19.

The final technical challenge will come from the ***changing need for computational hardware*** as the training of massive master models asynchronously, fast inference, and fast asynchronous training might become substantial bottlenecks. Furthermore, some theoretical advances, such as UQ or automatic differentiation, will benefit from and, in some cases, require new hardware developments (Chapter 15).

10.2.3 MANUFACTURING AND CERTIFICATION

Grand Challenge: Significantly reduce the time required to field new weapons systems with adaptive manufacturing and automated qualification and testing.

Introduction. AI-enabled autonomous control for additive and advanced manufacturing would be a revolutionary capability for the DOE national laboratories and U.S. manufacturing industries. It would accelerate the design, build, and test phases of large-scale DOE science experiments (e.g., National Ignition Facility, Z-machine, Fermi National Accelerator, Advanced Photon Source). Simultaneously, production capabilities for the NNSA nuclear stockpile program would be accelerated, enabling fundamental national security objectives. Traditional custom design, fabrication, testing, and qualification of components and integration in systems often take a decade or longer. AI-enabled digital engineering holds the promise of reducing these production lifecycle times by one-half or more through greater use of virtual design/simulate cycles on HPC systems, identifying the most promising candidates to reduce the number of build/test cycles, which are costly and time consuming [3].

Challenges. We describe here a grand challenge problem that demonstrates key benefits of applying AI capabilities, including autonomy and robotics (Chapter 05), to advanced manufacturing from the early conceptual stages through deployment in certified systems. Achieving this transformation will significantly accelerate facility or system deployment, enabling associated programs to compete with agility in an environment that is rapidly evolving technologically. This grand challenge has four components.

The first component is the ***development of AI-enhanced manufacturing technologies*** where ML techniques enable unprecedented improvement in the timescales required for developing parts and components. ML would be used to create fast surrogate models from high-fidelity physics simulations (Chapter 01, and previously in section 10.2.1). These fast surrogate models would be incorporated into the manufacturing process monitoring and control system. Multimodal data observations of Non-Destructive Test and Evaluation (NDTE) sensors would produce data to continuously train ML models that would be used to monitor manufacturing of components and to certify that they already meet all qualification requirements—without the need for further time-consuming inspection (i.e., the components are said to be "**born qualified**").

Manufacturing processes would then be scaled up using AI techniques through ***data-driven "digital twins"*** (Chapter 04) for manufacturing entire components, assemblies, and ultimately the manufacturing facilities themselves. This effort could also leverage new AI programs in energy and advanced manufacturing initiatives (Chapter 08). ML models generated from data collected during the manufacturing processes will be used to understand and optimize performance, as well as to train surrogate models or generate new configurations for training data generation for surrogate models. An ML framework could optimize a specific design for functionality, performance, or a consistent and reliable manufacturing yield, or any combinations of these.

Third, manufacturing technologies in this grand challenge would take into account NNSA needs to ***optimize components containing hazardous materials***. ML would be used to create surrogate models from high-fidelity physics simulations of the materials and the manufacturing processes. These surrogate models would enable broad exploration of the design space for chemical, radiological, mechanical, thermal, and constitutive properties. ML methods would also be used to fuse these surrogate models with the limited experimental data from facilities such as the Z-machine, Lawrence Livermore National Laboratory's National Ignition Facility, and the Los Alamos Neutron Science Center (LANSC).

These new AI/ML technologies must span the complete range of NNSA manufacturing needs. The fourth component of this grand challenge problem is ***AI-enabled manufacturing and co-design***. NNSA has unique manufacturing facilities that differ from industry and, as such, require specialized development and application of AI techniques. For instance, the NNSA has the only remaining trusted microelectronics fab for producing the NNSA's strategically rad-hardened microelectronics, and this facility is used to create custom integrated circuits (ICs) for nuclear deterrence electrical systems (NDESs). A challenge for domains such as AI-enhanced microelectronic co-design is the coordination with the highly developed electronic design automation (EDA) industry. This ~$30 billion/year industry is also deploying AI/ML within its tools, though its focus is not necessarily on the trusted strategically rad-hard (TSRH) microelectronics critical to the NNSA mission. NNSA-critical microelectronic products are currently designed using commercial tools supplemented by custom NNSA multiphysics codes. Supporting and complementing the industry ecosystem progress, while furthering our unique needs, will necessitate deep scientific understanding of the foundations and vulnerabilities of this AI-enhanced approach as well as continued coordination with, and evaluation of, commercial EDA software.

Semiconductor design and manufacturing is perhaps the penultimate example of process optimization: fabrication of CMOS chips with 100 million to more than 10 billion transistors of nanometer dimensions typically requires more than 700 individual process steps (lithography, pattern transfer etching, thin film deposition, planarization, cleaning, etc.). Each of these steps uses multiple $10 million tools guided via advanced metrology and statistical process monitoring and control. A modern fabrication may accumulate more than one terabyte (TB) of process data associated with a single wafer lot progressing through the full CMOS process flow (700+ steps), where this data may track minute variations of metal line widths, etch depths and roughness, film thicknesses and planarity, nanometer particulates and lithographic blurring. The ultra-high-volume throughputs (100–150 wafers/hour through each step) in modern fabrications can drive a rapid descent of the experience curve for most new products. This drives up wafer yields (fewer defects/errors) and enables the shipment of hundreds of millions of parts annually, as required to justify facility costs that are often in excess of $20 billion. The challenge here for NNSA is that the volumes required for its unique TSRH chips are ~100,000x smaller. Consequently, the descent of the experience curve is much slower and prone to setbacks due to manufacturing yield and qualification performance variability, leading to a much slower product realization than for commercial consumer chips. The solution here is to develop and employ AI/ML techniques that exploit the TBs of design and manufacturing data gathered during even low-volume fabrications to "virtualize" the rapid learning cycles otherwise achieved in ultra high-volume consumer device manufacturing. This will provide game-changing benefits to NNSA and other low-volume national security microelectronics customers (e.g., U.S. Department of Defense), with additional potential benefit to small U.S. companies in the microelectronics industry that have not yet achieved high volume deliveries.

Microelectronic design relies on models at multiple length and time scales to capture not only the theoretical performance of a given circuit design, but also the impact of minute variations in the fabrication of the transistors, wiring, and power delivery, especially in the presence of extreme environments (e.g., temperature, radiation, and high voltage). ML may be used to create compact device models at multiple levels of fidelity, which can be used to accelerate the co-design of microelectronic components in NDESs. AI methods may also be used to model and alter integrated circuit fabrication parameters to meet evolving design specifications, including the extreme environment performance not covered in commercial electronics. Design and fabrication of new, cognitively aware, and cyber-secure microelectronic devices could be enabled using ML techniques that evolve microelectronic design with anticipated hostile environments. Similarly, AI methods may be used to enhance acceptance criteria, the inspection process, and material use at PF-4 or additive manufacturing tooling at the production agencies. In-situ monitoring with AI-aided analysis is expected to enable detection of anomalous builds in real time and aid in non-destructive testing and evaluation for increased certainty in as-built parts. This analysis phase will then be used to inform and aid designers in designs that are easier to manufacture (increased acceptance rates) while still meeting demanding constraints.

Risks. The risks associated with not investing in this area are two-fold. First, the U.S. could not keep up with emerging threats to national security. Second, the time and cost to design and build our nuclear stockpile could become unsustainable. Without AI-driven manufacturing and certification, each step in the manufacturing process is a costly and time-consuming near-custom job. Neither of these risks is acceptable. Therefore, we believe that the U.S. must embrace and build upon the AI/ML capabilities being developed not only at DOE national laboratories but also in universities and throughout U.S. industry. The national laboratories will provide mathematical rigor, verification and validation, and UQ to AI/ML techniques; commercial AI applications and systems have more relaxed, or no, requirements in these areas. This will enable these techniques to be applied to higher-consequence applications. Working together, the national laboratories, universities, and U.S. industry will advance the state-of-the-art in AI/ML to improve small-lot manufacturing capabilities within the U.S., which in turn will reduce reliance on foreign manufacturing and improve national security.

10.2.4 DEPLOYMENT AND SURVEILLANCE

Grand Challenge: Develop a digital twin for every deployed system in the stockpile to assess health and aging under field conditions.

Introduction. The NNSA laboratories annually assess the safety and performance of the nuclear weapons stockpile and report the stockpile assessment to the President of the United States in an annual assessment report (AAR). The collection of new surveillance data is often limited by the availability of funding, support, and hardware for testing. In addition, the AAR is conducted on testers typically different from the testers that were used for original product acceptance due to rebuilds and upgrades driven by obsolescence or other requirements changes. New AI methods are needed to evaluate and predict component and system performance in the face of these challenges. Such methods may include:

- The development of new ways to apply advanced data analytics to existing data and/or generate synthetic data for minority classes of defects having insufficient, naturally occurring data for predictive analysis (rare defects).
- The ability to classify defects or signs of aging using data from measurements including microstructures, CT scans, images, and other available measurement modalities and the development of methods to verify and validate datasets, tester performance, and modeling validity.
- The development of the ability to forecast and predict manufacturing defects from incomplete production data (causal models).
- The creation of a more open data environment and analytics environment for widespread adoption of data science.

The NNSA surveillance mission would greatly benefit from new measurement capabilities and methods of manufacturing process control using physics-informed advanced data analytics and ML to support existing manufacturing capabilities and anticipated life extension programs (LEPs).

Opportunities. Although surveillance data collected annually may be sparse for some components, sufficient surveillance data have been collected over time to enable the present-day application of AI data analytics. In addition, high-fidelity, physics-based models have achieved a level of maturity sufficient to generate synthetic data that may be needed to predict rare defects and to reach critical mass with respect to sufficient data to train large models using HPC systems. Tools for advanced data analytics have also matured and are widely available to mine existing surveillance data and develop new capabilities for ensuring confidence in data quality. Finally, there is a sufficiently large and skilled workforce available to execute the data science for surveillance mission. These factors provide a starting point.

Risks. Adoption of more advanced, AI/ML-enabled data science methods such as those described in Section 01 to support the DOE's surveillance mission would lead to faster data-driven decision-making, repeatable and reliable decision-making with archived pedigree, and a reduction in the time and resources needed for stockpile evaluation (i.e., in the AAR). At the same time, operational systems require a level of confidence that underscores the requirements outlined throughout Section 01 and in mathematics and foundations (Chapter 12). If new AI methods are not explored and implemented, the surveillance of the nuclear stockpile could become prohibitively expensive and unsustainable.

Challenges. Many HPC codes and existing AI methods are available to make rapid progress on this problem. Existing and in-development ASC models of various components (Aleph, Aria, etc.), data analytical software (TensorFlow, Unscrambler X, MATLAB, R, Python, etc.), computing hardware (high-performance data analytics platforms with graphics processing unit [GPU] acceleration and distributed file systems), and experimental apparatuses are available on the restricted or classified network to verify and validate the component, system, or manufacturing processes and products. Beyond current systems, there is also a rapidly growing need for physics models for aging processes, ideally harnessing scientific progress within NNSA and in the Office of Science (Chapter 07).

The biggest challenge is related to performing big data analysis on multivariate sparse data and performing a causal analysis that links signs of defects and aging to the root cause. Natural language processing techniques, including foundation models (Chapter 02), could be used to automatically scan through hundreds of reports of significant findings to make fine correlations between symptoms and possible causes.

Another challenge is to create an AI-powered resilient knowledge ecosystem (RKE), as knowledge management continues to be a significant issue. We discuss this at length in Data Ecosystem (Chapter 14), based on requirements related to assembly, curation, evaluation, and encoding of training data (Section 01). The human ability to generate tremendous amounts of information is rivaled only by the complementary limit on any human's ability to digest that information, exponentially increasing the problem of not knowing what information is available and relevant to a given task at hand. This confluence requires that we *intentionally* manage our knowledge, data, and analytics. Through the use of various ML techniques (e.g., natural language processing, text analytics, various forms of ML) and other AI approaches (e.g., induction, reasoning by analogy), the RKE will enable easy preservation, curation, and dissemination of critical artisanal knowledge as our workforce, workflows, and work products continue to evolve. The DOE workforce (e.g., scientists, engineers, managers) will not only have access to but will rely on the RKE for recommendations, knowledge and resources (at whatever classification level) appropriate to the tasks they are performing as they perform them. Likewise, staff knowledge and decision provenance will be captured by the RKE as a function of the staff interacting inside the RKE without adding extra burdens to the NNSA workforce.

Last, we need to instrument deployed systems and their environments with sensors to vastly increase the amount of data collected for surveillance. This need not be standard surveillance data but can include a host of new measurement modalities that can be used to train and continuously optimize digital twins and simulations (Chapter 04) to predict aging and other effects. An added challenge addressable with the tools of AI is sensor signal discernment/inference that minimizes and quantifies anomalous information. Part of this capability will include the use of AI capabilities within edge computing devices to process observational data in situ for rapid onsite assessment.

The last grand challenge with the related theme is to develop digital twins of aging devices. Components of (critical) complex engineered systems often fail due to aging, as material microstructures evolve, material properties change, and material response to thermal, mechanical and radiation stimuli deviate from design specifications. Naively, it should be possible to design classifiers that detect outliers or anomalous behaviors via continuous monitoring and/or non-destructive testing. However, many of these complex systems are few in numbers and examples of failed behavior are few. Consequently, empirically collected datasets are insufficient to serve as training datasets for classifiers. It should be possible, in principle, to construct and train classifiers on synthetic datasets, assembled out of simulations of devices with aged material models (with further "tuning," e.g., via transfer learning, using scarce empirical measurements), but such material models (also known as "subgrid models," constitutive models, or closures) and digital twins of aging devices do not yet exist.

There are high-fidelity models, such as crystal plasticity models, that can be used to construct the training datasets for (aged) materials. However, challenges lie in the architecture of the material models, their "tuning" with multimodal data (images of microstructure, spatiotemporal measurements of macroscale responses to stimuli, field measurements of stresses and strains from load tests), and the incorporation of uncertainties in the trained models and their qualification (they are data-driven and can suffer from out-of-distribution errors). In addition, these data-driven closures must satisfy physical constraints (e.g., Galilean and rotational invariance). In addition, the incorporation of these new closures into device models (i.e., the digital twin of the aged device) may introduce numerical issues (e.g., stiffness) in current models. This use-case poses some of the requirements for surrogate models (Chapter 01) as closures and devices' digital twins will rely on surrogates, as well as material property estimation (Chapter 03).

If successful, these closures for aged materials can be used to develop training datasets for aging classifiers. A digital twin of the aged device could in turn be used to predict device lifetimes, which are fundamental for predictive maintenance.

10.2.5 NONPROLIFERATION

Grand Challenge: Develop an AI system to rapidly detect, locate, and characterize foreign activities related to fuel cycle and weapons development, movement of nuclear materials, and nuclear explosions across the globe.

Introduction. The NNSA Office of Defense Nuclear Nonproliferation (DNN) works to prevent state and non-state actors from developing nuclear weapons or acquiring weapons-usable nuclear or radiological materials, equipment, technology, and expertise [6]. The long-term effectiveness of U.S. methodology is confounded by advances in nuclear technology and adversary efforts to hide illicit activity, making it feasible for a nation to produce significant quantities of special nuclear materials, specialized explosives, rad-hard electronics, and other critical technologies with a minimal facility and personnel footprint. Despite the unparalleled amount of data being collected by ever-increasing and evolving sensing capabilities, it is doubtful that we will be able to collect significantly more actionable data than we have now, especially against a sophisticated low-profile proliferator. This situation highlights the need for even more sophisticated means of sifting and correlating the flood of data to extract the unique signatures associated with nuclear proliferation activities.

Data analytics and signature extraction processing needs include:

- Patterns of life: processing of open data (social media, industry supply chain data, scientific publications) [7].
- Centralized/Ground station data processing of sensor data.
- Distributed sensor ("edge") onboard processing (satellites, terrestrial, seismic).

AI techniques such as those described in Section 01 offer dramatic improvements in signature extraction in all of these areas independently, and perhaps even more impact by correlating across all three datasets.

The DNN R&D program advances the nonproliferation mission through leveraging investments in unique subject matter expertise and facility testbeds. These testbeds represent critical pieces in the nuclear fuel cycle and are ideal targets for exquisite remote sensing data collection to help in research and development efforts to detect and monitor foreign nuclear fuel cycle and weapons development activities, special nuclear material movement or diversion, and nuclear explosions.

Over the last decade, DNN R&D has made significant investment across multiple AI-enhanced programs with the goal of accelerating analysis timelines to detect, localize and characterize foreign nuclear proliferation activity. These same capabilities support nuclear arms control treaty monitoring and verification, operational interdiction and other nuclear security efforts across NNSA and government.

Key programs include:

- Multi-Informatics for Nuclear Operations Scenarios (MINOS): use of diverse physical measurements for high-fidelity detection, location, and characterization of proliferation activities.
- Advanced Data Analytics for Proliferation Detection (ADAPD): combination of data and physics models to enable early detection of low-profile proliferation.
- Persistent Dynamics: real-time optimization of proliferation detection.
- Steel Thread: use of foundation models to address proliferation challenges.
- Low Yield Nuclear Monitoring (LYNM): use of multiple sensing phenomenologies to increase detection sensitivity.

Ranging from large, coordinated multi-modal data collects at testbeds to robust multi-modal data analysis to establish patterns of life for event prediction to building large-scale foundation models for unique sensing approaches and phenomenologies, these AI-enhanced methods offer the promise of enabling nuclear proliferation analysts to perform deeper, more timely, and more comprehensive assessment of a foreign state's nuclear enterprise. Furthermore, these investments aim to enhance the teamwork effectiveness between nuclear proliferation analysts and AI systems to produce next-generation AI-augmented experts for global nuclear assessment.

Opportunities. To date, much of the DNN mission space has relied on subject matter experts and trained analysts to comb through ever-larger troves of data, searching for key "tells" that an adversary is working toward nuclear proliferation. As we shift from monitoring known large nuclear-capable states to include global detection of small-scale nuclear proliferation activities, this approach will be increasingly unsustainable. Enhanced sensing capabilities (direct) and other closed-source and open-source (indirect) data (Figure 10-2), such as publications, bills of lading, and social media, produce exquisite data at larger volume and velocity (and with more diversity) than humans can ingest. Concurrently, nuclear technologies have advanced to a degree that makes it easier for potential proliferators to hide their activity even among this deluge of data.

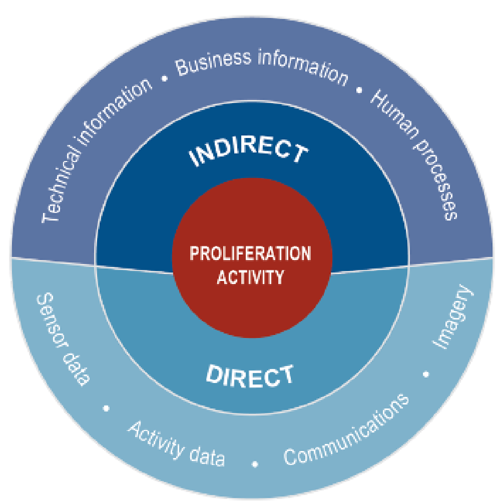

Figure 10-2. Indirect and direct sources of proliferation-related data.

The breakthrough insights that the AI community has developed in pattern matching, anomaly detection, natural language interfaces to query-answer tools, and the ability for sophisticated AI foundation models (Chapter 02) to synthesize cogent responses are unprecedented. Commercial applications—with limited application to DOE mission requirements, much less non-proliferation needs—often have access to millions of events of interest, and the AI methods they use rely on that volume to ensure reliable performance. By contrast, in nonproliferation there are typically very few events of interest even within the enormous amounts of data collected against those few events. Thus, to achieve parity with industry breakthroughs in AI, novel methods must be developed that combine the knowledge of highly skilled NNSA subject matter experts with sparse data across three axes: sparse observables, volume and variety of training data, and missing data modalities for inference.

There is little opportunity (or incentive) for the AI industry to invest heavily in the domain-specific adaptation that will be necessary to make these techniques successful on our mission data. However, DNN has invested in larger and larger measurement campaigns, presenting an opportunity for development of rich collections of data for algorithm development. In addition, DNN has invested in academic collaborations through multiple university consortia, presenting an opportunity and motivation for methodological breakthroughs that support reliable detection, location, and characterization of proliferation even when the number of events is limited. Going forward, strategic investments in AI have the opportunity to extract critical puzzle pieces from the nearly infinite streams of data being collected on a daily basis to accelerate time to insight.

One area that may be able to leverage AI industry progress is in onboard sensor AI engines for event detection and point-of-sensing data reduction (e.g.- for satellite or unattended terrestrial sensors). For example, the size, weight, and power (SWaP) constraints on the AI hardware in these applications share many requirements with hardware developed for the autonomous vehicle market, though the unique non-proliferation algorithm and application software development will require close multiscale co-design with the evolving AI hardware ecosystem to ensure reliability, auditability, etc. for this high-consequence application.

Risks. Without increased and sustained investment in this area, we will be left behind by foreign actors, and our analysts will continue to be overwhelmed by the data deluge that they face. The ultimate result will be unmetered global nuclear proliferation by unknown and unstable foreign actors.

Challenges. Cross-modal search and retrieval—between images, video, and text, for example—are among the bedrocks of advancements in this field. Large-language models and emerging foundation models (Chapter 02) that can demonstrate emergent properties on new tasks provide another key innovation supporting this mission space. These technologies demand substantial computing resources for both training and inference (Chapter 18). Truly enabling these technologies on the unique data sources within the nonproliferation mission space will place unprecedented demand on existing and planned computational resources for the myriad of missions and models that will be developed. Meeting this demand will itself require new advances in AI hardware architectures, software tools, and frameworks, as well as in fundamental mathematical techniques (Chapter 12). Furthermore, critical algorithmic challenges include the trustworthiness and auditability of a model's predictions, as well as the imminent threat posed by an adversarial AI system. In order for the output of these models to support actionable decision-making, new approaches and methods for auditability of the model's construction, training, and predictions are required. Lack of persistent data collection over facilities or other places of interest, including oversubscribed or paucity of sensors, inability to place sensors, and other denial of data streams, present additional challenges that may require new or additional sensors, new techniques to support AI at the edge, or novel methods to do better "tip and cue" to increase persistence.

10.3 Investment Needed for Achievement

To leverage the methods and techniques laid out in Section 01 of this report, it is paramount to prepare the necessary training datasets from simulations and experiments within the DP and DNN mission spaces. For example, some of the tasks that are required for developing foundation models (Chapter 02) are enumerated below. These tasks are required for each dataset, guided by a subject matter expert from within the DOE laboratory complex. This cannot be delegated to any other organization.

- Identify self-supervised learning tasks on broad categories of multi-modal data relevant to DDMD weapons lifecycle.
- Understand the nature of each modality within a dataset and create tokenization schemes that are required for ingestion into foundation or surrogate models.
- Articulate fundamental physical constraints and correlations between a sample's data fields that provide key conservation properties and provide models with elements of physics-informed deep learning.
- Curate labeled datasets for specific downstream learning tasks and model adaptation / transfer learning.

Systematic investments in software to develop AI systems, master models, and composable systems of AI systems with physics models are all needed. The grand challenges identified here all require use of hardware acceleration for training and inference (Chapter 15). Investments in co-design efforts in algorithm and architectures—and ultimately in materials and chips that comprise processor architectures—will be needed to be successful. Finally, investments in AI-empowered data collection from the experimental facilities will be key steps to realize solutions to the grand challenges identified here.

10.4 References

[1] U.S. Department of Energy, 2021. *U.S. Department of Energy FY 2022 Congressional Budget Request*, National Nuclear Security Administration, Office of Chief Financial Officer, Vol. 1, DOE/CF-0171, May.

[2] National Nuclear Security Administration, undated. *National Nuclear Security Administration FY 2023 Congressional Budget Justification*. https://www.energy.gov/sites/default/files/2022-04/doe-fy2023-budget-volume-1-nnsa.pdf, accessed October 18, 2022.

[3] National Nuclear Security Administration, 2022. *Accelerating Product Realization: Aligning the NNSA Nuclear Security Enterprise with Industry Best Practices*, Office of Defense Programs, Science Council, April.

[4] Ellis, J.A., Fiedler, L., Popoola, G.A., Modine, N.A., Stephens, J.A., Thompson, A.P., Cangi, A., and Rajamanickam, S., 2021. Accelerating finite-temperature Kohn-Sham density functional theory with deep neural networks. *Physical Review B* 104(3): 035120.

[5] Zuo, Y., Chen, C., Li, X., Deng, Z., Chen, Y., Behler, J., Csányi, G., et al.,2020. Performance and cost assessment of machine learning interatomic potentials. *The Journal of Physical Chemistry A* 124 (4), pp. 731–745.

[6] National Nuclear Security Administration, 2022. https://www.energy.gov/nnsa/nonproliferation, accessed Nov. 22, 2022.

[7] CNBC Technology Executive Council, 2022. How using analytics and AI can help companies manage the semiconductor supply chain. https://www.cnbc.com/2022/10/19/how-ai-can-help-companies-manage-the-semiconductor-supply-chain.html, accessed Oct. 19, 2022.

SECTION 03: TECHNOLOGICAL CROSSCUTS

Critical crosscutting technology challenge areas must be addressed to harness the promise of new AI methods (Section 01) accelerate progress across the diverse domain areas detailed in Section 02. This effort will require bridging the gap between traditional domain-driven methods and new, AI-based data-driven methods; developing the underlying mathematical and foundations of scientific machine learning (ML); and creating new integrative systems. These systems are crosscutting, comprising workflows, software and frameworks, data, and new types of hardware. In each of these areas, we detail Advanced Research Directions (ARDs), their importance, the gaps that prevent forward progress today, the urgency and timeliness of addressing those gaps, and what is needed to start now.

11. SOFTWARE AND FRAMEWORKS

Scientific software encompasses not only modeling and simulation applications but also analysis codes and system software (see also Chapter 13, Workflows). These software systems play an increasingly vital role in all areas of science, energy, and security. To date, production and research scientific software has followed a path independent of mainline artificial intelligence (AI) and machine learning (ML) frameworks such as PyTorch [1] and TensorFlow [2]. Advancing the complex approaches described in Section 01 will require significant progress in software, frameworks, and their integration. Tighter integration between scientific software and frameworks will not only facilitate such integration but will also improve the productivity of scientists and software/framework developers.

An effective software stack is needed to bridge the chasms between mathematical foundations, data, workflows, and hardware. Different science, energy, and security domain applications entail different constraints, such as assurance requirements, compute/data latency, energy consumption, inference time, resource availability, and knowledge distillation. Under any combination of constraints, software and frameworks should be high performing. The goal is thus to ensure that AI and domain capabilities, efficiency of computational and data resources, and domain and developer expertise are not sacrificed.

11.1 Advanced Research Directions in Software and Frameworks

Harnessing the new approaches described in Section 01 will require highly advanced and modular software ecosystems. Here, we amplify five Advanced Research Directions (ARDs) along which key innovations are needed for software and frameworks to enable diverse breakthroughs in science, energy, and security on high-performance computing (HPC) systems. These ARDs apply to the full software stack and also involve computational science domains ranging from advanced simulation to programming languages. These ARDs exemplify what is needed so that AI capabilities can be quickly and easily built, tested, deployed, continuously optimized, and trusted for applications critical to the U.S. Department of Energy (DOE). Moreover, they position the DOE enterprise to adapt to and harness the continued evolution of diverse AI workflows as AI capabilities.

11.1.1 ARD 1: COMPOSABILITY OF SCIENTIFIC SOFTWARE, HARDWARE, AND AI FRAMEWORKS

We will need an infrastructure that provides unified, interoperable, efficient organization and communication among multiple AI and physics-based models and simulations across scales, control systems, and sensors. It must be agnostic to changing hardware needs for autonomous systems and to changing ML software, and it must allow us to leverage new community- and vendor-provided tools as they emerge. Composable hardware (e.g., discussed in Chapter 15) will enable the underlying system architecture to be optimized at runtime, enabling massive-scale AI applications to map fine-grained computations to the most efficient microarchitecture. We must develop new software and frameworks that facilitate a wide range of AI models for the edge-to-HPC computing continuum (supercomputers, near edge clusters, and edge devices; see also Chapter 13: AI Workflows). The frameworks that enable efficient processing of large-scale datasets and continual learning for real-time control will also be required for scientific instruments and facilities, as discussed in Chapters 04 and 05 of this report. Software frameworks that integrate large language models for integrating domain-specific scientific knowledge from scientific literature into AI models will enable the creation of more accurate and robust models for scientific research.

11.1.2 ARD 2: UBIQUITOUS DIFFERENTIABILITY OF SCIENTIFIC SOFTWARE

End-to-end differentiability for composing simulation and inference in a virtuous loop is required to integrate first-principles calculations and advanced AI training and inference, as discussed in the context of HPC surrogate models in Chapter 01. Continuous integration of differentiable programming capabilities will ensure that computational domain capabilities are AI-ready for the future. AI-optimized hardware that supports differentiability as discussed in Chapter 15 will require deep co-design across algorithms, the software stack, and the underlying hardware. We need differentiability in the scientific software to enable verification and validation (V&V) of scientific software (simulation and AI) and to provide capabilities for analyzing their correctness and reliability—as discussed also in Chapter 12, Mathematics and Foundations. We must develop software frameworks that enable robust and reliable differentiability of large parallel and distributed applications in the presence of noisy experimental data or complex systems.

11.1.3 ARD 3: PORTABLE USABILITY OF DOE SOFTWARE ON EXASCALE AND POST-EXASCALE HETEROGENEOUS AI HARDWARE

Production HPC systems are complex engineered systems comprising many software layers that need to be tuned for each new hardware configuration and workload and for which optimization choices must be revisited as the hardware, software, and/or workloads evolve. For new platforms, as developed in ECP and anticipated with quantum computing architectures or new AI-oriented hardware (Chapter 15), portability is essential. The development of AI-enabled software frameworks and programming models to automatically provide these capabilities across this complex landscape—with a software framework that enables quick and easy sharing, deploying, diagnosing, and testing across systems and models—will significantly improve development and execution time, as well as allow predictable resource forecasts (execution time and memory) to inform real-time control. We must develop intelligent software tools with proactive and reactive capabilities to optimally distribute workloads across various hardware components with different hardware characteristics. The software frameworks will provide functionalities to manage and mitigate the complexity of using exascale and post-exascale systems, employing intelligent automation and predictive analytics. We must develop interfaces to enable users to easily access the full capabilities of exascale and post-exascale systems using natural language processing and visualizations.

11.1.4 ARD 4: REPRESENTATION FLEXIBILITY AND EXTENSIBILITY FOR MULTIMODAL SCIENTIFIC DATA

Science, energy, and security data take many forms and modalities, and these data are central to the creation and training of fundamental new AI capabilities described throughout Section 01 of this report. We will need software and frameworks that readily address concurrent forms of data (including graphs, grids, point clouds, and unstructured data), enable fast computation with native representations, facilitate expressive features and outputs, and allow for certain data to be protected (e.g., due to privacy concerns or to protect intellectual property [IP]). Considering the deluge of data (discussed in Chapters 14 and 19), the integration of data from different sources and with diverse formats (multi-modal) will open a new front on the problem of data reduction: what data is critical to keep considering whole multi-modal data sets to keep opportunities for scientific discoveries? This problem is beyond classic reduction techniques and must consider semantic aspects of the data (which is not the case with current data reduction methods). We must develop advanced data integration tools with semantic technologies to facilitate new methods such as foundation models (Chapter 02) that build on large language models, as these tools will enable multimodal scientific data to be integrated and combined across different domains and applications. These also require support so dynamic and adaptive representations of multimodal scientific data can be developed. New scalable software frameworks are also needed for the exploration and visualization of multimodal scientific data with interactive, easy-to-use interfaces, enabling insight as well as evaluation of the efficacy of various representation schemes. Furthermore, we need service-oriented software frameworks and tools to enable seamless exchange and sharing of multimodal scientific data. These frameworks must support ever-larger teams working across different research teams, organizations, and communities using scalable, open-data platforms and repositories.

11.1.5 ARD 5: TRUSTWORTHY AND SCIENTIFICALLY RIGOROUS AI SYSTEMS

Complex problems in discovery science and high-consequence applications demand ready availability of advanced uncertainty quantification (UQ) and V&V capabilities (further discussed in Chapter 12). Such capabilities include the end-to-end propagation of probability distributions throughout a software stack, the quantification and attribution of errors and approximations, and V&V in settings where hardware, operating system, or algorithms are nondeterministic. Advances in hardware technology such as UQ-optimized microarchitectures necessitate deep co-design with these architectures and will enable orders-of-magnitude improvements in these AI applications. Trustworthy AI systems are needed for near-real-time evaluation of correctness and accuracy at an experimental facility. We must develop AI software and frameworks that enable users to easily understand how AI algorithms make decisions and predictions and explain them in an easy way through interactive language models. Software frameworks that hide the complexities of reproducible research practices and robust experimental design will improve trustworthiness of AI systems. Concurrently, as discussed in more detail in Chapters 14: Data Ecosystem and Chapter 19: Data Infrastructure, these capabilities must be integrated into the entire data lifecycle, given the intimate interdependence between training data and trustworthiness.

11.2 Why Is It Important?

Advances in AI-enabling software and frameworks are critical for meeting the needs of grand challenges in science, energy, and security such as those highlighted in Section 02 of this report. These advances are also critical for ensuring that DOE's excellence in computational and mathematical science is fully leveraged to realize the long-term, AI-based breakthroughs (see Section 01) necessary to achieve these grand challenges.

For software and frameworks to catalyze advances on grand challenges, there must be a virtually seamless integration across model authoring, simulation, data, and compute infrastructure. When advances are focused on any component in isolation, the limitations of all AI-enabling components multiply and propagate throughout. For example, coupling multiple systems on heterogeneous, emerging architectures where multiple AI models are being trained and used for inference on multiple tasks requires a level and complexity of composition to span a large number of types of simulations, problems, and software systems. We will need a modular, standardized, and readily extensible application programming interface (API) for resilient, plug-and-play interaction with legacy and emerging technologies. The composability enables more efficient and effective analysis of large and complex datasets tailored to specific science domains. Without such composability, we risk duplicative and costly piecemeal integration of ML models and AI workflows that will impede fast progress on grand challenges.

A primary feature of existing ML frameworks is built-in differentiation capabilities (see 11.1.3). The availability of derivatives for training, through automatic differentiation (autodiff) techniques such as backpropagation, has been vital to the success of deep learning and beyond [3]. Differentiable programming capabilities can especially impact relevant domains that have benefited from forward simulation advancements but have seen fewer developments for inverse design, control, and other derivative-heavy outer processes [4].

In a typical scientific software development cycle, the rate at which the first-principles science, energy, and security models change is often slow. Consequently, scientific software requirements and specifications remain constant over a longer period, and the steps involved in the development cycle—such as testing, validation, verification, and scaling—remain relatively stable. In contrast, the rapid iterative nature of AI/ML model training and inference, combined with continual learning, pose a number of unique challenges. Specifically, not all of the data required for training the model are available in advance; instead, the data may be acquired over time. As new data become available, models must be retrained, validated, verified, and rapidly deployed in production. The development of software frameworks to enable such rapid model iteration is critical to improving the usage and effectiveness of overall AI-enabled approaches.

The diversity and constant evolution of hardware architectures and compute and data environments also require that software and framework solutions must perform across platforms and use cases. Such portability is fundamental to building user trust in the capabilities and reliability of AI-enabled processes in science, energy, and security. However, there is a natural tension between software portability and performance, and future hardware architectures that are highly optimized for specific tasks will require major advances to achieve performance, such as just-in-time compilation coupled with dynamic hardware reconfiguration. AI software and frameworks that are developed for science, energy, and security can be customized to the specific data and goals of a domain, leading to improved accuracy and performance of AI models.

DOE mission domain-driven AI models often have unique requirements and challenges that are not well served by generic AI solutions. For example, scientific software frameworks in use today were not designed to contemplate distributed, federated data injection and collaborative and interactive model development at scale (using DOE supercomputers with AI accelerators). Significant advances in the scale of frameworks have been demonstrated in industry, but these have been predominantly focused on very different applications, such as involving text and image data. This significantly affects the AI development cycle and overall scientific productivity. We must develop software tools and frameworks with enhanced collaboration and interoperability.

Reproducibility of the AI models requires software tools and frameworks with improved data and model management capabilities for large and complex datasets that are frequently encountered in the domains outlined throughout Section 02 of this report. This reproducibility is today nascent with the generic AI frameworks; addressing this will be important for assuring scientific integrity and correctness. It is crucial that these frameworks support but hide the complexities associated with large-scale data and model provenance.

Science, energy, and security applications are rarely concerned with a single prediction or decision in isolation. Instead, we test multiple hypotheses, confront multiple scenarios, and account for sources of uncertainty. Efficiently performing ensembles of computational tasks and producing probabilistic outputs are key to building confidence in AI-enabled advances.

11.3 Why Can't It Be Realized Now?

Traditionally, scientific software efforts have been focused primarily on forward simulation, that is, being able to develop digital twins of phenomena and systems encountered in science, energy, and security. To this end, various research sectors and industry domains converged on standards for model representation, simulator exchange, and distributed co-simulation [5, 6]. Development efforts have focused on achieving performance and scalability at ever higher levels of fidelity and for ever larger, more complex systems. In many areas, however, the focus on fidelity in such forward models has come at the expense of consideration for high-level tasks such as inverse design and autonomous discovery. The complex performance and fidelity optimizations of mission-driven software stacks will need to be reexamined in the context of new drivers and approaches like those noted in Section 01 of this report. Similarly, the requirements of existing frameworks have been driven by data, hardware, and uses that necessitate more than simple adoption of the advanced approaches envisioned here.

Current **composition** strategies often come at the expense of limiting capabilities to the weakest link in the composition. For example, straightforward composition in a software stack will often come at a significant overall performance expense, because we are limited to employing particular instantiations of the various components, and these components are often optimized for inputs, data flows, and hardware resources outside of the environments in which a fully composed stack is deployed. The complexity of assembling different components is a bottleneck even in current approaches to compose software with respect to correctness, validation, and verification (as well as safety and security for many applications). These factors prevent the community from achieving composable software for science.

Another major bottleneck in the broader adoption of AI technology is the lack of AI/ML frameworks that enable findable, accessible, interoperable, and reusable (FAIR) data and model artifacts (discussed further in Chapters 14 and 19). FAIR AI/ML frameworks have the potential to drive rapid adoption of AI technologies within DOE mission domain areas and to enable synergies and partnerships across diverse areas. *Currently, there is no science-centric AI/ML framework that adopts a systematic approach to relate data, models, and tasks within any particular scientific domain.* The resulting discord between the data and the model increase inefficiency in applications involving large volume of data. Software stacks used in industry do not reflect scale, diversity, and unique characteristics of the DOE mission domain areas. A particularly acute issue is the significant knowledge and technological gap with respect to the emerging AI-driven software development lifecycle because no related prior research exists within DOE or elsewhere (Chapter 16 provides further assessment of DOE's workforce).

Increasingly, there are efforts to employ popular ML frameworks for simulations to realize **differentiable programming** capabilities. However, for many science, energy, and security applications, this approach tends to come with significant costs. First, simplifications tend to be made to the forward models so they can be **expressible** in the existing framework syntax. Second, accuracy and performance compromises are often made in such implementations. These include substituting smooth approximations for known regime changes and fixing the mode of automatic differentiation (e.g., backpropagation) independent of the output and input dimensions. Furthermore, differentiation throughout a production scientific software stack today is often enabled by one-off efforts, such as differentiation through a single LAPACK routine. In addition, current frameworks struggle to retain valuable information when confronted with multiple data modalities. When a differentiation or representation limit is reached, it is rarely the case that such a change in control flow or knowledge is propagated up the software stack, which would otherwise facilitate exploitation in higher-level operations.

Usability in DOE mission-related, grand challenge computational problems, usability tends to be limited to intersectional (hardware-software-framework-data-algorithm-problem) specialists (see also Chapter 16, Workforce). There are many limitations on higher-level software, frameworks, and applications due to hardware and low-level software constraints. Current strategies for **adapting software to new systems** tend to rely on applying heuristics for each piece of software individually—and typically only once. For instance, the memory performance experienced by a computational workload can be affected by multiple layers of memory management policies, from the operating system kernel to runtimes. Systems are becoming increasingly energy-constrained and cannot supply full power to every hardware component at all times. The question of which components to prioritize—and at the expense of which other components—can have significant performance implications, yet it is frequently workload-dependent. Furthermore, while HPC systems have complex job schedulers, individual nodes also have task schedulers, input/output schedulers, and network schedulers, each of which is highly capable and configurable yet rarely adjusted to changing workloads. Floating-point implementations have also evolved, creating an additional knob and an additional source of complexity as domain specialists port scientific applications from one hardware–software stack to another. Current programming models and language choices are also largely incompatible with the emerging AI hardware technologies.

Today, UQV&V tasks such as formal **verification** tend to be limited to narrow classes of operations and implementations. Although the predominant use of ML frameworks is in building models through empirical risk minimization, making use of distributional knowledge, whether from data or first principles, is difficult in general. Furthermore, most scientific software is still devoted to outputting point estimates: At best, simplistic uncertainty indicators such as individual error bars are obtained. Examples of where richer distributional information or end-to-end error and uncertainty propagation are employed in software and frameworks are typically found in niche fields or problems lacking the full complexity envisioned here.

11.4 Why Is It Reasonable to Start Now?

Despite software and framework-related challenges associated with realizing the promise of the approaches in Section 01, an environment is emerging in which dedicated focus on the identified ARDs could enable the sea change needed for transformational advances in AI-driven scientific discovery. This is critically important, as developing and advancing the building blocks described in Section 01—from surrogate and foundation models to digital twins, inverse design, autonomous laboratories, or automated coding— will rely on the software and framework ARDs described here. Below we take the pulse of this emerging environment in other crosscuts and ecosystems detailed in this report.

An increasing number of science, energy, and security domains are employing AI/ML techniques, in some cases through popular ML frameworks. As noted in Chapter 12, the theory underlying ML and its foundations is advancing to make some capabilities provided by today's frameworks increasingly ready for adoption into select high-consequence science applications and settings. A growing understanding of AI/ML techniques' domains of applicability and limitations is emerging. This understanding is allowing practitioners to move beyond the full factorial combination of methods and problems to a reduced, more principled set that better facilitates performant execution.

Vendors and hardware are significantly addressing data-driven settings that facilitate computational performance for AI-enabling technologies. As noted in Chapter 15, these developments are advancing both on general-purpose and highly specialized emerging architectures. Although floating-point conventions will continue to evolve, standards are emerging along with a better understanding of the implications, both in accuracy and performance, for different levels of precision. Algorithms and software for mixed and variable levels of precision have also seen significant development.

Recent improvements in the integration and dynamic nature of the HPC software stack show great potential for quickly delivering and testing more configurable software layers. Concurrently, the application of reinforcement learning and control theory for computer systems has made significant progress.

DOE's Exascale Computing Project (ECP) has hardened a software technology infrastructure [7] that increasingly represents a bridge between the emerging computer hardware and the science, energy, and security specialists tackling large-scale, complex problems (Figure 11-1). These efforts have enabled increased composability across the software stack and are addressing new challenges associated with massive scale and heterogeneity of data. Composability is a critical driver in the research and development of new ML frameworks [8].

In some domains, forward simulation fidelity has improved to the point where leading errors are now associated with unknown parameters, uncertain states, and the like. In others, fidelity levels have improved to the point where control, robotic automation, and targeted design can take advantage of AI/ML. Endowing such applications with UQV&V capabilities and pursuing the approaches in Section 01 are high-potential opportunities.

Probabilistic programming languages (PPLs) are also increasingly mature and provide proofs of concept for propagating probability distributions across a software and programming hierarchy [9–12]. Differentiable programming is also seeing adoption beyond ML frameworks to new classes of applications [13].

There is also an increasing appreciation and understanding of the science of team-based software and frameworks—with many important lessons and success stories from the ECP itself, which has more than 1,000 participants. DOE near-term priorities include the following [14]: understanding practices, processes, and tools that can help improve the development, sustainment, evolution, and use of scientific software by teams; developing next-generation tools to enhance developer productivity and software sustainability; and developing methodologies, tools, and infrastructure for trustworthy software-intensive science.

FAIR AI/ML frameworks are crucial to overcoming the challenges of developing AI/ML models for DOE applications. The DOE Advanced Scientific Computing Research community is in a unique position to develop these frameworks by leveraging past successes in developing

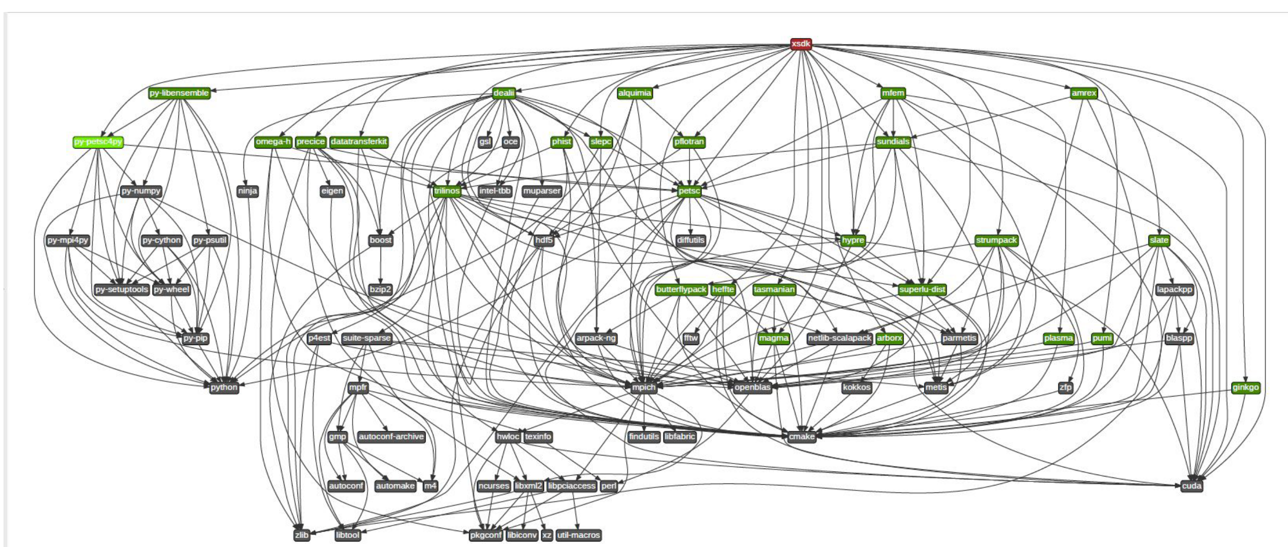

Figure 11-1. An example build tree for ECP's math library, SDK, which illustrates the complex interdependencies among different libraries. Source: Satish Balay, Argonne National Laboratory.

scalable and efficient data and workflow management software tools. These tools enabled researchers from various DOE programs to handle massive amounts of data from simulations, experimental facilities, and observational instruments.

The nature of current and future AI-enabled DOE applications also demands a high degree of autonomy in data generation and model development. AI/ML models require continuous adaptation; and manually doing so will slow AI-enabled scientific discovery. We are at the cusp of access to tremendous exascale computing power capable of designing self-driving AI systems (e.g., as discussed in Chapter 04), which can be scaled to zettascale systems and beyond. Automating critical functions in building and refining AI systems will be critical given the expected diversity of AI-enabled DOE applications that require extreme customization and continuous adaptation. Simply put, accelerating DOE's discovery processes in science, energy, and security will rely upon the availability of robust software tools and frameworks that enable a wide range of AI/ML models across different applications.

11.5 What Is Needed to Start Now?

The ARDs cut across the grand challenges and approaches identified and are an indication of the broad needs for critical advances. Key efforts needed in the near term include the following:

- Develop extensible, **large-scale evaluation suites** for science, energy, and security:
  - Create open abstractions and pipelines for grand-challenge problems to spur community engagement and advances across the software stack.
  - Engage multiple frameworks and ecosystems to understand trade-offs and to accelerate future advances.
  - Create testing and validation suites, standards, and APIs.
- Develop **standards and APIs** to enhance greater composability across scientific software and ML frameworks:
  - Increase the modularity and ease-of-use of sub-monolithic framework blocks into an ecosystem of interoperable and composable microservices.
  - Facilitate the communication of requirements and AI-ready capabilities as technologies evolve.
  - Sustain efforts to automate capability discovery and composition of software blocks by AI technologies.
- Expand **differentiable programming** in scientific software so that it is endowed with properties similar to those of an artificial neural network:
  - Propagate known switches/conditionals up the stack to enable differentiation.
  - Express known dependence structures to be exploitable by the rest of the software–hardware stack.
  - Enable seamless interoperation of autodiff for scientific simulation and differentiable programming for AI that accounts for resource constraints for complex workflows.
- Further the degree of **performance portability and interoperability:**
  - Expose additional hardware-software-workflow configurations.
  - Provide performance models and simulation capabilities for virtual testbeds of emergent hardware architectures and environments.
- Facilitate **performant re-use** of energy-intensive, leadership-class ML capabilities:
  - Train and store large-scale AI models.
  - Enable the ability to recommend a base model and retrain for downstream application and software–hardware instantiation.
- Establish **"born qualified" trustworthiness** for software and framework artifacts:
  - Increase PPL adoption and development.
  - Accelerate adoption of UQV&V-ready capabilities.
- Escalate **extensibility and representability** beyond what consumes current development:
  - Enable software-generating environments.
  - Facilitate computational resources (measured in "Machine Learning Operations," or MLOps) for continuous deployment on and refinement of new problems and environments.
- **Sustain interaction** among the scientific software development community; ML framework developers; computational facilities; and emerging science, energy, and security opportunities.
- **Enable the rapid design, development, and training of fast-learning and reusable AI/ML models** for DOE scientific data and to make the reusable models FAIR by leveraging recent advances in the open-source data and model management tools.

These ARDs are crosscutting and should not be viewed in isolation. We have illustrated key steps for acceleration in compact activities, and advancements along multiple ARDs have the potential for multiplicative effects when realized in concert.

11.6 References

[1] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32.

[2] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. {TensorFlow}: A system for {Large-Scale} machine learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283.

[3] Baydin, A.G., Pearlmutter B.A., Radul, A.A., and Siskind, J.M., 2018. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research,* 18(153), pp. 1–43. https://www.jmlr.org/papers/volume18/17-468/17-468.pdf, accessed May 12, 2023.

[4] Goodrich, C.P., King, E.M., Schoenholz, S.S., Cubuk E.D., and Brenner, M.P., 2021. Designing self-assembling kinetics with differentiable statistical physics models. In: *Proceedings of the National Academy of Sciences* 118(10), e2024083118. https://doi.org/10.1073/pnas.2024083118

[5] Krammer, M., Schuch, K., Kater, C., Alekeish, K., Blochwitz, T., Materne, S., Soppa, A., Benedikt, M., 2019. Standardized integration of real-time and non-real-time systems: The distributed co-simulation protocol. In: *Proceedings of the 13th International Modelica Conference*, pp. 87–96. https://doi.org/10.3384/ecp1915787

[6] Blochwitz, T., Otter, M., Arnold, M., Bausch, C., Clauss, C., Elmqvist, H., Junghanns, A., Mauss, J., Monteiro, M., Neidhold, T., Neumerkel, D., Olsson, H., Peetz, J.-V., and Wolf, S., 2011. The functional mockup interface for tool independent exchange of simulation models. In: *Proceedings of the 8th International Modelica Conference*. https://doi.org/10.3384/ecp11063105

[7] Heroux, M.A., McInnes, L., Li, X.S., Ahrens, J., Munson, T., Mohror, K., Turton, T., Vetter, J., and Thakur, R., 2022. *ECP Software Technology Capability Assessment Report*. https://doi.org/10.2172/1888898

[8] JAX: Composable transformations of Python+NumPy programs, v. 0.3.13, 2018. http://github.com/google/jax

[9] Baydin, A.G., Shao, L., Bhimji, W., Heinrich, L., Meadows, L., Liu, J., Munk, A., et al., 2019. Etalumis: Bringing probabilistic programming to scientific simulators at scale. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19)*, Article 29, pp. 1–24. https://doi.org/10.1145/3295500.3356180

[10] Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N.D., 2019. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(1), pp. 973–978. https://doi.org/10.48550/arXiv.1810.09538

[11] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A., 2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), pp.1–32. https://doi.org/10.18637/jss.v076.i01

[12] Salvatier J., Wiecki T.V., Fonnesbeck C., 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science,* 2, pp. e55. https://doi.org/10.7717/peerj-cs.55

[13] Schoenholz, S.S., and Cubuk, E.D., 2020. JAX, M.D.: A framework for differentiable physics. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS'20)*, Article 959, pp. 11428–11441. https://doi.org/10.48550/arXiv.1912.04232

[14] Bernholdt, D.E., Cary, J., Heroux, M., and McInnes, L.C., 2022. *The Science of Scientific-Software Development and Use*. https://doi.org/10.2172/1846008

12. MATHEMATICS AND FOUNDATIONS

The current science and engineering paradigm is rooted in mathematical models that are validated against experimental data. These mathematical models are derived by scientists and engineers based on first principles understanding and well-defined unifying concepts. Importantly, this is very different than modern machine learning, in which mathematical model forms are highly flexible and applicable to many domains [1]. First-principles models are naturally predictive beyond the datasets used to learn and validate the theories, because they are designed to be consistent with established science. In contrast, artificial intelligence (AI) and machine learning (ML) models may have challenges generalizing beyond their training data because they are much less constrained. However, first principles often are too complex to work with directly, necessitating approximations derived for certain domains of applicability and to fit various constraints. As scientists and engineers, we can use well-established, domain-driven methodologies to evaluate the validity of these models, recognize inconsistencies, and identify improvements. Additionally, we have rigorous statistical and mathematical tools to work with domain-driven models and infer conclusions [2, 3]. Through the scientific process, these models evolve to fit new data and better reflect reality.

The promise of AI/ML approaches, as detailed in Section 01 of this report, is that they offer a pathway to develop principled, data-driven models to extract insight with data-driven methodologies. These methods will complement domain-driven methodologies, and they will do so at the unprecedented scales of data generation we see today. Further, AI/ML models can be used to create systems that make decisions and perform inference with limited human input and involvement. These models complement traditional first-principles scientific models, as their flexible mathematical structures and learning methods enable the development of building models where first principles understanding does not exist or is too complex to practically leverage. Therefore, to deliver on the promise of AI/ML, methods are needed to bridge, in a fundamental way, domain-driven methods and data-driven methods.

Among other advantages, AI/ML methods can automate the learning process while reducing dependence on scientists and engineers—humans with limited availability and capacity for scale and computation—to specify constraining assumptions. The more we can relax these constraints, the more we can harness AI systems to learn from highly diverse data sources—including those that scientists have not thought to leverage and those at scales beyond human capacity to leverage.

In addition, the rapid advances in scale, capabilities, and applications of AI models in recent years have created new systems, many with emergent properties. However, the inner workings of these models are opaque—raising challenges in explainability, trust, and uncertainty quantification (UQ).

Harnessing the opportunities possible with AI to advance the U.S. Department of Energy's (DOE's) scientific and engineering mission (Section 02 of this report) will require developing the mathematical foundations of scientific AI/ML, combining traditional domain-driven methods with newer data-driven methods in principled ways. This will build on and complement the foundations of AI/ML more generally, enabling us to ground new developments with the same mathematical rigor as has undergirded traditional scientific and engineering exploration, design, and operation. We want AI/ML that can predictably generalize, have understandable approximations with clear domains of applicability, integrate with other sources of knowledge, and propose improvements to the models when theory and/or data are inconsistent.

12.1 Advanced Research Directions in Mathematics and Foundations

Our understanding of mathematics and foundations for AI/ML ranges from the philosophy of science and epistemological foundations of AI/ML designing AI models and algorithms for efficient training. We have identified four Advanced Research Directions (ARDs) where current mathematics and foundations are not yet prepared to meet the needs for future AI necessary to support DOE science, energy, and security mission areas. We structure the balance of this chapter around these four ARDs, here describing them at a high level and in subsequent sections addressing their collective importance, the challenges that must be overcome, why it is urgent to begin now, and what steps are needed.

12.1.1 ARD 1: DEVELOP FOUNDATIONAL PRINCIPLES AND ALGORITHMS FOR SELF-GUIDED LEARNING OF AI SYSTEM WORKFLOWS

AI must be more *self-guided*. These systems must be able to tune and optimize themselves to meet abstract specified goals by adjusting the AI implementation (e.g., learning methods, structure, models, hardware), select informative data, recognize/adapt to changing environments, and provide self-certified notions of trust. Progress in this direction will result in gradually relaxing the specifics of design constraints users must provide for the model, data, and task.

12.1.2 ARD 2: DEVELOP INFORMATION THEORETIC MECHANISMS TO INTEGRATE SCIENTIFIC PRIOR KNOWLEDGE, THEORIES, AND MULTIMODAL DATA

Scientific AI systems must be able to synthesize existing scientific knowledge (e.g., physics or mathematics properties), heterogenous big data (e.g., multi-fidelity, multiscale, multi-phenomenology), and limited small data (e.g., rare events, expensive simulations/experiments) (see [4] and Chapter 04 for the case of digital twins). Integrating solutions for these challenges is critical for robust and trustworthy inference using AI. Further, new AI systems must be able to build on existing AI systems much as new scientific theories build upon the existing science.

12.1.3 ARD 3: DERIVE FOUNDATIONAL PRINCIPLES AND THEORY FOR DECISION-MAKER TRUST IN AI

DOE has been at the frontier of UQ and verification and validation (V&V) research for science and engineering modeling; adapting these methods to AI/ML and associated workflows will be central to addressing explainability, correctness, and trust. Together, UQ and V&V broadly include three sets of capabilities. First, UQV&V entails theory, methods, and algorithms that learn with uncertainties (e.g., Bayesian inference and ensemble methods). Second, they assess sensitivities to inputs, data, assumptions, model forms, and approximations (e.g., global sensitivity analysis). Finally, UQV&V methods validate against other sources of data (e.g., cross-validation), evaluating the correctness of algorithm (e.g., formal methods), or integrate human knowledge as an additional layer of validation (e.g., explainable/interpretable AI). If AI/ML models cannot improve, quantify, and communicate their robustness, they will lack the fundamental underpinnings necessary to be embedded in systems involving mission-critical decisions and processes (see specific examples in Section 02: Scientific Domains). Because AI systems will interact with humans, they must provide suitable evidence, as judged by the decision-maker, to establish confidence in the AI's assertions. This demands new UQ, V&V, and explainable/interpretable methods [5] to communicate reliability and uncertainty, perform predictably (importantly, over multiple scales, different domains, and compositions with other models), adapt to changing environments, and operate securely.

12.1.4 ARD 4: DEVELOP THEORY AND ALGORITHMS TO QUANTIFY AND OPTIMIZE TRADE-OFFS IN THE IMPLEMENTATION OF AI SYSTEMS UNDER RESOURCE, PERFORMANCE, AND ROBUSTNESS CONSTRAINTS

New AI systems required by DOE missions must also scale in complexity to meet resource and robustness constraints. This ranges from scaling AI down—to operate within individual components of an experiment or instrument—to scaling AI up to support distributed learning in systems harnessing multiple DOE computing and other user facilities. Constraints force trade-offs within multiple dimensions including *resources* (e.g., cost, computation time, power, bandwidth, and data), *performance* (e.g., learning metrics, Quantities of Interest (QoI), accuracy, and rewards), and *robustness* (generalizability out-of-domain, stability, adaptability, representations of uncertainty, and integration of knowledge). The trade-offs among these dimensions—resources, performance, and robustness—must also be quantified. This will require new mathematical principles to explicitly translate resource, performance, and robustness specifications into metrics for the AI model. In turn, novel algorithms that can efficiently explore the Pareto front defined by these trade-offs will be required to support the design of AI systems to meet design criteria.

12.2 Why Is It Important?

Research in the mathematical foundations of AI/ML has been active in topics ranging from foundational questions of epistemology and statistical learning theory to theories of representation complexity of different learning models, to theories of optimization algorithms. This theoretical basis shares some foundations with existing domain-driven scientific learning methodologies (e.g., Bayesian epistemology), but in some respects differs significantly (e.g., domain-agnostic models and extreme over-parameterization). It is thus of central importance to develop a foundation for scientific ML, integrating both domain-driven and data-driven approaches. This is necessary to develop and apply the science and engineering discoveries of a scale and complexity that is demanded by the DOE science, energy, and security missions. Because these scales and complexities exceed the limitations of human domain knowledge and expert judgment, AI/ML methods are not only critical to the scientific computations but also to their evaluation and certification.

DOE mission challenges involve incredibly complex systems applied in high-consequence domains, involving a wide range of challenges stemming from either a paucity or deluge of data, integrating existing knowledge, computational and experimental resource constraints, and robustness and trustworthiness. For example, AI deployed for problems like climate prediction, nonproliferation, power grid operations, and complex system operations in inhospitable environments will present unique challenges.

In contrast, AI research today is largely dominated by social media and internet industries, such as those dealing with scale in terms of millions of consumer devices or aggregate workloads comprising relatively small, and independent, applications and which focus on a very different set of challenges. These different downstream goals lead to different problem formulations, different notions of model quality, and different technical requirements, often in important but subtle ways. For example, driven by goals of high-quality predictive models, industry has developed highly impactful but opaque AI methods that significantly outpace our ability to rigorously understand them. For many intended industry applications—for example, consumer services such as facial recognition in photo libraries or interpreting voice commands—UQ, V&V, or explainability are not required. Consequently, in order to adapt and leverage the rapid pace of industry AI innovation for advancing the DOE mission areas—which demand quantified robustness and explainability—we must develop the requisite mathematics and foundations. This requires substantial investments in the foundations of scientific AI/ML to complement the applied mathematics foundations underlying scientific computing, where DOE has a large body of expertise. Absent such investments, AI/ML methods will likely fail to support the robustness and complexity required for DOE mission areas of science, energy, and security.

Finally, as complex AI systems demonstrate robustness and correctness, they will become integral to many processes that will inform designs of materials, components, or critical engineered systems, including complex systems and infrastructure operations. Here, robustness and correctness certifications that are not grounded in solid mathematical foundations and derived from theory-based tools (e.g., for UQV&V) would create false confidence. This would render them vulnerable to unanticipated failure modes, such as those associated with errors, overfitting, or even data poisoning by adversaries. In a real sense, this would be worse than having no certifications. Beyond the mission impact, AI/ML model failure in such cases would erode confidence in the use of AI/ML in the future and result in lost opportunities to fully realize the benefits such as outlined throughout this report.

12.3 Why Can't It Be Realized Now?

Here, we discuss the current barriers in context of the four ARDs outlined above.

ARD 1. The promise of autonomous discovery (Chapter 05) and complex systems and infrastructure control (Chapter 04) through ubiquitous AI requires tackling the challenges of self-guided learning. We require AI to rapidly adapt and respond to large amounts of streaming heterogenous data from highly dynamic and nonstationary systems. One example of such a system is the future smart grid composed of millions of autonomous AI actors (e.g., systems within components or control infrastructure) making decisions for control through demand response, at different time and geographical space scales, from appliances to electrical distribution networks (Chapter 08). These AI systems need the ability to learn autonomously from partial information and to adapt and evolve in response to rapidly changing conditions. Self-guided learning will enable AI actors to make (or recommend) decisions to the degree that they can develop full situational awareness and evaluate multiple potential responses and outcomes (Chapter 04). In the worst-case, a poorly self-guided AI system will be highly confident but wrong because it is acting on an outdated and/or inadequate understanding of the system, which can cause unreasonable and potentially catastrophic decisions. Therefore, we need to be able to apply physical constraints on the operation of self-guided AI for safety (e.g., closed-loop stability in control).

Active learning, optimal experimental design, control theory, and reinforcement learning (RL) provide a strong foundation for self-guided AI. One critical challenge, however, is learning subject to multiple objectives or with poorly defined objective functions [6]. Specifically in autonomous discovery, it becomes difficult to define the task and cost functions that guide these algorithms. Therefore, research is needed to identify new, goal-oriented, and information theoretic learning paradigms for self-supervised learning that can learn, in effect, everything interesting that can be learned from the available data. A second challenge centers around the data and computational complexity of existing self-guided methods. RL training is often computationally expensive, requiring large volumes of training data and many training iterations to effectively navigate in the high dimensional optimization space [6]. This is particularly true in online settings where the algorithms must balance exploration and exploitation and where learning must be done sequentially. We must develop new algorithms for training, more compute- and data-efficient RL, hierarchical models that learn at different levels of abstraction and spatial-temporal scales [7], and methods to leverage prior information (e.g., physical constraints, ARD 2).

ARD 2. We must learn how to incorporate prior knowledge from science and engineering theory into AI systems. This will entail work in areas including first principles theory (e.g., physics), mathematical models, structure preservation, and models of uncertainty [8]. This encompasses developing useful data representations for common science and engineering data, like those that exist for natural language processing (NLP), that can be used to integrate scientific data into common models. This is particularly critical for foundation models (Chapter 02) that often rely on transformers that, in turn, rely on tokenizers, embeddings, and positional encodings [9, 10]. This means defining notions of concepts and context for scientific data, as tokenizers and embeddings for scientific data would segment the data into concepts with defined relationships while the positional encoding retains important contextual information about how and where those concepts occur in the data. We also must learn how to reduce scientific datasets to be efficiently ingested by AI. Scientific experiments and simulations often have very large output (e.g., snapshots from decadal climate simulations for many different choices of parameters). Current training methods and models such as transformers scale poorly with the dimension of the input space, making it essential to develop algorithms for reducing datasets and scalable training.

Conversely, we must also learn how to extract interpretable knowledge from AI and translate it into scientific theories. A good example of this is when AI is used for autonomous discovery of novel physics (Chapter 05). Similarly, we need to develop a more robust theory of transfer- and multi-task learning that identifies commonalities between data, models, and tasks to enable robust information fusion. Self-supervised learning frameworks (ARD 1) should improve learning by using AI capabilities to autonomously seek out and utilize extant data and other models. Data reduction also plays a role as AI will learn to optimally reduce past datasets into salient summaries to train future models when new data is available. Reconstructing past datasets from the models and data summaries is also needed when the original data is lost. Data summaries will impact models, particularly foundation models, derived from data that is too large to store in totality but will need to be periodically updated. Therefore, foundation models that are designed for efficient sequential updating both in terms of adding new data and prototyping new model structures for improved performance will be necessary.

ARD 3. We need to extend and adapt current V&V frameworks for application to AI/ML models. Throughout this report, particularly in Sections 01 (AI Approaches) and 02 (Scientific Domains), large-scale and/or complex AI systems are discussed. V&V will be critical to underpinning trust in every step of the ML pipeline by certifying the performance of each step in the pipeline and identifying those that are problematic. Such V&V frameworks would, for instance, isolate the providence of poor performance to identify a faulty training algorithm, an inappropriate ML model, or an issue with data [11]. This will require not only developing methods that test a given AI algorithm for generalizability and prediction accuracy but also assess the reliability of data, modeling assumptions, and even implementations of the AI algorithms themselves on novel hardware/software environments. Ultimately, we need rigorous mathematical theories that can provide quantifiable assessments of the suitability of various AI methods to address a specific problem, quantify sensitivities to errors and adversaries, and provide certifiable performance bounds. Additionally, theory must be developed to quantify the utility of a dataset (particularly for any that are small) and determine if it is sufficient for the intended learning objective. V&V-like methods must also be developed to quantify the operational envelope of a given AI system, which is a critical aspect in creating composable AI systems for tasks like control of cyber-physical systems (Chapter 04). Rigorously addressing these AI and V&V concerns using current algorithms such as Bayesian UQ requires many assumptions and approximations in how we represent information such as priors (ARD 2) and how we solve the UQ problem tractably [12] where addressing resource demands has significant impact on quality [13]. Research is needed to better understand these and similar trade-offs (ARD 4). Additionally, algorithms and approaches that provide UQ for cutting edge architectures must be explored because it is unclear to what extent UQ methods developed for one architecture translate to new architectures. For example, introducing Bayesian UQ for attention-based deep learning models is a nascent area of research. Existing approaches must be adapted in order to best fit our conceptual understanding of the self-attention mechanism, and they must also still be efficiently trainable with back-propagation [13, 14].

In order to facilitate adoption, stakeholders require not only UQV&V capabilities, but also methods for effectively integrating them into decision-making processes. This will mean addressing questions such as how to present UQ in a way that is actionable and understandable in specific operational contexts from the standpoint of decision-makers. Explainability and interpretability will be critical ingredients to trust, particularly as AI models become increasingly complex and otherwise opaque. New methods must be developed that can identify the type of information (e.g., modalities, datasets, task similarities) that is being used to inform decisions, particularly in the adoption of AI foundation models (Chapter 02). Finally, it will be important to explore connections with advances in self-guided learning (e.g., optimal experimental design) to identify new ways that AI system users can identify potential weaknesses and suggest improvements in AI systems. This will both mitigate concerns and provide stakeholders with the information necessary to support operational use. This will require theory-based tools for parsing stakeholder needs and translating them into criteria for the AI to present evidence of trust and improve itself.

ARD 4. DOE science, energy, and security missions face particularly challenging operational requirements, such as those associated with very short timescales, high-consequence decisions, or inhospitable operational environments. These resource constraints force trade-offs that must be understood between resources (e.g., cost, compute time, power, bandwidth, data), performance (e.g., learning metrics, QoI accuracy, rewards), and robustness (generalizability out-of-domain, stability, adaptability, representations of uncertainty, and integration of knowledge). Quantifying these trade-offs–and developing solvers for designing AI in the face of these constraints–is a grand challenge that will uniquely impact DOE missions. One specific need that we foresee is simplifying large foundation models. These models often have billions of parameters, making them too large for hardware constrained problems. One strategy would be to tune them for specific tasks.

12.4 Why Is It Reasonable to Start Now?

ARD 1. As we move to increasingly complex and automated systems for discovery and control, AI must be more *self-guided*. The success of large industry models (e.g., as discussed in Section 01 of this report, ranging from foundation models to property inference and inverse design) suggests that we embrace even greater expectations for how AI will affect science and engineering. This means broadening our expectations for what AI can tune via exploring and identifying meaningful prior knowledge, data modalities, model structures, learning algorithms, training hardware, and UQV&V methods. This ultimately will make all steps in the AI pipeline self-guided.

ARD 2. Without tackling problems of information fusion, our AI/ML methods will be limited to standard, supervised, and often data-intensive learning approaches where AI learns only from data gathered from the target task. Many problems of interest cannot provide these datasets due to limitations such as that the data is too expensive, does not exist for the exact target process, or is too unstructured with poorly understood relationships between observables. Recent advances in AI/ML have illustrated the power of going beyond the standard view of learning (supervised, single-task, data-centric) to facilitate learning particularly in limited data settings [15, 16, 17]. These advances increasingly replicate the critical human capability of leveraging prior and disparate knowledge sources to draw inferences.

The possibility of integrating prior knowledge either encoded by scientists (e.g., physics-informed neural networks [18, 19]) or captured by prior AI tasks (e.g., transfer learning, foundation models) is a significant opportunity and necessary for three reasons [12]. First, prior knowledge, when appropriately applied, fills in gaps in the data, providing much better generalizability in AI. Second, prior knowledge can constrain (e.g., with physical laws, multi-fidelity models [20]) AI systems to make them more trustworthy and predictable, as we know they will behave in certain desired ways. Third, building upon prior knowledge allows for the scaffolding of knowledge that is central to science and engineering.

Another profound opportunity is the integration of heterogenous data from a variety of tasks. Multi-task learning methods, like those foundation models, demonstrate the single models that learn many different tasks by leveraging latent commonalities in tasks [15, 21]. This even allows them to perform tasks for which they have not been trained (e.g., zero-shot learning). This integration of diverse data sources is exactly what we are looking for in autonomous discovery to identify novel processes and relationships in complex science and engineering data (see Chapter 04). While individual task-specific data may be limited, DOE facilities are generating exponentially more data from a diversity of tasks, and these can be integrated and leveraged to train such multi-task learning models.

ARD 3. DOE has a long history of leadership in UQV&V for science and engineering. Leveraging this expertise and integrating it into state-of-the-art AI provides a significant opportunity to uniquely contribute to AI and harness its potential to support ever-increasing demands spanning DOE mission areas (see Section 02). We have identified three specific directions where the DOE can contribute in the near term to maximize opportunities: identifying principled UQ and V&V, quantifying and communicating trust for stakeholders, and certifying composability.

First, *principled and certifiable UQV&V* are central to using AI on challenging problems, particularly those in DOE mission spaces that involve limited data, out-of-domain predictions, and high consequence decisions. Secondly, the lack of adequate basis for trust limits the adoption of and ultimately investment in AI capabilities by stakeholders. By understanding the components of trust necessary to enable stakeholders to rely on AI systems (i.e., to quantify and minimize risk), we will be able to increasingly integrate it in DOE mission spaces. Finally, we see specific opportunities in providing robust and certifiable composability of AI systems to enable the systems-level thinking that is a central part of many DOE mission spaces, ranging from the certification of the nuclear stockpile (Chapter 10) to the design of future power grids (Chapter 08) to the control of complex systems using digital twins assembled from individual component models [4] (Chapter 04).

ARD 4. Resource-constrained problems are common in many DOE challenge areas where AI is being applied or considered, including control and optimization of complex engineered systems, autonomous discovery, AI-at-the-edge [22], large-scale foundation models, federated learning [23], and surrogate models in high-performance computing (HPC). If we do not tackle foundational challenges in understanding and navigating trade-offs, DOE will find it increasingly difficult to leverage private industry's rapid advancements in AI due to differences in operational requirements. For example, large AI models, which have become popular in industry, will require significant optimization to fit DOE mission constraints like robust operation to adversarial attacks or operating on limited computing hardware. Additionally, as AI becomes increasing intensive, resources required to train models could become unsustainable [15] in terms of data collection, computational resources, and efforts needed for V&V.

12.5 What Is Needed to Start Now?

12.5.1 GOALS FOR 1–3 YEARS

Goals include creating AI algorithms, especially for federated learning and foundation models, with defined performance and computing (e.g., bandwidth, memory, and computation time) constraints at scales spanning from embedded systems to HPC.

- Theory and methods to assess data requirements for an AI task.
- Foundational studies on key aspects of stakeholder trust through both AI/ML and cognitive science [24].
- Empirical research into scientific data representations and multi-task learning for foundation models in science and engineering to guide future theories.
- Algorithms and model forms that allow sequential updating of foundation models with new datasets and prototyping of new model structures.

12.5.2 GOALS FOR 3–5 YEARS

Goals include creating theory and methods to translate requirements (e.g., V&V, resource constraints, and explainability) from natural human descriptions for self-guided and resource-constrained AI.

- Theory and methods to translate human descriptions of prior knowledge into AI models and cost functions.
- Domain-specific representation of scientific data for science and engineering foundation models.
- Theory and methods to predict transfer learning/multi-task learning success.
- Theory of robust AI that can quantify expectations of composability, operational envelopes, out-of-domain performance, etc.
- V&V for the AI/ML pipeline.

12.5.3 GOALS FOR 5–10+ YEARS

Goals include creating a common representation of scientific data for science and engineering foundation model.

- Framework for end-to-end self-guided AI for autonomous discovery/control that can adaptively tune itself to fit different high-level design requirements and resource constraints.

12.6 References

[1] Pion-Tonachini, L., et al., 2021. Learning from learning machines: A new generation of AI technology to meet the needs of science. *Technical Report, Preprint*: arXiv:2111.13786.

[2] Coveney, P.V., Dougherty, E.R., Highfield, R.R., 2016. Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* 374, 2080, 20160153.

[3] Lee, E.A., and Sirjani, M., 2018. What good are models? *International Conference on Formal Aspects of Component Software*. Springer, Cham.

[4] Niederer, S.A., Sacks, M.S., Girolami, M., Willcox, K., 2021. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science*, 1(5), 313–320.

[5] Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J., 2021. Unsolved problems in ml safety. *arXiv preprint,* arXiv:2109.13916.

[6] Dulac-Arnold, G., et al., 2021. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9), pp. 2419–2468.

[7] LeCun, Y., 2022. *A Path Towards Autonomous Machine Intelligence*. Version 0.9, 2, 2022-06-27.

[8] Burt, D.R., Ober, S.W., Garriga-Alonso, A., van der Wilk, M., 2020. Understanding variational inference in function-space. *arXiv preprint,* arXiv:2011.09421.

[9] Cao, S., 2021. Choose a transformer: Fourier or galerkin. *Advances in Neural Information Processing Systems*, 34, pp. 24924–24940.

[10] Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., Catanzaro, B., 2021. Efficient token mixing for transformers via adaptive Fourier neural operators. In: *International Conference on Learning Representations*.

[11] Broderick, T., et al., 2021. Toward a taxonomy of trust for probabilistic machine learning. *arXiv preprint,* arXiv:2112.03270.

[12] Gal, Y., et al. 2022. Bayesian uncertainty quantification for machine-learned models in physics. *Nature Reviews Physics,* 4(9), pp. 573–577.

[13] Abdar, M., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, pp. 243–297.

[14] Zhang, S., et al., 2021. Bayesian attention belief networks. *International Conference on Machine Learning*, PMLR.

[15] Bommasani, R., Hudson, D.A., Adeli, E. et al., 2021. On the opportunities and risks of foundation models. *arXiv preprint,* arXiv:2108.07258.

[16] Chen, Z., Liu, Y., and Sun, H., 2021. Physics-informed learning of governing equations from scarce data. *Nature communications,* 12(1), pp. 1–13.

[17] Zhuang, F., et al., 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE,* 109(1), pp. 43–76.

[18] Cuomo, S., Di Cola, V.S., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F., 2022. Scientific machine learning through physics-informed neural networks: Where we are and What's next?, *arXiv preprint,* arXiv:2201.*05624*.

[19] Karniadakis, G.E., et al., 2021. Physics-informed machine learning. *Nature Reviews Physics,* 3(6), pp. 422–440.

[20] Chakraborty, S., 2021. Transfer learning based multi-fidelity physics informed deep neural network. *Journal of Computational Physics*, *426*, 109942.

[21] Brown, T., et al., 2020. Language models are few-shot learners. *Advances in neural information processing systems,* 33, pp. 1877–1901.

[22] Wang, S., Tuor, T., Salonidis, T., Leung, K.K., Makaya, C., He, T., Chan, K., 2018. When edge meets learning: Adaptive control for resource-constrained distributed machine learning. In: *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, April, pp. 63–71.

[23] Kairouz, P., McMahan, H.B., Avent, B. et al., 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2), pp. 1–210.

[24] Speed, A., and Stracuzzi, D.J., 2020. *Research Needs for Trusted Analytics in National Security Settings*. United States.

13. AI WORKFLOWS (EDGE, CENTER, CLOUD)

Any substantial artificial intelligence (AI)-enabled application requires many distinct interconnected components, including software and systems to collect, process, and prepare data needed to train AI models; and processes to update those models and make them available to operate on diverse platforms, from HPC systems and edge devices [1]. AI “workflows” comprise many different programs on multiple computing platforms: not just “AI” programs but also computational simulations; data discovery, preparation, and curation systems; and others. Effectively harnessing the advances outlined in Section 01—from surrogate and foundation models to inverse design or automated laboratories—will not only entail much larger and diverse data flows and sources relative to traditional modeling and simulation, but will also introduce opportunities for the use of AI to optimize, automate, and accelerate the workflows themselves. To develop and train such workflows effectively will require the creation of digital twins (discussed in Chapter 04) of the workflows and the underlying scientific infrastructure—which is itself a complex engineered system—with AI-based control systems to design, optimize, and operate end-to-end scientific experiments, and innovations in workflow system software.

AI workflows facilitate monitoring and control of experimental apparatus (computational and observational), software (including AI models), and data sources and flows. A particular scientific “campaign” will involve many iterations and experiments using these resources in various combinations and sequences. The realization of an AI-enabled campaign thus typically involves a collection of *workflows*, each responsible for the orchestration of elements of the campaign’s data and control flow—engaging a variety of computers, storage systems, scientific instruments, and other devices, from the edge to the exascale. A workflow supporting such a campaign is an encoding of the scientific method and may ultimately be instantiated in an AI foundation model (Chapter 02). A workflow may also be considered the broad realization of “programming in the large.” It provides the connective tissue to coordinate and manage computing and data resources and is the integrative glue for the software infrastructure. AI systems for discovery will require novel ways to compose workflows, capabilities to coordinate computational and data resources, and software services, with the ability to support new and diverse components, such as post-exascale system architectures or quantum computing systems. These novel workflows can in turn enable new breakthroughs by automating the lifecycle of AI-driven discovery. In this chapter, we identify the requirements, capabilities, and challenges as well as a conceptualization of a path to accelerate development.

Workflows and AI are inextricably linked. Workflows directly enable AI campaigns in their execution during the inference phase (using trained models), but they are also critical in setting up training phases to develop AI models and to collect the raw material (such as programming language traces) to instantiate the AI harnesses needed to develop new models. Workflows can include crucial functions such as those necessary to determine when AI models drift outside of their trained regime and need to be retrained, including generating or collecting training data on demand in active learning. Concurrently, the AI models themselves can be used to optimize future workflows. In addition to fine-grain resource tuning, an AI-enabled workflow can include models that evaluate and inform coarse-grained resource allocation and job placement, determining an appropriate mix of edge, high-performance computing (HPC) center, and cloud resources to complete a federated science campaign. These components, and workflows, may in turn use AI models to be automated, or “self-driving,” and eventually autonomous [2] as detailed in Chapter 05. Moreover, AI models that orchestrate workflows and learn failure patterns will enable the workflow to be self-adapting and self-healing, providing resilience to changing conditions in both the computing and communication systems and in the science domain.

Critical research directions in AI workflows from edge to HPC center to cloud are discussed below. We expand on why

PROJECT SPOTLIGHT

Project Name: Autonomous workflow for single crystal neutron diffraction

PI: Junqi Yin

Organizations Involved: Oak Ridge National Laboratory, National Center for Computational Sciences, Computer Science and Mathematics and Neutron Scattering Divisions

Goal: Create an AI-based autonomous workflow at the SNS DEMAND instrument for single-crystal neutron diffraction studies.

Significant Accomplishment: Combines an edge-inference capability with continuous integration to update AI models on the Summit supercomputer and present them in a user dashboard to control the workflow.

In the News: Junqi Yin, J., Zhang, G., Cao, H., Dash, S., Chakoumakos, B. C., Wang, F., 2022, *Toward an autonomous workflow for single crystal neutron diffraction*, presented at the Smoky Mountains Computational Sciences and Engineering Conference, Kingsport, TN, August 23–25.

AI-enabled and AI-driven workflows are important for the U.S. Department of Energy (DOE) mission, what is needed to bring developments in this field into full realization, and why this is the ideal time to accelerate the work.

13.1 Advanced Research Directions in AI Workflows

The science of workflows—enabling applications with functions distributed among multiple networked resources—has been an active research topic for over three decades. The maturity of workflow systems provides insight into how AI innovations can address emerging challenges and opportunities, such as those arising from unprecedented complexity and/or scale or those associated with new approaches to AI (Section 01 of this report).

13.1.1 ARD 1: ESTABLISH DIGITAL TWINS FOR DOE APPLICATIONS AND FACILITIES WORKFLOWS

Digital twins (Chapter 04) for workflows will enable AI models to be developed to represent, analyze, and optimize the operation of facilities and system workflows across the DOE complex. Digital twins comprise models for subsystems and their interactions within complex engineered systems—such as the power grid, an HPC center, an experimental instrument—or the resources making up a scientific workflow. Establishing digital twins as frameworks for workflow development will enable the design, testing, and adoption underpinning AI workflow systems with tools, methods, and policy parameters to connect facilities more efficiently.

13.1.2 ARD 2: INSTANTIATE AI SYSTEMS OF WORKFLOW CONTROLLERS

The potential to create foundation models (Chapter 02) trained by workflow execution data suggests the potential for general-purpose foundation models that can be used to create new workflows—that is, a master model that will provide control and optimization while using operational data from workflows as training data for continuous refinement. The workflow master model will include AI-reasoners (predictor, classifier, optimizer) for various categories of workflow campaigns, including optimization of workflows such as control, domain-dependent semantics, resilience to disruptions, and resource-constrained operations. Reasoners will also monitor science exploration, detect outlier results, classify new phenomena, and respond with appropriate actions such as launching new tasks for analysis.

13.1.3 ARD 3: DEVELOP AND DEPLOY AI BUILDING-BLOCKS AND WORKFLOW CODE GENERATORS

The DOE complex will need AI-driven cross-facility workflow code generators, leveraging the emerging approaches described in Chapter 06. We formulate in this activity an approach and methodology for science-based AI-driven code generators. These will establish specific activities such as data collection and reduction at a facility, time-dependent and data-dependent processing, and support for autonomous feedback loops. The scientific intent of a campaign is in this way realized in an instantiated workflow.

13.1.4 ARD 4: CAPTURE DOE COMPLEX-WIDE WORKFLOWS SYSTEM STATE

Data repositories for edge-to-center operations are critical to capturing the programming environment and runtime monitoring information of workflow data, and they allow expansion to science-driven domain-specific modalities and their influence on data. This is vital to improving digital twins and setting up training environments for automatic instrumentation as well as to the ability to gather information for programmatic (workflow-driven) control.

The broader workflows area of research depends on but also drives the realization of approaches described throughout Section 01. For instance, collected data will help inform how we might construct a workflow to train a surrogate model, create a foundation model, or adapt a workflow developed within one domain to be applied in a different domain.

13.1.5 ARD 5: INNOVATE TRUSTWORTHY WORKFLOW TECHNOLOGY FOR AI-ACCELERATED SCIENCE

Modern science campaigns are iterative, nonlinear ensembles of thousands of activities in a complex search space that cannot be realized without commensurate breakthroughs in workflow science itself. Workflows must self-describe, self-drive, self-adapt, and self-heal with minimal human effort, thereby providing dynamic initialization, execution, switching, and termination of tasks in support of active, continual, reinforcement, and foundational learning. Workflows of the future will drive the multimodal exploration of a problem space, federating foundation models, surrogate models, computational models, physical experiments, and observational data across multiple sites. Perhaps most importantly, workflows must enable trust in their outcomes by validating models, flagging uncertain results, and retraining models before potential errors are propagated downstream.

13.2 Why Is It Important?

Workflows and workflow frameworks capture optimized practices for creating, executing, and optimizing scientific experiments, enabling campaigns involving many experiments. Without explicit workflow support, these practices manifest as bespoke systems for individual scientific teams. Thus, absent an intentional, comprehensive workflow development initiative, DOE's investments in the application of AI systems will involve many redundant efforts producing an inefficient collection of custom software and tools to support the immediate needs of each individual science campaign. The availability of systematic, domain-agnostic workflows that are easy to deploy will give scientists a straightforward path to designing experiments and executing campaigns, accelerating and effectively reinventing DOE science and engineering practices using AI tools and techniques.

Building new, world-leading AI systems entails more than simply training a single AI model. Every innovation outlined in Section 01 of this report and every application detailed in Section 02 require a broad range of tasks, from data acquisition, aggregation, and curation; to model design, development, hyperparameter studies, large-scale training and validation studies; and model comparisons, deployment, and continuous learning. For the many campaigns that require observational data from user facilities, field laboratories, and other instruments, AI workflows also include integrating the AI models running in edge systems (e.g., providing in situ data analysis and real-time control)—potentially involving the orchestration of hundreds or thousands of such components and their data flows.

Many and varied needs are converging on not only optimizing workflows for AI but also using AI to optimize the workflows themselves:

- Domain communities must compose dynamically updated AI models for systems control and surrogate model development.
- Workflows are needed to couple multiple spatial and temporal scales, from real-time control of observational sensors (beamlines, scopes, radars, etc.) in experiments, to scheduling computing facilities, to responding to disruptions and load demands in nationwide energy grids.
- AI is making it possible—indeed, necessary —to mix and match hybrid models that require auto-selecting control set-points and appropriate surrogates, forward simulations, data proxies, hyperparameter optimizers, and so on. Campaigns will be optimized on the fly, requiring a deep understanding and improvement of the state of the art in workflows.
- For data protection in edge-to-multi-exascale campaigns, federated learning models are required, and these will be implemented as workflows in and among protected and sensitive data facilities.
- In many instances, scientific communities will collaborate on building, training, and using foundation models (Chapter 02), which will require workflows for training from diverse, multimodal data sources, with tasks ranging from data provenance to evaluation and training for specific downstream tasks (see Chapter 19: Data Infrastructure and the concept of *active collective memory*).

There are significant data and model management challenges for workflows [3] as they orchestrate the data, system, middleware, and applications, functioning in a real sense as the operating system of a set of related complex, distributed AI functions and resources. These functions of AI workflows can be illustrated by considering a sample of the capabilities described in Section 01 for the new AI approaches, requiring workflow technology that can:

Create and deploy surrogate models. Training and incorporating surrogates [4] in forward simulations or data—integral to digital twins—requires effective workflows to incorporate multimodal training data and AI models for inference, updating them systematically with on-line training.

Deliver foundation models and move toward general applicability. To create and refine foundation models that are generally applicable across domains will require adapting the end-to-end workflow with training data from a diversity of experiments within a particular domain or set of domains. Here, general-purpose workflow frameworks are essential to support the equally diverse community of scientists and teams collaborating to build and use shared foundation models.

Address questions of inverse design. Inverse design models capture data representing prior experience to improve system processes and rules–effectively playing a generative workflow in reverse. This is an unexplored area with transformative potential detailed in Chapter 03.

Design, engineer, and execute complex experiments and manage complex engineered systems. Timescales of workflow control and execution may vary from minutes to weeks. Operating with diverse data types and modalities, geographically distributed facilities from edge to center require AI models that are continuously trained by data from experiment iterations incorporating deep "understanding" of the dynamics of various classes of experiments as necessary to optimize, respond to disruptions, and ultimately make structural and procedural improvements to the workflow.

Develop autonomous laboratories. The use of AI to automate laboratory experiment campaigns will entail AI workflows involving not only traditional components such as data collection, analysis, and operation but also the operation of traditionally stand-alone laboratory equipment. This will require incorporating new application programming interfaces and even instrument operating systems into the control, monitoring, and adaptation workflow functions.

Create AI for (and through) programming automation. Creating AI models that can assist with, or carry out, software engineering and programming tasks will also require workflows that manage and prepare training data. The critical nature of software throughout every layer of infrastructure and experiment also underscores the importance of workflow functions that evaluate correctness, robustness, and security vulnerabilities.

13.3 Why Can't It Be Realized Now?

Workflows are currently static in definition. We do not yet have the methodology to respond to changing science needs, the AI models to inform the workflow, or mechanisms to harness data about experiments for training the AI model responsible for optimizing the workflow. These building blocks, in turn, are needed to fully realize the potential for AI models that can design new workflows, predict or detect flaws, and optimize workflows over time. Consequently, the use of AI in workflows would require individual domains to create their own bespoke systems, resulting in a number of challenges, including:

- Surrogate generation and construction are tied closely to the particulars of the domain and need deeper generalization research to apply to other domains.
- Complex systems are diverse in resource use and heterogeneity; their control through workflows driven by AI needs a further systematized definition of workflows and their operations.
- Foundation models are only beginning to be used for particular functional tasks; how we might have foundation models support multiple functional domains is an unsolved problem.
- Large-scale AI for workflows is a field hampered by the manner in which traditional AI currently operates for simulation campaigns and well-known datasets. Edge-to-center workflows will need a training-to-inference loop, and this is still an area of active research.
- Autonomous systems and their control workflows are unable to generate training data because most systems—both software and hardware—are not sufficiently instrumented.
- Code generation with AI is in its infancy, and development of programs for heterogeneous resource-constrained platforms is as yet a nascent area of research.

13.4 Why Is It Reasonable to Start Now?

AI models outperform humans on tasks that range from the mundane to those that were only recently regarded as uniquely achievable by humans—such as on-the-fly language translation or prompt-based image generation. The stunning pace of these advances surprises even researchers familiar with the underlying mathematics and recent history. Applying these principles to scientific systems—which differ significantly from natural language processing—could enable DOE to create AI models with the goal of outperforming humans in efficiently programming supercomputers, analyzing results, and even in formulating promising scientific hypotheses, thereby automating workflows for a significant fraction of the process of computational scientific discovery. These advances, and others described in Sections 01 and 02 of this report, are decadal in nature.

The AI models described throughout this report will require extensive computational resources for training and execution, with the potential for inverse design capabilities that could themselves be used to propose both improvements in resource use and new designs for resources—from instruments to supercomputers. These designs could, in turn, drive AI-enabled automated design and manufacturing to orchestrate their construction, operation, and use.

For the scale and uniqueness of DOE mission areas, the realization of these advances will require building infrastructure to support the embedding of AI in workflow systems, incorporating performance and results data to continuously self-train, and advancing workflow technology to enable further breakthroughs in the use of AI for DOE mission areas. Ultimately, an AI model for adapting, or creating new, workflows for a scientific campaign will incorporate the coding of its subsystems as well as execution provenance and workflow descriptions. We may imagine an AI-driven workflow controlling scientific exploration of a single problem on a single supercomputer. This trajectory might begin with a scientific seed prompt (SSP), from which AI identifies three component classes to satisfy the prompt: (1) the known—parts of the problem that have been solved and can be reused, (2) the discoverable—the parts of the problem known to be solvable but for which the answer must be sought by generating code and running it on a supercomputer, and (3) the unknown—the parts of the problem that cannot currently be solved and in the immediate term require human intuition. As much as possible, AI would operate this workflow independently and generate discovery artifacts for review by scientists. Human-machine collaboration will be required to explore the unknown and advance beyond it.

We are at the start of a decade in which we find workflows proliferating across the DOE complex, while at the same time pockets of AI-driven work are appearing at specific steps within these workflows. We must create research and development activities that connect workflows and AI (Figure 13-1). The workflows community has matured and is converging on an action plan [5, 6, 7, 8]. Integrated research infrastructure (IRI) needs across DOE's Advanced Scientific Computing Research (ASCR) community are driving the expanded deployment of workflows. The need for AI-driving and AI-driven interfaces, encapsulators, and descriptors to be composed flexibly—to be tracked and trained for prediction and altering of campaign trajectories—will grow significantly. It is most fruitful to start now to guide the tools and technologies as they emerge.

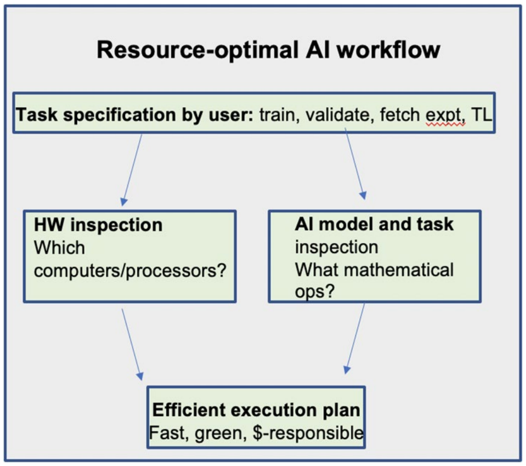

Figure13-1. Resource-optimal AI workflow.

13.5 What Is Needed to Start Now?

Workflows will be the glue to connect facilities and systems across the DOE complex. These workflows will be operated by large-scale AI foundation models that will be continuously trained by data from the execution and results of workflows. This transformation will require the following immediate steps.

- Establish digital twins and virtual environments for a set of several specific scientific workflows that are characteristic of experiments already operating in distributed fashion today. The digital twin will emulate the various parts of each pilot system, including both edge-to-center and cross-facility campaigns. The objective of these pilots will be to create an initial set of AI models that execute and learn from experiment iterations, and in turn can be evaluated for use in other campaigns with similar characteristics. In addition, it will drive the development of descriptors and operators specifying operations and characteristics of the workflow's constituent computing and experimental as well as the operations associated with data sources, flows, and curation. This will enable experiments to evaluate the mechanisms for defining workflows, including languages used, methods for expressing science goals, performance evaluation methods, and composition frameworks to identify and capture opportunities for autonomy.
- Design and deploy the first AI-based workflow controllers for DOE facilities. This effort will create several initial foundation models and associated systems for various categories of workflow campaigns. This would include optimization of workflows to include control, domain-dependent semantics, and resource-constrained operations.
- Research and develop next-generation workflow software systems capable of dynamic control and the dynamic data services needed to support data generation, model generation, model training, inference, and analysis with a maximum of autonomy and resilience.
- Instantiate a data repository of edge-to-center operations to capture programming environments and runtime monitoring information for workflow data and allow expansion to science-driven domain-specific modalities and their influence on data (Figure 13-2). This will address data collection, curation, and generalization challenges and shed light on the closed-loop need for datasets for AI, which will in turn modify and control workflows.

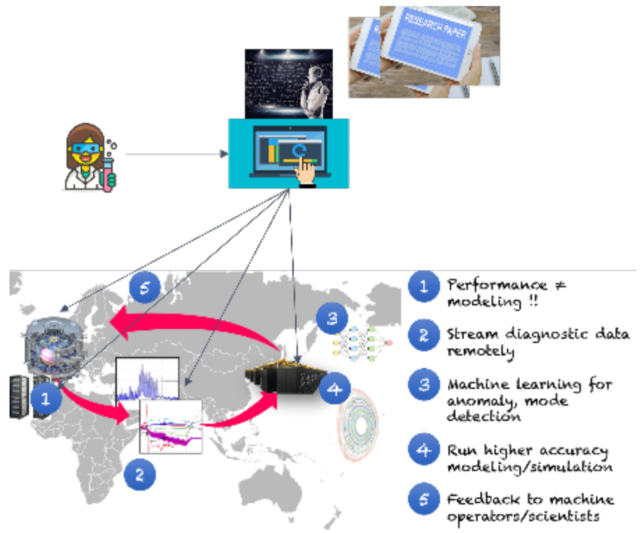

Figure13-2. Edge-to-center operation (image courtesy of R. Churchill et al. 2021 [9]).

These initial experimental digital twins of intra-facility workflows (e.g., to create an associated foundation model) and inter-facility workflows connecting DOE facilities will help create the smart workflow systems required for the DOE community to achieve the potential breakthroughs fueled by the approaches described in Section 01. The resulting workflow systems will free the scientist from committing to hard choices early in the campaign (guiding the campaign with data/compute/surrogate choices) and create the capability to traverse resource-limited and sensitive (e.g., national security) environments (restricted data, edges, low-power).

The advances in workflow definition methodologies will also be essential to accelerating progress in AI-generated code (Chapter 06).

In five to ten years, we will need to establish ways for data from workflows in the field to be collected in a repository to feed AI models. Foundation models will be able to operate in a test environment to explore their applicability. We will move to a deeper specification of complex workflows that can be explored and analyzed to bridge data-driven insights with the physics-driven observations (Figure 13-3). This richer understanding will allow new developments in the domain and in AI that both enable workflows to adapt to emerging needs (dynamically as well as in their design) and allow them to become a seamless part of the scientific discovery process.

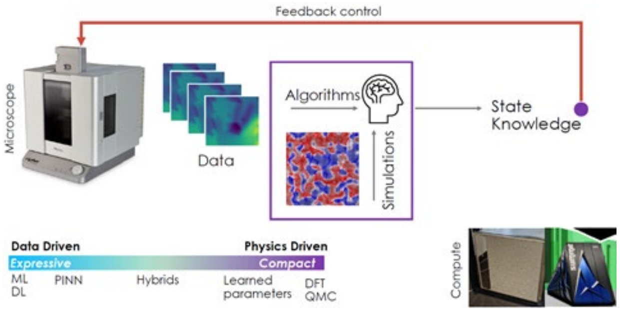

Figure 13-3. Bridging physics principles and observations with workflows. Image courtesy of Rama Vasudevan [10].

13.6 References

[1] Beckman, P., Sankaran, R., Catlett, C., Ferrier, N., Jacob, R., and Papka, M., 2016. Waggle: An open sensor platform for edge computing. *2016 IEEE Sensors*, pp. 1–3. https://doi.org/10.1109/ICSENS.2016.7808975

[2] LeCun, Y., 2022. A path towards autonomous machine intelligence. https://openreview.net/pdf?id=BZ5a1r-kVsf, accessed May 12, 2023.

[3] Ali, A., Sharma, H., Kettimuthu, R., Kenesei, P., Trujillo, D., Miceli, A., Foster, I., Coffee, R., Thayer, J., and Liu, Z., 2022. fairDMS: Rapid model training by data and model reuse (preprint). https://doi.org/10.48550/arXiv.2204.09805

[4] Yin, J., Wang, F., and Shankar, M.,2022. Strategies for integrating deep learning surrogate models with HPC simulation applications. United States: N. p., Web. doi:10.1109/IPDPSW55747.2022.00222

[5] da Silva, R.F., et al., 2021. A community roadmap for scientific workflows research and development. *2021 IEEE Workshop on Workflows in Support of Large-Scale Science (WORKS)*, pp. 81–90. https://doi.org/10.1109/WORKS54523.2021.00016

[6] National Academies of Sciences, Engineering, and Medicine, 2022. *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop*. https://doi.org/10.17226/26532

[7] U.S. Department of Energy (DOE), n.d. *Advanced Scientific Computing Research*. Office of Science. https://science.osti.gov/ascr, accessed May 12, 2023.

[8] Stevens, R., Taylor, V., Nichols, J., Maccabe, A.B., Yelick, K., and Brown, D., 2020. *AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science*. https://doi.org/10.2172/1604756

[9] Churchill, R.M., et al., 2021. A framework for international collaboration on ITER using large-scale data transfer to enable near-real-time analysis. *Fusion Science and Technology* 77(2), pp. 98–108, Feb. doi: 10.1080/15361055.2020.1851073

[10] Vasudevan, R., 2022. Machine learning for materials characterization and visualization, *2022 Gordon Research Conference on Computational Materials Science and Engineering*, October 6, Newry, Maine, USA. Zenodo. https://doi.org/10.5281/zenodo.7153303

14. DATA ECOSYSTEM

Fully realizing the potential of artificial intelligence (AI) for U.S. Department of Energy (DOE) missions requires mission-relevant data in forms and formats that can enable the next generation of AI systems. This is challenging, as DOE data are complex, combining simulations, observations, and experiments across a vast array of facilities, disciplines, and security requirements, and in many cases spanning decades of experiments, observations, and multiple generations of instruments. This heterogeneity of sources, disciplines, scales, and data types limits our ability to fully use DOE data for the development of present and future AI systems. Moreover, the high volume of data produced by DOE facilities is already too large to fully analyze. This dilemma will only intensify as future facilities come online, producing mission-relevant data that will be too large, too complex, and too fragmented to use effectively.

Consequently, there is urgent need for DOE to develop an AI-driven data ecosystem as a comprehensive solution for the many aspects of managing and using this critical data to fully exploit the potential of AI and drive advances in strategic areas of research and economic competitiveness. This data ecosystem must be structured around using AI to manage the complete lifecycle of data, comprising:

- A DOE complex-wide and accessible data universe with open standards, intelligent archiving, and built-in safeguards for security and privacy.
- AI-enabled data librarians that assimilate new data while identifying gaps in the completeness of that data.
- The use of AI for curating, annotating, and maintaining these data and their provenance to help ensure their longevity and usefulness.
- AI-enabled methods for data navigation, visualization, transport, integration, and delivery that enable these data to be easily utilized and leveraged.
- AI-enabled data search to find the data relevant to training or driving an AI model.
- Machine readable interfaces to enable automated access to interpretation of, and use of the data.

This comprehensive, AI-driven data ecosystem would have profound impact on the DOE as it is requisite for the development of AI systems harnessing any and every capability detailed in Section 01 of this report. Acquiring the ability to automatically manage and intelligently stage federated and distributed data will enable groundbreaking results in both scale and impact. The breakdown in barriers to data will democratize data and fully engage the DOE workforce, enabling a strong sense of mission and engagement. Novel methods for AI-driven maintenance, curation, and modernization of the data will drastically reduce data wrangling costs, thereby enabling a more efficient AI model development cycle. The use of a data-historian and AI-driven search will enable both legacy and currently generated data to be used and visualized more effectively. The DOE scientific community, empowered by an AI-driven data ecosystem, will make, keep, and find the right data at the right time for the right problem.

14.1 Advanced Research Directions in Data Ecosystem

Developing an AI-driven data ecosystem to manage critical scientific and economic data presents considerable scientific and technical challenges. The role of AI is fundamental to the success of this ecosystem due to the ability of AI models to perform tasks at a scale that is well beyond what manual data librarians can achieve. The information that must be managed has already scaled to exabytes in magnitude, creating the need for a holistic data strategy that enables capturing the potential that this data represents. The

PROJECT SPOTLIGHT

Project Name: ARtificial Intelligence-focused Architectures and Algorithms (ARIAA)

PIs: Roberto Gioiosa, Siva Rajamanickam, Tushar Krishna

Organizations Involved: Pacific Northwest National Laboratory, Sandia National Laboratories, Georgia Institute of Technology

Goal: Co-design algorithms, programming models, mini-applications, simulations tools, and hardware for machine learning and scientific simulations.

Significant Accomplishment: ARIAA has established collaborations with data flow hardware companies such as Cerebras, Graphcore, NextSilicon, and SambaNova and demonstrated Advanced Scientific Computing Research (ASCR)-focused mini-applications on these hardware systems through co-design of algorithms and programming models.

In the News: Publication: R. Garg, E. Qin, R. Gioiosa, S. Rajamanickam, T. Krishna, et al., 2022. *Understanding the design-space of sparse/dense multiphase GNN dataflows on spatial accelerators*. https://doi.org/10.2172/1821960. Best paper candidate: IEEE IPDPS.

Advanced Research Directions (ARDs) that follow focus on key areas necessary to create, optimize, and leverage such a data ecosystem.

14.1.1 ARD 1: THE DOE DATA ECOSYSTEM

The primary need is for a complex-wide data universe, built upon open-standard hardware and software, that brings the profusion of DOE data to the AI researcher's fingertips. Such an exabyte-scale data universe requires novel global data management and data infrastructure that can locate and deliver relevant data in usable forms for modern workflows. As detailed in Chapter 13, these workflows often couple multiple independent analysis codes, experiments, or simulation models. They are often distributed across multiple platforms from supercomputers to edge processors and must adhere to strict security and privacy concerns. These requirements greatly increase the complexity of creating this universe, which means that AI systems are necessary for optimizing the layout and management of the ecosystem. Simply put, the data ecosystem is a "complex engineered system" with the properties, and AI approaches, described in Chapter 04.

14.1.2 ARD 2: AI DATA LIBRARIANS THAT IDENTIFY GAPS AND COLLECT DATA

Given a comprehensive data universe, research is needed to develop an AI system that will locate existing data across multiple sources, assess its relevance for a given task, and in the process identify (and help to fill) any gaps in coverage. Prior to running an experiment, a DOE researcher should be able to know whether it has been carried out before and if the data exist already.

14.1.3 ARD 3: AI DATA LIBRARIANS THAT CURATE, MANAGE, AND ANNOTATE DATA

Data in the DOE ecosystem will also require annotation with metadata to enable rapid AI-driven searches. When data are incomplete, the AI librarian will generate complementing data and construct a data production workflow that integrates experiments and simulations through code composition. The data ecosystem must also support workflows that automatically and intelligently move data to where it is needed, for instance, from a user facility at one laboratory to a computational facility at another, and from there into a DOE data ecosystem storage cache that may be at a third location.

14.1.4 ARD 4: AI FOR DATA NAVIGATION AND MEANING

AI approaches such as those described in Section 01 will greatly facilitate the requirement that data are searchable and visualizable, thus bringing meaning/importance to the researcher. These data will have open standards, while safeguarding both proprietary and security concerns. Workflows will prioritize the ability to bring in wide varieties of data to develop a comprehensive view of the research direction. An AI-powered data navigator should automatically highlight important features and help the researcher make the most of the data.

These ARDs will require the use of every approach outlined in Section 01 and will, in turn, be critical to the scientific and engineering mission objectives laid out in Section 02.

14.2 Why Is It Important?

The lack of a comprehensive AI-driven data ecosystem is already weakening national competitiveness in several ways. Currently, high-performance computing (HPC) systems do not support the data usage patterns needed for AI at scale, for instance, to train surrogate (Chapter 01) or foundation (Chapter 02) models. This gap prevents basic capabilities such as controlling and optimizing complex systems (Chapter 04), or developing inverse design methods (Chapter 03), all of which require high volumes of multi-modal data for model training. Absent a data ecosystem as described above, our capacity to harness multiple datasets into a comprehensive and more accurate view of the problems of interest is limited.

The lack of a common data ecosystem infrastructure creates structural barriers that impede research. A unified ecosystem with a built-in transferability and portability will streamline the process of managing data and workflows. This ecosystem, to be built on commonly accepted standards, will enable code development (required for the application of AI methods to programming and software engineering, described in Chapter 06) and workflow. This ecosystem will, in turn, improve reusability, ultimately increasing efficiency and accelerating scientific discovery. Given the rich datasets that are currently managed by DOE, investments will be needed to ingest the wide variety of existing formats and thus make them more widely accessible to broader research communities.

The scale and velocity by which data are being generated compounds these structural problems. With the data ecosystem infrastructure described above, the data generated will be fed directly into surrogate models to evaluate and modify experiments, optimize the operation of user facilities, or even to improve the efficiency of an autonomous production plant. Directly coupling data creation and evaluation and use through AI models will provide new opportunities, as well as challenges that differ from the traditional approach of archiving data before its use. Supporting a move to this new paradigm will require new methods to facilitate tighter integration between the computational and experimental instruments that produce data, the AI models that perform analysis, and the AI-enabled data storage and management systems comprising the data ecosystem. Lacking these integrated capabilities, the AI-based tools and platforms that are currently being developed will be limited in scale and application, at best addressing narrow problems or those that are of limited relevance to DOE's science, energy, or security mission needs.

Data curation is the most resource-intensive component of AI research, requiring many experts, as selecting the right data requires considerable knowledge of science goals and AI techniques. For many research domains, wrangling data into useful forms can dominate the timeline of work required to create, optimize, and train an AI model. As detailed in the context of software development in Chapter 06, the use of AI systems for these labor-intensive tasks will not only improve productivity and reduce timelines, but will also reduce errors and ultimately result in models that can create new and more effective methods (as described in Chapter 02 with respect to foundation models).

The use of AI systems to resolve research or production questions more quickly will create agility in several areas. Often, experiments (whether with laboratory instruments or computational models) are duplicated or employ sub-optimal methods due to the inaccessibility of data and optimization insights from similar experiments. Realizing efficiencies in these areas will help remove barriers to accessing AI-based knowledge creation—barriers that DOE and its researchers face in rolling out technology developments to industry and other partners.

Below are examples of the benefits of a comprehensive data ecosystem.

Pervasive data collaboration and increased transparency. Projects with smaller databases need data improvement to use AI effectively. Activation barriers will be reduced by making more effective use of effort and expertise. Better utilization of archival experience across the DOE complex will enable DOE to parlay this expertise to yield benefits years after initially applied.

Low latency between data and decisions. Active learning, as required for nearly every AI approach discussed in Section 01, requires responsive, intelligent data sources. The amount, size, and rates of data will vary between challenges, making one-size-fits-all solutions unrealistic. The AI-enabled data ecosystem will enable the mixing of multiple sources and allow the level of effort required to be more easily recognized to realize a research goal. The mixing of data from many sources requires coherent interfaces for quality assurance. The data ecosystem must also track provenance and detect potential vulnerabilities such as the accidental inclusion of bad or intentional insertion of "poison" data [1]. The data management systems must acquire the ability to learn when and how to trust the data as a filter.

Reduced data wrangling for surrogate modeling. The ecosystem will significantly reduce the time it currently takes to prepare data for model training and improvement. Steps that need to be optimized include:

- Gathering the data needed across disparate sources.
- Transforming the data.
- Evaluating the data/building validation datasets.
- Connecting to active learning data.
- Dealing with data of different scales and modalities.
- Building persistent databases and data movement in surrogate computational infrastructure

Hardware and software capable of handling massive datasets. The infrastructure needed to handle massive datasets has several requirements. The data ecosystem will provide hardware and software that enables:

- Filesystems capable of massive and random-access reads.
- In-transit processing capabilities.
- Smart storage: that is, computing that is devoted to managing the data.
- Autonomous learning that makes the data subsystems work better.
- Anticipation of data needs for new applications.
- Intelligent search capabilities and automatic metadata inference.

Foundation models for DOE experimental facilities. A potential benefit of AI foundation models (detailed in Chapter 02) is the ability, once trained to critical mass, to capture, maintain, and preserve all of the experimental data from user facilities. Establishing this element eliminates the loss of usable information, in that it preserves the entire storage of generated experimental data at today's production rates. This effort will require:

- Development of good, heterogenous data production test beds that assimilate the results from multiple complementary experimental facilities and domains.
- Foundation models that operate on heterogenous and distributed computing and storage infrastructure.
- Comprehensive data policy that preserves privacy and security concerns.
- Resiliency and consistency for data storage across the facilities.
- The leveraging of industry innovations where possible.
- Reproducibility and validation capabilities.
- Clear definitions of what constitutes a self-supervision learning task for each modality of multi-modal data.
- Invariants between data fields—rules for physics constraints within multi-modal data.

14.3 Why Can't It Be Realized Now?

There are several barriers to developing this comprehensive data ecosystem. Much of this challenge arises due to the wide variety of DOE research and production efforts. This diversity creates several requirements that must be managed. Security considerations involving classified, proprietary, and scientific results further complicate data usage. These considerations and others have resulted in individual facilities developing their own methods of collecting, curating, and archiving data. DOE currently lags behind industry standards and best practices; however, with significant investments, the opportunity exists to leapfrog industry. One common concern with respect to promoting data access across multiple fields is that domain scientists are not accustomed to sharing data due to a lack of protocols and standardized tools. This state of the practice is complicated by current solutions, which are typically ad-hoc and domain specific.

Current repositories do not enable real-time access to data, do not capture the dynamic nature of data that changes over time, and cannot access data in different granularities. In addition, there is no current way to query data repositories efficiently to bring relevant data to the forefront that can help researchers with novel situations. Another major concern is data sparsity, where many surrogate models train on incomplete datasets. This is the case when there are only a few data points measured. The lack of access to large datasets is of particular concern to developing foundational models. Uptake of findable, accessible, interoperable, and reusable (FAIR) data standards has been slow, in large part due to the large investment in resources needed to modify current data and data collection pipelines [2–5].

Because metadata and data standards are not uniform across the DOE complex, an AI-enabled data ecosystem (if it existed today) would suffer from data sparsity. The current repository systems do not capture enough information to enable users to determine whether a dataset has been generated by experiments or is a synthetic dataset, with many lacking even basic information like authorship, origin, and data types and limitations. Solutions need to integrate a multilayer metadata approach to enable users to handle sparsity in training data.

14.4 Why Is It Reasonable to Start Now?

The need for developing a DOE-wide, comprehensive, AI-enabled data ecosystem has become acute. Across the DOE complex, there is a push for DOE assets (whether from microscopy to HPC assets or from neutrons/light sources to HPC assets) to become more interconnected [6, 7]. Expanding integration with other federal agencies and partners is also driving the need. The great strength of DOE is its ability to create large interdisciplinary teams and pair them with cutting edge infrastructure to solve problems that span the needs of the federal government. The comprehensive data ecosystem will greatly facilitate the ability of DOE to help our federal partners.

DOE expertise with HPC is one such asset. DOE computational facilities at the Office of Science (SC) and Office of Defense Programs (DP) are widely used in scientific and national security areas for addressing a wide range of problems. Their experience with high-end computing enables them to manage exascale and similar large data flows. By lowering the data barriers to helping our partners, we enable utilization of larger datasets from more sources. This expansion will provide unique solutions to the data needs, of which foundation models are an important example. These facilities offer performant HPC for achieving faster training, higher-accuracy models. This capability enables training with enough speed that these surrogate models can be used in ongoing simulations/experiments.

DOE is well poised to construct an AI-enabled data ecosystem through success in multiple avenues of data science. It already has preliminary success applying its expertise in areas of national impact. Following are areas where DOE is making impacts through the use of AI and machine learning (ML). All of these grand challenge areas require advances in this data ecosystem to realize success.

- Optimization of manufactured/synthesized material microstructure and properties [8].
- Nuclear deterrent systems that are survivable in radiation environments.
- Optimization of electrical grid operation under evolving demand environments. Grid storage field data are fed back to design efforts at DOE labs.
- National Institutes of Health (NIH) interactions, particularly in response to infectious diseases, such as Covid-19.
- Climate solutions such as carbon sequestration and better understanding of climate processes.

DOE's current involvement in these foundational areas is strong motivation for beginning immediately on developing an agile and robust AI-driven data ecosystem. This data ecosystem will greatly assist current DOE mission areas.

14.5 What Is Needed to Start Now?

Following is a roadmap of near-, medium-, and long-term goals that need to be realized to accomplish the objectives outlined in this chapter.

14.5.1 GOALS FOR 1–3 YEARS

Short-term goals include efforts to:

- Establish policy involving data standards and privacy, proprietary, and security concerns for both experimental and simulation data.

- Begin creating a database and training a data curation engine by leveraging existing standardized databases in select vanguard fields, such as astronomy and high-energy physics.
- Research low-latency data retrieval and movement systems, including advanced data compression algorithms.
- Begin building AI-based query models for scientific data that can locate and assess data relevance from new user queries on select, curated databases.
- Create a data validation framework and capabilities that can detect faulty and/or incomplete data and perform testing on curated databases.

14.5.2 GOALS FOR 3–5 YEARS

Medium-term goals include efforts to:

- Extend curated databases and engines to account for multimodal data, including published data.
- Develop automated pipelines and infrastructure for continual data imputation and augmentation.
- Develop interfaces for enabling scientific data search using context-aware natural language queries.
- Develop visualization tools for exploring the available data.
- Establish common data interchange formats.

14.5.3 GOALS FOR 5–10 YEARS

Longer-term goals include efforts to:

- Test and validate data retention policies as facilities upgrade.
- Couple data infrastructure and AI librarians to enable self-improving search models.
- Incorporate data across the DOE complex and across classification levels.
- Build configurable data preparation and augmentation pipelines, including a recommender model that can interrogate the available data for data selection/recommendation/classification/labeling [9].
- Complete the DOE data ecosystem, searchable via natural language queries.

14.6 References

[1] Baracaldo, N., Chen, B., Ludwig, H., and Safavi, J.A., 2017. Mitigating poisoning attacks on machine learning models: A data provenance-based approach. In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 103–110, November.

[2] Wilkinson; M.D., Dumontier, M., Aalbersberg, I.J., et al., 2016. The FAIR guiding principles for scientific data management and stewardship, *Scientific Data,* 3(1), 160018. DOI: 10.1038/SDATA.2016.18

[3] Dunning, A., De Smaele, M., and Böhmer, J., 2017. Are the FAIR data principles fair?, *International Journal of Digital Curation*, 12(2), pp. 177–195.

[4] Ravi, N., Chaturvedi, P., Huerta, E.A., Liu, Z., Chard, R., Scourtas, A., Schmidt, K.J., Chard, K., Blaiszik, B., and Foster, I., 2022. Fair principles for AI models, with a practical application for accelerated high energy diffraction microscopy, *arXiv preprint*, arXiv:2207.00611.

[5] Research Data Alliance (RDA). https://www.rd-alliance.org/, accessed October 5, 2022.

[6] Orr, L., Goel, K., Ré, C., 2022. Data management opportunities for foundation models. *12th Annual Conference on Innovative Data Systems Research (CIDR '22)*, January 9–12, Santa Cruz, Calif., USA.

[7] Thirumuruganathan, S., Tang, N., Ouzzani, M., Doan, A., 2020. Data curation with deep learning, In: *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT)*, March 30–April 2, Copenhagen, Denmark.

[8] Hiszpanski, A.M., Gallagher, B., Chellappan, K., Li, P., Liu, S., Kim, H., Han, J., Kailkhura, B., Buttler, D.J., and Han, T.-Y., 2020. Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge, *Journal of Chemical Information and Modeling,* 60(6), pp. 2876–2887. https://pubs.acs.org/doi/10.1021/acs.jcim.0c00199, accessed October 12, 2022.

[9] Yang, H., et al., 2019. Pipelines for procedural information extraction from scientific literature: Towards recipes using machine learning and data science, In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2., IEEE.

15. AI-ORIENTED HARDWARE ARCHITECTURES

Over the past several years, artificial intelligence (AI) has begun to show significant potential to enable a sea change in computational science and engineering, allowing scientists to address critical questions in national security, energy security, and leadership science with a level of agility and accuracy that will fundamentally change how we address risks in an uncertain world. These advances, detailed in Section 01 of this report, also require fundamental changes in the nature of scientific applications and workflows, both exploiting new hardware architectures and involving new forms of data flows and shifts in computational methods, such as the use of surrogate models. The scientific, energy, and national security challenges to which AI can make groundbreaking contributions are numerous. Fully realizing them—supporting these new forms of applications and workflows—will require revolutionary advancements in AI-oriented hardware architectures. These advances are driven by requirements spanning the new approaches described in Section 01 as well as the crosscutting areas detailed in Section 02.

The U.S. Department of Energy's (DOE's) approach to developing and deploying computational resources also must be revisited, particularly with respect to the "co-design" methodology. Deep co-design for the Exascale Computing Project (ECP), has resulted in breakthroughs in cluster-level and even node-level architecture, system software, workflow tools, and applications, but with limited influence over the central/graphical processing unit (CPU/GPU) design—designed and produced for consumer workloads. To fully harness new AI approaches (Section 01) to reinvent the broad and diverse scientific, energy, and security domains outlined in Section 02, the co-design process and timeframes of interactions must extend to the CPU/GPU design and beyond to encompass new materials and techniques necessary for future zeta-scale machines, which are to be constructed within rational constraints with respect to costs, especially power.

15.1 Advanced Research Directions in AI-Oriented Hardware Architectures

Each of the fundamental AI approaches described in Section 01 promises unprecedented advances across the entire suite of DOE scientific domains detailed in Section 02. These six AI approaches present unique challenges with respect to the hardware architectures underpinning those advances. Preliminary analysis of the requirements in each of these areas, for example a growing number of large-scale industry-driven AI models, suggests a need for *three orders of magnitude* improvement in computational efficiency over the next 15 years. This is driven by the need to support the magnitude of processing required for training of brain-scale neuro-symbolic models, such as surrogate (Chapter 01) and foundation (Chapter 02) models. These advances will only be realized through Advanced Research Directions (ARDs) targeting optimizations for unique DOE needs and major improvements in energy-efficient computing from the edge to the largest-scale high-performance computing (HPC) facilities.

15.1.1 ARD 1: ARCHITECTURES OPTIMIZED FOR DOE

DOE has unique hardware architecture needs that are driven by the complexity of our HPC and AI applications, which comprise massive multi-scale modeling and simulation and the integration and analysis of experimental data necessary for training of AI models for specific domains (many unique to DOE). Architectures that support these capabilities will push the limits of extreme heterogeneity, reconfigurability, and DOE-specific optimizations:

- True hardware reconfigurability, enabling frictionless composition of discrete components of the hardware;

PROJECT SPOTLIGHT

Project Name: Flexible neuromorphic computation in networks of superconducting oscillators

PI: Christoph Kirst and Co-PI: Dilip Vasudevan

Organizations Involved: Lawrence Berkeley National Laboratory and the University of California–San Francisco

Goal: Design and evaluation of superconducting oscillatory networks using collective dynamics principles of neural activity.

Significant Accomplishment: Designed hardware for superconducting oscillatory neural network with pixel error detection for image recognition and software for modeling the superconducting oscillatory computing.

In the News: One of the five teams selected for DOE neuromorphic computing funding (awarded for two years). R. Cheng, C. Kirst, and D. Vasudevan, 2022. *Superconducting-Oscillatory Neural Network with Pixel Error Detection for Image Recognition*, presented at Applied Superconductivity Conference (ASC 2022), Hawaii, October.

- Specialized micro architectures and components (chiplets, analog, non-von Neumann, compute in network/storage);
- Symbolic and probabilistic computing;
- Able to run in harsh environments (radiation and/or vacuum);
- Uncertainty quantification (UQ)-capable processing elements;
- Hardware-enabled trust.

15.1.2 ARD 2: ENERGY-EFFICIENT COMPUTING (EDGE TO HPC)

DOE must lead the nation in energy efficient HPC and edge computing. Current technology trends are realizing a slowing in energy efficiency that will cause our competitiveness in science, energy, and security to stagnate. This must be addressed through a focused set of research topics:

- Edge computing platforms that are as capable as today's multi-petaflop systems.
- An ability to dynamically control numerical precision, frequency, and resiliency for total dissipated power.
- Alternative hardware (analog, non-von Neumann).
- Massive increases in compute density within a fixed power budget.
- Differentiable computing hardware from the gate to component level.
- New materials and approaches to support ultra-low-power computation both in the aggregate (HPC systems) and at the edge (battery/solar powered devices).

Deep co-design of each of the DOE AI approaches alongside the ARDs will be necessary to meet the performance, scalability, resiliency, and reliability requirements they impose. To date, efforts and investments from industry have been driving rapid advancement in AI-oriented hardware targeted at a limited number of general-purpose use cases, such as recommender systems, speech and image recognition, and language translation. The financial benefits have driven industry to create more specialized AI-oriented hardware for these specific use cases (e.g., tensor cores, bfloat16 data format, low-precision/low-bandwidth data processing units [DPUs], etc.).

This hyper-optimization for divergent workloads in isolation will have limited benefit for DOE's diverse, unique set of science, energy, and security grand challenges. Advancements driven by strategic national priorities, such as many of those outlined in Section 02, will benefit from some facets of industry work, such as data management and workflow systems for training large-scale surrogate or foundation models, conceptual architectures for digital twins, or advanced transformers, but their application to DOE domain areas will require significant adaptation and, in some cases, a complete refactoring. Such efforts are under way in other countries, notably in China, where industry, academia, and government laboratories are inextricably connected. China's "New Generation Artificial Intelligence Development Plan" was established in November 2017 and coordinated by the Ministry of Science and Technology [1]. As of November 2022, China is the world's largest producer of super-computers, and Chinese supercomputers dominate the TOP500 rankings—hosting 160 systems, nearly twice the number of U.S. systems [2]. DOE's ECP program has kept the U.S. competitive for traditional simulation and modeling, but today must be augmented to pivot to new AI approaches, without which the United States will most certainly decline in national competitiveness [3]. China, Japan, and the European Union continue to make bold bets on AI, which represents an opportunity to erase, if not leapfrog, decades of U.S. leadership—largely reliant on the DOE complex.

15.2 Why Is It Important?

Here, we walk through the six new and emerging AI approaches detailed in Section 01 of this report, noting the hardware architecture demands unique to both the approach and the DOE target domain areas (Section 02).

AI and Surrogate Models for Scientific Computing (Chapter 01) require rapid inference using large, complex models [4][5][6] that are trained on data from instruments and simulations on a massive scale. UQ [7] and training robust models bring unique requirements that require hardware innovations to support multi-path AI training [8] in which data is labeled with probability distributions. and training is conducted across multiple discrete samples of the distribution. Large-scale training data with uncertainty distributions can be generated with microarchitecture advances that are transparent to the simulations used to generate them [9]. These active learning workloads will require frictionless composition of discrete components of the hardware, such as UQ-capable processing elements, neural network accelerators, and high-performance memories that are shared across these components, driving the need for memory-rich, chiplet-based architectures that are composable at runtime (Figure 15-1). These advances will enable massive scale UQ ensembles to be run in line with active learning workflows that train models on the distributions generated by these ensembles, ultimately to generate AI-based surrogates capable of achieving multiple orders of magnitude higher performance than traditional fine-grained modeling techniques.

AI Foundation Models for Scientific Knowledge Discovery, Integration, and Synthesis (Chapter 02) outlines some of the most demanding computational and data requirements in existence. Just a few years ago, the BERT [10] model, a forerunner of modern foundation models, was the largest sequence-to-sequence model in existence, with 110 million parameters. Google broke the one billion mark in 2016.

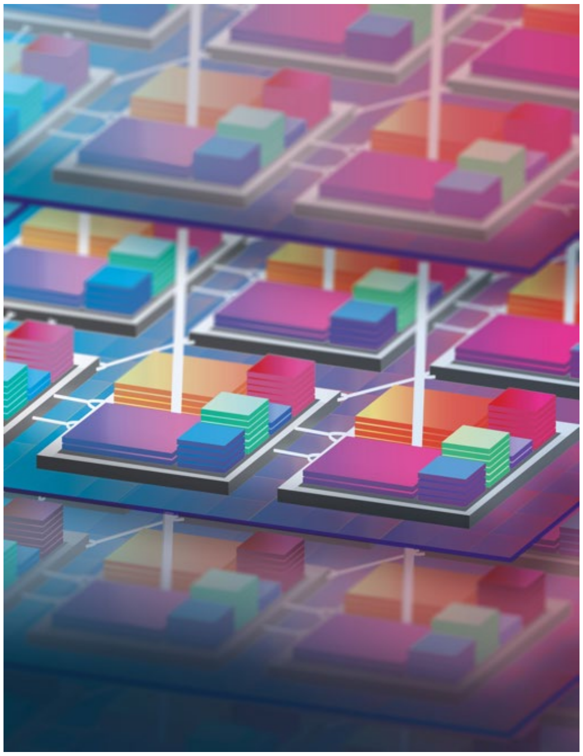

Figure 15-1. Conceptual view of an architecture optimized for DOE, composed of optically interconnected AI accelerators, HPC processors, and advanced memory technologies coupled via 3D organic and silicon integration.

Trillion-parameter models are now commonplace, and models with hundreds of trillion parameters are not unheard of [11]. As the complexity of these models and the data on which they operate continue to grow, so also does the need for radically new hardware architectures that go beyond simply increasing throughput and focus on accuracy and latency to meet the needs of active learning with timely feedback. Specifically, the differentiation between the self-supervised network core and the network periphery supporting task adaptation provides a unique opportunity for AI hardware architecture co-design.

Computational systems are complex, multi-layer, engineered systems, the complexity of which—illustrated by today's exascale machines--demands new design techniques. AI for surrogate or foundation models, applied to challenges such as AI for Advanced Property Inference and Inverse Design (Chapter 03), requires the ability to operate on massive datasets that tie structure/organization to desired properties. These data can span simulations, observations, experiments, publications, and more. Training for inverse design will require exabytes of simulation data coupled with many more exabytes of imaging or other experimental data. The data is often sparse, presenting unique requirements to efficiently manipulate these data structures at a level of performance and scalability that is relevant for large inference engines.

AI-Based Design, Prediction, and Control of Complex Engineered Systems (Chapter 04) drives the need for ultra-fast predictive control, allowing decision making that is anticipatory rather than reactive for everything from hypersonic vehicles to fusion reactors. Intelligent edge devices that can handle the massive data volume and velocity will require rethinking how sensors integrate with these devices. Real-time inference with quantified uncertainties will be required for these high-consequence scenarios. This will require such innovations as differentiable computing elements that enable global loss-function optimization that is orders of magnitude more efficient in space and power. Hardened and resilient computing architectures that can withstand harsh environments and can degrade gracefully over time while continuing to meet threshold performance limits are required. Constraints on the power envelope and operating environment will necessitate deep co-design of these processor-in-sensor devices.

AI and Robotics for Autonomous Discovery (Chapter 05) brings major challenges in pushing high-intensity computing capabilities deep into scientific instruments and facilities. The compact muon solenoid experiment at the large hadron collider, with 1 billion detector channels, will generate a petabyte per second of data that must be processed *in situ*. This will necessitate edge computing platforms that are as capable as today's multi-petaflop systems. Perhaps more challenging will be the need to run these systems in harsh environments (radiation and/or vacuum) and/or remote locations with limited communication capacity (e.g., climate or ecological observatories). Massive increases in compute density within a fixed power budget may necessitate entirely new process technologies and major advances in cooling and radiation hardening techniques. AI for autonomous discovery will require synchronization of DOE computing with data resources that span the DOE complex. This organization of compute and data resources will drive the co-design of AI hardware architectures that facilitate distributed workloads, which in turn incorporate modeling/simulation and AI training/inference, with data access and control systems at the edge.

AI for Programming and Software Engineering (Chapter 06) promises to fundamentally change how we approach computational science and engineering. With the potential to reduce multi-decade efforts in code development and validation to a few short months or even weeks, AI in programming and software engineering will allow us to answer questions of national importance in a truly agile way. The impact of such a capability cannot be overstated: It provides a means to gain an understanding of complex systems instead of being limited to often superficial levels of detail. Accomplishing this will require the ability to routinely train massive neuro-symbolic models that combine knowledge of algorithms, methods, programming languages, and architectures. These neuro-symbolic models will integrate symbolic languages for knowledge representation, neural networks for pattern recognition, and probabilistic inference to establish causal relationships between entities [12].

All of this will require several critical AI hardware advances. For the neural network component, hardware will need to scale to support AI models of up to a quadrillion parameters as well as training times in days rather than months—far beyond what is feasible today. The ability to dynamically control numerical precision, frequency scaling, and total dissipated power during training will be the path to achieving this goal. Coupling this with symbolic reasoning and probabilistic methods will require even more aggressive architecture changes [13].

15.3 Why Can't It Be Realized Now?

Today's AI hardware technologies—built from commodity materials and components, albeit assembled and operated through co-designed architecture efforts—are simply incapable of scaling to the level of throughput required for our most pressing challenges, particularly in foundation models, surrogates, complex systems, software development, and engineering. Each of these areas requires a function step change in the nation's ability to train on massive datasets and parameter spaces that dwarf what is possible in even one-off "hero calculations" today. Analysis of the requirements in each of these areas indicates a need for at least three orders of magnitude improvement in efficiency over the next 15 years. This is driven by the need to support routine training of brain-scale neuro-symbolic models with an agility and responsiveness 100 times more than the current state-of-the-art while achieving a 20-times improvement in energy efficiency. Currently, the largest model requires approximately 175 to 540 billion parameters to describe its state space and took the equivalent of half of the combined computing resources of the top ten supercomputers in the world for over a month.

Future systems must be capable of routinely training models with over a quadrillion parameters. Incrementally training models this large in real time will require hardware with adaptive resilience, differentiable resources, lattice-structures of memory and computation, and dynamic precision. Adaptive resilience will enable architects to place resiliency on control paths while relaxing it on data paths where there is more error tolerance (due to the inherent error of the underlying data). Dynamic precision will enable massive improvements in efficiency while improving model robustness. Differentiable hardware resources will enable global optimization among compute elements in loss function minimization, further improving performance and the ability to backtrack when errors are encountered that would otherwise result in ad hoc "solutions" to convergence that are currently based on trial and error rather than causal analysis. Lattice structures will enable major gains in efficiency and the ability to drive massive hybrid parallelism (model and data) that are out of reach on today's systems. These technology advances coupled with deep co-design are the path to meeting our grand challenges.

New hardware architectures will also be required to address daunting challenges in computing at the edge. This is particularly true for complex systems and autonomous discovery, where hyper-local decision making is often needed. Harsh and inaccessible environments will prevent the routine use of a remote HPC environment for active learning, inference, and control, necessitating a level of AI computational density that far outstrips what is available today. These requirements will drive further advances in critical dimension shrink, integration, and novel architectures to achieve power/performance requirements while improving the latency of response by several orders of magnitude. These architectures will need to be engineered for graceful degradation over time with predictable reliability and performance. In essence, we will need to achieve a "resilient petaflop and petabyte at the edge" over the next 15 years. Such a capabilities would catalyze an entirely new ecosystem of technologies to support the diversity of workloads from the largest scale HPC environments to embedded computing at the far edge.

15.4 Why Is It Reasonable to Start Now?

The ECP [14][15] has driven major technological advances in HPC and AI to meet the scientific and national security goals of the project. The Frontier supercomputer achieved the first sustained exaflop of double-precision floating point performance in the world. Beyond the remarkable power efficiency of this system for this level of performance, Frontier has delivered new networking, storage, and packaging technologies that will provide the foundation for an entire ecosystem of supercomputing technologies moving forward. The El Capitan supercomputer is driving major advances in processor manufacturing; the MI300A accelerated processing unit (APU) is "a 3D chiplet design with AMD CDNA3 GPUs, Zen 4 CPUs, cache memory and HBM[3] chiplets" [16]. The Aurora system is driving advances in 3D packaging (Foveros) and chiplet interconnects (EMIB). These processor and packaging advances have revolutionized how computing technologies are designed and built and will enable major breakthroughs in systems moving forward.

While major advances have clearly been achieved, some compromises had to be made along the way. The first generation of exascale systems are highly optimized for dense and uniform computations, which are important—but insufficient to meet our future goals. Advances in the ability to

[3] HBM = high-bandwidth memory.

handle sparsity due to an explosion of parameter space and highly divergent workloads that will be needed in neuro-symbolic AI will require new hardware technologies, some of which are beginning to emerge in prototype form.

Moore's law (technology shrink) has slowed significantly over the past two decades, and Dennard scaling (frequency increases) has halted entirely. This has driven scientists and engineers to tackle the continued demand for performance improvements in other innovative ways, such as 3D chip stacking and interconnecting of multiple chiplets. Wafer-scale manufacturing of tightly interconnected processing and memory fabrics is another approach that is being actively explored.

We expect these trends of tight integration of heterogeneous chiplets to continue and hybrid approaches to emerge. Advances in silicon photonics coupled with advanced packaging technology will result in highly connected discrete multi-package modules, each of which will be capable of many petaflops of performance and orders of magnitude higher memory bandwidth than today's most capable computing systems. This massive increase in connectivity between computing technologies will enable the first truly reconfigurable systems, in which complex workflows with divergent computing and data requirements will adapt the hardware on demand to their requirements. The implications of this change are profound, enabling supercomputers to be designed with a variety of processing and memory technologies that are individually optimized for particular components of a workflow but can then be assembled and operate in unison as if they were a single piece of silicon. Other innovations, such as cryogenic complementary metal oxide semiconductor (CMOS) design, which cools the entire system to around 77 K, present a tremendous power-saving potential for centralized computing facilities. Joint optimization of device and design will further enable total power savings of more than 30%, including the cooling power for the complete system [17].

Concurrent with these technological advances, there has been a tectonic shift in the economics of hardware design and manufacturing [19][20]. In the past, technologies providers and foundries were vertically integrated organizations. This is no longer the case. Technology providers are now separate from foundries, either in completely different companies or in different business units. Fabrication through these foundries is now much more accessible, particularly at larger process nodes, which has enabled an explosion of technology vendors pursuing highly optimized, near-specialized accelerators for extremely specific workloads. The availability of licensable and open intellectual property (IP) (ARM and RISC-V) is further lowering the barrier of entry for hardware designers. Custom processing technologies no longer cost $400 million for each design and tape-out cycle; customization can be done for as cheaply as a few million dollars today. As a result, over the past decade, we have witnessed over 140 fabless (i.e., without foundries) design companies emerge to take advantage of the confluence of changing economics in hardware design and the need for customized technologies to meet our grand challenges. These technologies include coarse-grained reconfigurable architectures, spatial streaming dataflow architectures, machine learning (ML) inference engines, ML training accelerators, graph analysis accelerators, processors in memory, and programmable network and storage devices. This diversity of computing technologies is currently the most likely path toward achieving the 2,000-times improvement in end-to-end efficiencies needed to meet our requirements.

While encouraging, many of these technologies remain quite immature and the marketplace is highly fractured. Architecture research, design, and fabrication still has a significant lead time, necessitating early engagement with commercial vendors, including not only system integrators, but also technology component vendors of processors, accelerators, memory storage systems, etc. The fledging marketplace also requires the timely development of standards for interoperability. DOE's involvement can ensure the neutrality and openness of the marketplace. The CHIPS and Science Act of 2022 will provide significant resources and support public–private partnerships to drive such interoperability standards [21].

Deep co-design [22] is needed to ensure that the most pressing science, energy, and security challenges are addressed by these component technologies. Technology maturation of these co-designed technologies will require enduring partnerships with scientists and engineers at the national laboratories. To meet our grand challenges, no single component technology is sufficient, necessitating a level of technology integration at a massive scale that no single organization can achieve. Similarly, the massive diversity of technologies, a veritable Cambrian explosion, will require deployments of multiple systems across major computing facilities within the DOE, where each system will focus on a set of technologies aligned with a broad, but potentially not exhaustive, set of grand challenge workloads.

The timeframe to achieve ECP advances and to transition from development to operation has been on the order of 7–10 years. The reinvention of DOE's modeling and simulation through adoption and development of AI approaches and the design and construction of the necessary AI system architectures will be a similarly long pipeline. With the ECP transitioning from development to operation, any delay risks stalling this pipeline.

15.5 What Is Needed to Start Now?

This project is a unique opportunity for DOE to lead the nation in developing new AI-aware hardware that addresses DOE and national missions. The time is ripe for deep engagement on component technologies (e.g., processors, memory, accelerators, network, storage) as the aforementioned market forces accelerate. Beginning immediately, DOE must invest in the following:

- Deep co-design activities that span fundamental technology design, from materials and processor architecture to algorithms and applications. While drawing upon prior experience in co-design before and throughout the ECP, this co-design process must encompass a much broader set of technologies and recognize our ability to shape technologies, up to and including the fundamental microarchitecture.
- Investment in fast-forward/path-forward activities driven by the national laboratories, which will require an ability to reason about and shape technologies, algorithms, and applications at a very deep level while preserving a high degree of productivity and agility.

Longer term, a roadmap for the developments in this area includes the following:

- AI driven hardware design and optimization to achieve 2,000-times improvement in efficiency (20-times power reduction and 100-times performance improvement).
- Exploration and evaluation of new edge device systems integrating edge-AI, sensing, and workflows for critical DOE mission environments, including experimental instruments/facilities and autonomy in complex systems, such as laboratories or vehicles.
- Tool building to enable computational scientists, applied mathematicians, computer scientists, and computer engineers to productively reason about and shape applications, algorithms, and architectures.
- Interdisciplinary centers that couple fast-forward/path forward-like activities with subject matter experts and advanced tools for application-, algorithm-, and architecture-based co-design.

15.6 References

[1] Roberts, H., Cowls, J., Morley, J., et al., 2021. The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society*, 36, pp. 59–77. https://doi.org/10.1007/s00146-020-00992-2

[2] Top500., n.d. List Statistics. *Statistics*. https://top500.org/statistics/list/, accessed May 12, 2023.

[3] Special Competitive Studies Project, 2022. *Mid-Decade Challenges to National Competitiveness*. https://www.scsp.ai/wp-content/uploads/2022/09/SCSP-Mid-Decade-Challenges-to-National-Competitiveness.pdf, accessed May 12, 2023.

[4] Wang, B., et al., 2020. Multi-physics-resolved digital twin of proton exchange membrane fuel cells with a data-driven surrogate model. *Energy and AI 1,* p. 100004. https://doi.org/10.1016/j.egyai.2020.100004

[5] Yin, J., Wang, F., and Shankar M., 2022. Strategies for integrating deep learning surrogate models with HPC simulation applications. In: *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 01–10. https://doi.org/10.1109/IPDPSW55747.2022.00222

[6] Blanchard, A.E., et al., 2021. Language models for the prediction of SARS-CoV-2 inhibitors. *International Journal of High Performance Computing Applications*, preprint. https://doi.org/10.1101/2021.12.10.471928

[7] Tripathy, R.K., and Bilionis, I., 2018. Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of Computational Physics,* 375, pp: 565–588. https://doi.org/10.1016/j.jcp.2018.08.036

[8] Tang, Y., et al., 2020. Uncertainty-aware score distribution learning for action quality assessment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp: 9836–9845. https://doi.org/10.1109/CVPR42600.2020.00986

[9] Tsoutsouras, V., et al., 2021. The LAPLACE microarchitecture for tracking data uncertainty and its implementation in a RISC-V processor. In: *Proceedings of 54th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 1254–1269. https://doi.org/10.1145/3466752.3480131

[10] Devlin, J., et al., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*: https://arxiv.org/abs/1810.04805v2

[11] Lian, X., et al., 2022. Persia: An open, hybrid system scaling deep learning-based recommenders up to 100 trillion parameters. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3288–3298. https://doi.org/10.1145/3534678.3539070

[12] Lake, B., Ullman, T., Tenenbaum, J., and Gershman, S., 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, E253. https://doi.org/10.1017/S0140525X16001837

[13] Rahimi, A., et al., 2017. High-dimensional computing as a nanoscalable paradigm. In: *IEEE Transactions on Circuits and Systems I: Regular Papers*, 64(9), pp. 2508–2521. https://doi.org/10.1109/TCSI.2017.2705051

[14] Messina, P., 2017. The exascale computing project. *Computing in Science & Engineering,* 19(3), pp. 63–67. https://doi.org/10.1109/MCSE.2017.57

[15] de Supinski, B.R., et al, 2022. *Non-Proprietary Companion to the Q1CY2021 Path Forward Final*

Assessment WBS 2.4.1, Milestone PM-HI-1040. https://doi.org/10.2172/1845203

[16] AMD, 2022. AMD Details Strategy to Drive Next Phase of Growth Across $300 Billion Market for High-Performance and Adaptive Computing Solutions (news release), June 9. https://ir.amd.com/news-events/press-releases/detail/1078/amd-details-strategy-to-drive-next-phase-of-growth-across.

[17] Chiang, H.L., et al., 2020. Cold CMOS as a power-performance-reliability booster for advanced FinFETs. In: *Proceedings of 2020 IEEE Symposium on VLSI Technology*, pp. 1–2. https://doi.org/10.1109/VLSITechnology18217.2020.9265065

[18] Saligram, R., et al., 2021. Power performance analysis of digital standard cells for 28 nm bulk CMOS at cryogenic temperature using BSIM models. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits,* Vol. 7, No. 2, pp. 193–200. https://doi.org/10.1109/JXCDC.2021.3131100

[19] Murray, C., et al., 2018. *Basic Research Needs for Microelectronics: Report of the Office of Science Workshop on Basic Research Needs for Microelectronics*, October 23–25. https://doi.org/10.2172/1616249

[20] Ang, J.A., Chien, A.A., Hammond, S.D., et al., 2021. *Reimagining Co-design for Advanced Scientific Computing: Report for the ASCR Workshop on Reimagining Co-design*. https://doi.org/10.2172/1822199

[21] The White House, 2022. Fact Sheet: CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China, Statements and Releases, August 9. https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china/, accessed May 12, 2023.

[22] Vetter, J.S., et al., 2018. *Extreme Heterogeneity 2018—Productive Computational Science in the Era of Extreme Heterogeneity: Report for DOE ASCR Workshop on Extreme Heterogeneity*. https://doi.org/10.2172/1473756

SECTION 04: INFRASTRUCTURE AND WORKFORCE REQUIREMENTS

Advancing and leveraging new AI capabilities, translating decades of investment and advancement of DOE's world-leadership in modeling, simulation, and infrastructure into world-leadership in AI-empowered science, energy, and security systems will require the DOE workforce, scale of operation, computational and data resources, and instrumentation to be similarly transformed to meet the challenges and achieve the vision captured in this report. We survey each of these areas in this section, noting the current state, the grand challenges, and the path forward to meeting those challenges.

16. WORKFORCE AND ETHICS

Progress in artificial intelligence (AI) for U.S. Department of Energy (DOE) mission science requires a growth in the workforce across the DOE, especially when considering what is needed with respect to the advanced research directions. Moreover, it is essential for DOE to lead in critical areas of ethics and safety, for instance developing and embracing principles such as accountability, which relates to AI researchers having a clear understanding of the liability involved with application of AI and potential unintended consequences. Concurrently, the impact of AI on workforce is multi-dimensional challenge. In this chapter we discuss the need for DOE workforce development as well as the importance of ethical considerations related to the use of AI.

16.1 Current State

Nationally, the demand for AI researchers and practitioners has grown rapidly. Achieving transformational artificial intelligence (AI) for U.S. Department of Energy (DOE) mission science requires not only robust AI methods but integration of AI research with advanced computational skills and methods in concert with domain-specific knowledge. As DOE expands its AI workforce, it is also essential to emphasize broadening participation among groups underrepresented in STEM fields and within DOE labs. In addition, DOE must stimulate and accelerate the development of AI expertise and experience within the existing DOE workforce through collaborations, training, and career development.

DOE has a data-rich environment often including major instruments. It possesses mature state-of-the-art computational models with access to the world's most advanced computers. Moreover, its existing scientific workforce is highly interdisciplinary and collaborative. These assets offer an attractive learning environment for new and existing staff to explore transformative AI for a range of DOE missions.

The DOE has a long and productive history of partnering with other agencies, universities, and industry to advance the nation's innovation leadership and stimulate technological breakthroughs. From fundamental to applied research, the DOE ecosystem offers a ripe environment for maturing the full complement of AI-related skills in individuals, teams, and institutions. These skills include but are not limited to computer and computational science, information science, statistical sciences and uncertainty quantification, applied mathematics, and theory of complex systems.

It is also important that the DOE workforce reflect U.S. demographics. When considering underrepresented communities in science, technology, engineering, and medicine (STEM, i.e., women, African American/Black, Hispanic/Latino, American Indian/Alaskan Native), the current DOE technical research staff consists of 20% female and 13% ethnic minoritized communities (African American/Black, Hispanic/Latino, American Indian/Alaskan Native) [1]. In contrast, the 2021 U.S. demographics indicate 32.2% ethnic minoritized communities and 51% female [2]. To provide the transformative science needed to ensure America's security and prosperity, it is critical to have a highly skilled DOE workforce that fully utilizes all the talent available in America.

Targeted alliances such as the Stewardship Science Academic Alliances (SSAA) Program, NNSA's Predictive Academic Alliance Program (PSAAP), Minority Serving Institution Partnership Program (MSIPP), and others will ensure that new opportunities to develop AI skills necessary to participate in DOE AI research and development is available to students.

Concurrently, the research, development, and application of new AI capabilities—including the adaptation of industry results where possible—requires new skills and experience that are in high demand not only within DOE but in industry. This competition for talent suggests that DOE must examine new models for collaboration with industry.

In addition to addressing the skills competition with industry, this report lays out several areas where AI models will perform some tasks that currently require “humans-in-the-loop,” whether in operating laboratory instruments, or in data management and curation, or even in software development. Such automation can have the effect of increasing the time for scientists to focus on creative and innovative tasks to advance the science resulting from the elimination of mundane tasks. Further, the job areas automated with, or assisted by, AI models are likely to have the greatest effect on entry-level jobs in the tech workforce, which require the least amount of experience or subject matter expertise. Consequently, DOE's workforce training efforts must accommodate this shift, including skills related to using new AI tools and frameworks, by developing new strategies for early-career staff, including engineers and scientists. The DOE will also need to work closely with partnering universities through their academic alliance programs (such as the Stewardship Science Academic Alliances (SSAA) Program, NNSA's Predictive Academic Alliance Program (PSAAP), Minority Serving Institution Partnership Program (MSIPP), etc.) to ensure that new hires have the AI skills necessary to perform these jobs.

With respect to the ethics of AI, the White House (Office of Science and Technology Policy) has released the “Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People." [15] as a starting point. However, with the release of large language models in the late 2022 and early 2023, it is clear that more work is needed. Here, the series of AI@DOE Roundtables held in late 2021 and early 2022 provide valuable guidance.

16.2 Grand Challenges

16.2.1 GROW AND FOSTER AI AT A DOE WORKFORCE THAT REFLECTS THE U.S. DEMOGRAPHICS

To achieve the AI that will foster transformative scientific breakthroughs for DOE science, energy, and security, it is critical to have the diversity of thought that comes from fully engaging a broad cross-section of the scientific workforce. For some communities, achieving representation that matches the U.S. demographics requires increases in unprecedented multiples of their current representations, that is, changes on a scale that could accurately be described as a grand challenge.

Ample evidence suggests that the benefits from a diverse workforce are broad and significant. A more diverse workforce does scientific research differently and does both different and more innovative scientific research.

- **Doing science differently:** Just as a black cosmologist who is also a jazz musician develops physics theories inspired by a black music tradition [3], a more diverse workforce can bring a diversity of approaches to using AI to accelerate advances in DOE science, energy, and security.

PROJECT SPOTLIGHT

Project Name: Sustainable Research Pathways (SRP)

PI: Mary Ann Leung

Organizations Involved: Sustainable Horizons Institute

Goal: Connect scientists at eight DOE labs with faculty and students from underrepresented groups at community colleges, four-year colleges, and doctoral degree-granting research institutions.

Significant Accomplishment: Since its 2015 inception at Lawrence Berkeley National Laboratory through its current eight-lab operation, SRP has fostered hundreds of new research collaborations between national lab scientists and faculty and students at a variety of institutions, including Historically Black Colleges and Universities (HBCUs) and Hispanic-Serving Institutions (HSIs).

In the News: *HPCWire* Workforce Diversity and Inclusion Award 2021, available at: https://www.hpcwire.com/off-the-wire/hpcwire-reveals-winners-of-the-2021-readers-and-editors-choice-awards-during-sc21/, accessed December 5, 2022.

- **Doing different science:** Just as a black research software engineer who has confronted stigma in other arenas might feel emboldened to embrace a widely stigmatized programming language [4], a more diverse workforce might be inspired to explore the benefits of novel language choices in DOE AI programming environments.
- **Inspiring innovation:** The application of machine learning to text analysis of the publications of a near-complete population of 1.2 million U.S. doctoral degree recipients between 1977 and 2015 demonstrated that the underrepresented groups that diversify organizations produce higher rates of scientific novelty [5]. Paradoxically, the same study showed that scientific researchers from underrepresented communities had less successful careers due to such factors as their work being taken up less by others.
- **Rethinking "entry level" skills:** The use of AI models for many rudimentary tasks will also change the nature of training and learning opportunities for early career individuals, both by introducing new workflows and tools and by raising the bar with respect to minimum job skills required and the content and extent of resources that will be required for effective training and onboarding.

16.2.2 DEMOCRATIZING AI FOR DOE SCIENCE

Developing a workforce ready to advance AI for the DOE mission requires broadly exposing and engaging both the existing computational and disciplinary workforce to AI as and future generations of the DOE workforce. This will entail democratization to consider the full educational ecosystem—K-12, two-year institutions, higher education, graduate programs, and alternative paths (e.g., code camps and other mechanisms for retraining). The DOE complex comprises 17 national laboratories, all of which have developed resources for the various components of the educational ecosystem [6]. To provide the training needed for AI for DOE mission areas, it is important that the associated research be made available in appropriate ways for the different levels of the educational ecosystem, with special attention given to significantly engaging students from underrepresented communities and the shifting of entry-level skills toward more advanced AI, science, and engineering requirements.

At the same time, the competition for talent in science, technology, and mathematics will only increase and DOE must increasingly focus on attracting and retaining early career individuals, particularly in AI, computer and computational sciences, and mathematics.

16.2.3 ETHICS OF AI SYSTEMS AND APPLICATIONS

The DOE AI Roundtable events in late 2021 and early 2022 included ethics discussions throughout many breakout sessions. These discussions identified the need for an advisory framework within DOE to “help guide and address AI R&D ethical questions, advise on concerns, and maintain awareness of social and technological challenges” [16]. The concerns arising from widespread adoption of large language model-based applications (e.g., OpenAI ChatGPT, Google Bard, Microsoft Bing) in early 2023 underscores and indeed increases the urgency of this recommendation. The need for developing effective ethics and safety guidelines and guard-rails underscores the urgency of advancing fundamental research understanding complex AI systems—as discussed in Chapter 12: Mathematics and Foundations and throughout Section 03 of this report.

16.3 Path Forward

16.3.1 DIVERSE LAB WORKFORCE FOR AI AT DOE: NEAR TERM

Success will require normalizing inclusion so that underrepresented communities become an integral part of the scientific enterprise from the initial spark of an idea to that heady moment when an experiment actually works. To address and retain a scientific workforce that is representative of the U.S. demographics, it is important to address issues related to recruiting (i.e., expanding the networks and partnerships leveraged to seek candidates), the hiring decision process (i.e., are inclusive factors considered in hiring decisions), and the need for an inclusive environment where all voices are valued and considered with respect to promotion and advancement.

To address the recruiting issue, we need to bridge the gap between multiple sectors, bringing researchers into close collaboration across institutional, geographic, and cultural divides. We need to expand the professional networks of AI at DOE researchers by supporting collaborations across institutions in ways that differ from long-established patterns [7–11]. It is important to establish long-term connections with minority-serving institutions, workforce development organizations such as the Graduate Degrees for Minorities in Engineering and Science (GEM) consortium [9] and Sustainable Horizons Institute [12], which organizes the Sustainable Research Pathways (SRP) program (see Figure 16-1), and as well as to have visibility at diversity conferences. The existing DOE diversity programs, such as Minority Serving Institutions Internship Program (MSIIP) and Minority Serving Institution Partnership Program (MSIPP), need to be strengthened and grown to promote hiring of diverse candidates. The individual national laboratories can leverage each other’s work by approaching events or partnerships with the aim of representing not only the lab itself but also the DOE complex. In this way, we significantly increase impact through a collaborative approach.

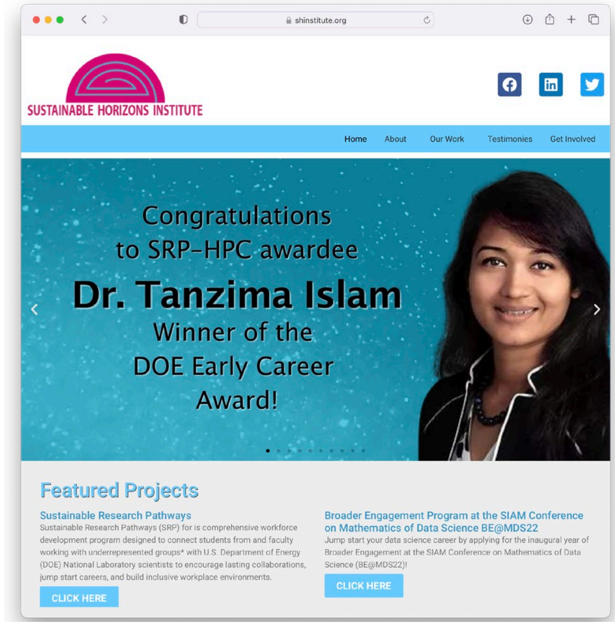

Figure 16-1. Sustainable Horizons Institute landing page (sample).

Further, it is important to continuously identify and diagnose bias in hiring and promotion practices and outcomes. A common activity is to provide bias training for staff. It is recognized, however, that one-time training does not result in cultural change. It is important to apply bias diagnosis in ways that affect hiring and promotions, for example, by analyzing the language used in job descriptions and revising as necessary to be more inclusive. Similarly, organization-wide (and laboratory-wide) climate surveys can be used to identify issues about the culture and practices of the organization. Disaggregated promotion and retention data can be analyzed to further identify issues in these areas.

It is recognized that affinity groups are important to creating inclusive environments, as such groups often provide a sense of belonging, informal mentoring, and a sharing of narratives. It is important to leverage affinity groups at all scales: within individual labs (e.g., employee resource groups), across labs (e.g., national laboratory women of color community forum), and via professional societies and organizations (e.g., National Society of Black Engineers and Advancing Chicanos/Hispanics & Native Americans in Science).

To gauge impact, such efforts must include assessments that measure *outcomes*—as distinct from more common measures such as regarding participation in activities. For inclusion to be normalized and for any new partnerships to be sustainable, such partnerships must ultimately be productive. To have sustained and transformative impact, there must be concrete evidence that diversifying the workforce involved in AI for DOE advances the cause of science. Researchers must see objective evidence that the partnerships and inclusive programs lead to breakthrough science. This ultimately suggests that the investment in this grand challenge must reflect expectations of transformative impact.

16.3.2 DIVERSE LAB WORKFORCE FOR AI AT DOE: LONG TERM

The next decade will witness a democratization of AI such that the high school students of tomorrow will conduct studies that today require the training and resources of Ph.D. students. It is important to consider the full educational ecosystem with respect to developing the needed workforce for AI for DOE science, energy, and security in the long term. Achieving this goal will mean partnering with organizations such as Level Playing Field Institute, which organizes the Summer Math and Science Honors (SMASH) program [13], and CSforAll, which is a national consortium working with state and city educational systems to provide resources to improve the quality of computer science education to all children [14]. Partnerships with such organizations can involve providing grade-appropriate materials relevant to AI for DOE science, energy, and security to excite students about STEM disciplines. Such investments today will have a long-term impact.

With respect to two-year institutions, higher education, graduate education, and alternative pathways, the labs offer internships and visiting faculty positions to help students and faculty, especially from underrepresented communities, engage in the research to advance AI at DOE.

16.3.3 ACCELERATING DEVELOPMENT

Accelerating the path to an AI for DOE workforce that reflects the U.S. demographics requires intentionality, which implies that it is a top priority for everyone and is woven into the way we conduct the science. Much as a fully connected neural network can encode rich information about complex inputs that escape the grasp of a single-layer perceptron, a research community in which ideas flow freely among diverse researchers enables the otherwise daunting challenges of leveraging AI for the most pressing problems in DOE mission science.

16.3.4 DEVELOPING EFFECTIVE AI ETHICS GUIDELINES FOR DOE

DOE's wide-ranging mission space creates a breadth and depth of scientific and technological innovation that has applications in scientific discovery, energy research and production, nuclear security, and environmental management. The AI ethical principles established for DOE will need to provide a framework for balancing scientific discovery against societal impact, while not compromising national security. It is important for DOE to be strategic in its AI investments, while determining policies to ensure the safe development and ethical application of AI technologies consistent with our Nation's values, policies, and priorities.

There has been considerable discussion around AI related ethics in both government and industry. Public statements on AI tend to address broad principles, but often neglect to discuss tools for implementation. Recurring themes include the need for principles that are specific to the organization, that principles should be implementable through actionable practices and procedures, and the importance of acknowledging that considerable scientific advances are necessary to fully enable ethical pursuit of AI. A future framework for ethical AI development and use at DOE should not only specify a set of principles to guide research and development of AI, but must also deliver guidance for implementing these principles.

16.4 References

[1] National Laboratory Directors' Council, 2022, Demographic Data for the National Lab, October, https://nationallabs.org/staff/diversity/, accessed May 12, 2023.

[2] USA Facts, 2022. Our Changing Population: United States, October, https://usafacts.org/data/topics/people-society/population-and-demographics/our-changing-population, accessed May 12, 2023..

[3] Alexander, S., 2016. *The Jazz of Physics: The Secret Link Between Music and the Structure of the Universe*. New York, NY: Basic Books.

[4] Leung, M.A., Rouson, D., and Curfman-McInnes, L., 2020. Increasing productivity by broadening participation in scientific software communities. *2020 Collegeville Workshop on Scientific Software*, July 21–23. https://collegeville.github.io/CW20/WorkshopResources/WhitePapers/leung-broadening-participation-cse-hpc.pdf, accessed May 12, 2023.

[5] Hofstra, B., et al., 2020. The diversity-innovation paradox in science. In: *Proceedings of the National Academy of Science USA*, 117(17), pp. 9284–9291. http://doi.org/10.1073/pnas.1915378117

[6] National Laboratories Directors' Council, 2022. The National Laboratories STEM Resources. https://nationallabs.org/our-labs/stem-resources/, accessed Nov. 28, 2022.

[7] Leung, M.A., 2020. Diversity and inclusion through leadership during challenging times. *Computing in Science and Engineering*, 22(6), pp. 92–96.

[8] Leung, M.A., Crivelli, S., and Brown, D., 2019. Sustainable research pathways: Building connections across communities to diversify the national laboratory workforce. Presented at: *Collaborative Network for Engineering and Computing Diversity (CoNECD)*, April 14–17, Crystal City, VA. https://monolith.asee.org/public/conferences/148/papers/24706/view, accessed May 12, 2023.

[9] The National GEM Consortium. https://www.gemfellowship.org/, accessed May 12, 2023.

[10] Whitney, T., and Taylor, V., 2018. Increasing women and underrepresented minorities in computing: The landscape and what you can do. *Computer*, 51, pp. 24–31. http://doi.org/10.1109/MC.2018.3971359

[11] Clyde, A., 2022. AI for science and global citizens. Patterns, 3(2), 100466. https://doi.org/10.1016/j.patter.2022.100446

[12] Sustainable Horizons Institute. https://shinstitute.org, accessed May 12, 2023.

[13] Level Playing Field Institute (LPFI) SMASH program. https://www.smash.org/about/our-story/, accessed May 12, 2023.

[14] CSforAll. https://www.csforall.org/, accessed May 12, 2023.

17. SCALE

17.1 Current State

Between 2022 and 2024, the U.S. Department of Energy (DOE) will successfully launch its first exascale computing machines. In 2022, it launched Frontier at Oak Ridge Leadership Computing Facility (OLCF, Figure 17-1). In 2023, it will add Aurora at Argonne's Advanced Leadership Computing Facility (ALCF) and then El Capitan at Lawrence Livermore National Laboratory (LLNL). DOE has also deployed the sixth generation of the Energy Sciences Network (ESnet6) and the National Energy Research Scientific Computing Center Supercomputer (NERSC-9, or Perlmutter). These new facilities will support commonly used AI/ML environments, and DOE will invest in development of additional AI-focused supercomputing capabilities in the next decade, along with environments facilitating development of large-scale AI models, (e.g., the foundation models), and real-time federated learning across multiple experimental facilities.

Recent advances in AI have emerged from the ability to use large high-performance computing (HPC) facilities to collect, store, and process large, labeled datasets. The DOE's HPC facilities represent some of the world's largest computational and data ecosystems to generate, move, and analyze experimental and simulation data. These facilities are uniquely positioned to be centers for advances in AI research and applications and must therefore prepare to fully support these capabilities in the next decade [1]. Improving integration among DOE experimental user facilities will ensure domain scientists have the resources to apply AI methods in their research.

Figure 17-1. Frontier Supercomputer, DOE Exascale Computing Project.

The unique position of DOE's HPC facilities to support the advances in AI discussed throughout this report stem partly from the scale at which DOE's Exascale Computing Project (ECP) deployed data and computing resources in 2022 [2]. In this regard, traditional rankings such as the Top500 list do not fully capture scale [3]. For example, the performance of the first ECP machine, Frontier, at DOE-OLCF is nearly as powerful as the *sum of the rest of the machines in the list's top 10 in June 2022*. At 1.102 exaflops, Frontier is 2.5x more powerful than Fugaku, ranked second at 0.442 exaflops. The sum of the remaining eight top 10 systems is 0.724 exaflop [3]. Nevertheless, a system at this scale is designed to support a relatively small number of the largest projects, and the integration of AI methods and approaches brings different types of scale challenges to HPC centers. The computational requirements for training large-scale models, whether surrogates, foundation models, or others, will increase the scale of a number of projects demanding HPC resources. The workforce requirements, particularly with foundation models, will increase the scale of teams from dozens to hundreds of scientists, each with unique training and execution workflows. We discuss these and other scaling challenges below. Simply put, the reinvention of DOE's modeling and simulation approaches—required to achieve the promise of new AI approaches outlined in Section 01 of this report—has the potential to overwhelm DOE's computational resources. For example, generating the training to create a surrogate model for an ECP application will in itself require ensembles of tens to hundreds of

PROJECT SPOTLIGHT

Project Name: CANDLE (CANcer Distributed Learning Environment)

PI: Rick Stevens, Georgia Tourassi, Fred Streitz, Tanmoy Bhattacharya, and Eric Stahlberg

Organizations Involved: Argonne National Laboratory, Oak Ridge National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Brookhaven National Laboratory, and Fredric National Laboratory for Cancer Research

Goal: Use deep learning at scale on DOE leadership computing resources to address molecular, cellular, and population-level problems in cancer research and beyond.

Significant Accomplishment: Members of the CANDLE team contributed to three Gordon Bell COVID-19 Special Prize Finalists during a pivot in 2021 from cancer to COVID-19 in response to a request by the Secretary of Energy.

In the News: CANDLE team members and others presented significant results on COVID-19 in the Gordon Bell COVID-19 special track, including: (1) Language Models for the Prediction of SARS-CoV-2 Inhibitors; (2) Intelligent Resolution: Integrating Cryo-EM with AI-Driven Multi-Resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action; and (3) #COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol.

instances—an illustration of the maxim, “things will get worse before they get better.”

17.2 Grand Challenges

The uniqueness of DOE’s missions on science, energy, and security requires not only supercomputing capabilities, but also scalable evaluation and benchmarking suites for assessing trustworthiness of AI models, integration of AI into experimental facilities for cross-lab AI-guided federated scientific experiments, and large-scale interdisciplinary AI teams to achieve the AI research objectives discussed in the previous chapters. The expected outcome will be a DOE-level scalable AI environment that provides the resources and environment that support the advances detailed throughout this report. The user community will range from scientists solving scientific problems to operational engineers controlling the power grid to National Nuclear Security Administration (NNSA) weapon development and maintenance.

17.2.1 IMPROVE PREDICTIVE CAPABILITIES OF HPC-BASED MODELING AND SIMULATION BY INTEGRATING LARGE-SCALE SCIENTIFIC DATA

DOE’s ECP has invested to leverage AI and scientific data to improve the accuracy and efficiency of scientific prediction on modern HPC [2]. A common routine for AI-based data model integration initially uses a scalable simulator to generate many training samples that complement the missing information in experimental data; then uses the combined training set to build a surrogate model (e.g., deep neural networks); and finally uses the trained surrogate to reduce time-to-solution in computationally expensive tasks, such as solving inverse problems. For example, the Cancer Distributed Learning Environment (CANDLE) project built a scalable, deep neural network code to solve large-scale machine learning problems for cancer-related pilot applications, such as the drug response problem and the treatment strategy problem. DOE’s ECP project, ExaLearn, has demonstrated initial applications of fast neural network emulators to computational cosmology, as well as AI-based inverse solvers to back out complex materials structure from neutron scattering data. However, even with the success of those pioneering efforts, a wide gap still exists between the current and the ideal situation. Large-scale scientific datasets are usually heterogenous and multimodal, and exascale computational models often consist of modules that simulate multi-scale, multiphysics processes. These differences present a major challenge in data-model integration. Blindly integrating data of an incorrect type into an exascale model may deteriorate the performance or undermine the accuracy of the original model. Moreover, unlike the reduced-precision HPC systems used in the industry, DOE’s AI mission requires a completely new co-design of HPC and AI systems that support mixed precisions with a significant portion of double precision machines, to support the accurate AI-based inference and prediction for high-risk scientific applications.

Thus, domain scientists need an AI-based, goal-oriented, data-model integrating system to help find the best model and the best data for their problems. This AI system can initially be trained by prior knowledge and then be actively updated by users’ experiences. The constant AI system update requires massive computing resources to automatically select informative data (e.g., data that are new to the AI model) and pre-processing data, update AI model parameters, and re-validate AI model predictions. Furthermore, each DOE supercomputer will support many of these AI systems for different scientific domains and user facilities. Building an effective hardware and software workflow infrastructure presents another challenge to fulfill this objective (Chapter 13).

17.2.2 EVALUATION AND TRUSTWORTHINESS OF AI-BASED DECISION-MAKING AT SCALE

DOE has a broad range of responsibilities, including managing and overseeing the U.S. energy sector (e.g., power grids and oil reserves) and nuclear arsenal. DOE often must make urgent decisions but may lack sufficient related data and appropriate models (or knowledge of the data and model on which to rely) to inform reliable decisions with high consequences. AI technologies have potential to play an important role in providing suggestions in this urgent decision-making process. However, such decisions rely on trust in the AI-assisted predictions. This means that every step to establish, select data, train, and evaluate AI models should integrate assessments of trustworthiness instead of adding such assessments as an afterthought.

Despite considerable industry advancement of AI models, the resulting methods to assess trustworthiness do not necessarily accommodate exascale simulation and large-scale scientific data. Risk and reliability of AI-assisted predictions are often measured using uncertainty quantification (UQ) or probabilistic training [4]. The calculation of a risk or reliability metric usually requires training and/or evaluating a large ensemble of AI models, therefore the computational resources required to assess trustworthiness of AI models may be dominant in future AI modeling, especially in making high-consequence decisions. For example, the electricity grid has more uncertainties when distributed generation, storage, and dynamics of use (e.g., private solar panels, whole-home batteries, and electric vehicles) are added [5]. An AI-based grid controller must consider the large-scale uncertainties in making operational decisions, but non-scalable UQ and trustworthiness methods may not provide a reliable solution in time. Therefore, building a large-scale evaluation suite to benchmark DOE’s AI models will require a major crosscutting effort across scientific domains, workflows, software, and hardware.

To address this challenge, DOE HPC systems will play a critical role in assessing AI-based predictions for timely decisions in urgent situations. In addition to the surge in computational resource requirements associated with AI model training, this urgency represents a significant departure from the traditional operational model for HPC systems, where no time-critical integration with experimental or operational instruments or facilities exists. We discuss this challenge next.

17.2.3 INTEGRATION OF AI INTO LARGE-SCALE EXPERIMENTAL USER FACILITIES

The concept of AI-enabled smart laboratories and facilities (Chapters 04 and 05) has attracted much attention in the scientific community, and small-scale, self-driving laboratories, from autonomous chemical synthesis to materials discovery, have been successfully demonstrated. However, integrating AI into the operation of DOE's large-scale experimental user facilities is much more challenging. To close the loop between experiments and AI modeling, we need to integrate many independent components into a single integrated platform. Most DOE user facilities were established many years before the recent AI advances, so experimental instruments were not designed to accommodate AI technologies, and do not have sufficient sensors to collect operational data for an AI agent to steer experiments. It may be feasible to upgrade a small-scale research laboratory to an AI-assisted autonomous laboratory by adding sensors and/or controllers to instruments and connecting them to a small computing cluster. However, it is orders of magnitude harder to upgrade a large-scale DOE scientific user facility to accommodate AI techniques.

For example, the DOE's entire Spallation Neutron Source (SNS) facility is an integrated system with a mercury target station, a neutron beam accelerator, and twenty different instruments. Integrating AI into SNS will require a systematic upgrade to all the components of the facility, including adding sensors to the target station to predict the mercury target's lifetime and adding controllers to the accelerator to automatically detect and maintain the neutron beam's stability. These upgrades require intrusive modification of the core hardware of the facility, which may be impractical due to potential safety issues.

This HPC integration is essential, however, as small-scale clusters cannot fulfill the computing and networking resources needed to enable AI at these large-scale DOE facilities. Leveraging DOE's HPC resources will be essential to providing sufficient computing power for autonomous experimental facilities. Because the HPC and experimental facilities are at different geographical locations (many across the country), making seamless connections between them presents another significant challenge. Experimental facilities typically generate a large amount of data for each experiment, and the current paradigm of sending all the data to HPC for analysis is already increasingly intractable; data reduction at the edge is needed. At the same time, the data cannot be reduced to the point of losing critical physics information. The expertise of domain scientists is needed to design a plan that balances data reduction and networking bandwidth to connect HPC and experimental facilities. Similarly, edge-analysis is required in many cases, particularly where AI models are used to control and optimize an experiment—where the latency of analyzing data on a remote HPC system is prohibitive. Therefore, enabling AI at DOE's existing large-scale experimental facilities presents significant challenges from the perspective of hardware, software, and infrastructure, which requires strategic investment.

17.2.4 BUILDING AND MAINTAINING LARGE-SCALE INTERDISCIPLINARY AI TEAMS

To conduct large AI projects in industry, AI research groups increasingly comprise more than a thousand team members. Scientific teams typically include researchers with diverse knowledge backgrounds and expertise, and, on scales involving dozens of participants, the knowledge and communication gaps between team members are relatively easy to manage. At scales of hundreds of participants, however, organization, team building, communication, tracking, operational security, and similar functions need much more sophisticated methods. Moreover, to fully achieve DOE's objectives of using AI to advance science, energy, and security technologies, large-scale interdisciplinary AI research teams with members from significantly different scientific communities with many facets of AI will be necessary. Unlike typical AI teams in industry, DOE's AI teams will need both AI experts and scientists with critical domain expertise but possibly very little AI or even computing knowledge. DOE has a long history of successfully performing research and development with large-scale interdisciplinary teams, from the Manhattan Project to ECP. The breadth of impact that AI can have, as outlined in earlier chapters, will encompass nearly every aspect of DOE's research and development (R&D) and operational missions. Consequently, DOE's future AI teams will include some of the most diversified and interdisciplinary teams ever assembled.

One example is development of a general AI foundation model for scientific discovery for use in a variety of downstream applications, such as material design, chemical synthesis, or drug design. This would require highly coordinated collaboration between physicists, chemists, biologists, mathematicians, and computer scientists. Thus, leveraging DOE's experience to develop organizational support functions to build and train DOE's AI teams presents a significant challenge. Moreover, all AI models evolve as they continuously learn, and need regular updates to adapt to and incorporate new hardware, software, and AI approaches. Even though DOE has extensive experience in managing large-scale codes, maintaining the AI models for DOE requires a sustainable approach to AI teams. Addressing this challenge requires a strategic plan on AI workforce development for prioritized scientific and security areas (Chapter 16).

17.3 Path Forward

The DOE's first three Advanced Scientific Computing Research (ASCR) facilities—Frontier (OLCF), Aurora (ALCF), and El Capitan (LLNL)—support the most popular AI platforms, which are typically much smaller scale than HPC systems. In the coming decade, DOE will develop more AI-focused HPC systems, as well as AI environments for sharing AI models, architectures, weights, and hyperparameters across the DOE complex. The deployment of scalable scientific data management systems to form the foundation for curating high-quality datasets is needed (Chapter 14). This work will continue with the deployment of ESnet-integrated data gateways, which in turn are controlled and optimized by AI workflow and data management models, that facilitate the transfer of data among instruments, experimental facilities, and computational facilities.

The outcomes of ECP, including exascale simulators for science, energy, and security applications; scalable software libraries; and the exascale data infrastructure, will be transformed and integrated into DOE's scalable AI ecosystem. These capabilities will be accessible by the broad scientific community. For example, the scalable simulators will be used to generate datasets to train and validate AI models, and data-driven AI models will be integrated into the exascale simulators to add missing physics to improve the predictive capabilities of the simulators. Concurrently, DOE's scientific experimental facilities will gradually integrate new AI capabilities. To achieve the goal of AI-based self-driving facilities within the next 10 years, AI technologies will be deeply integrated into facility daily operation.

To realize AI-assisted federated facilities and self-driving experimentation across facilities in the DOE complex requires at least several additional exascale machines. These additional machines will train AI foundation models, fully integrate edge computing at experimental facilities for AI data processing and steering experiments, and further the interconnection of all facilities with an upgraded, next-generation ESnet. Without the support of DOE HPC facilities and these resource expansions, the scientific community will struggle to take advantage of the promise of AI. With appropriate direction, funding, and cross-facility cooperation, DOE can achieve a seamlessly interconnected DOE complex.

17.3.1 LARGE-SCALE FEDERATED LEARNING

Training a large-scale AI foundation model on an HPC system will in most cases require experimental and simulation data at the scales of peta- to exabytes, generated from multiple DOE experimental facilities and computational models. Often, regulations or limited network bandwidth prevent sharing of some data. Here, AI-based federated learning techniques can accelerate AI model development by training a high-quality, centralized model, where the training data remains distributed over many locations. For every iteration in training, each experimental facility computes an update of the current model based on its own data and then pushes the update to an HPC facility at another location, which aggregates all the updates from different experimental facilities to obtain a new globally, optimized model [6]. Federated facilities can enable scalable information fusion and decentralized control of assets in a reliable fashion. The realization of federated facilities requires not only exascale computing systems, but also scalable data infrastructure (Chapter 19) and AI workflows (Chapter 13), in order to enable on-the-fly training or updating of AI models using large streams of data generated in situ at multiple DOE facilities.

Usually, raw experimental data cannot be used directly to train AI models. Both AI practice in industry and experience in the scientific community show that significant effort is required to prepare data for each AI project. Currently, we lack the infrastructure and policies to facilitate curating high-quality AI-ready datasets at the scale needed for the AI projects outlined in this report, which are critical to fully realize the potential of AI [7]. The findability, accessibility, interoperability, and reusability (FAIR) data principles provide guidelines to reach this goal, but the effort needed to implement such systems is daunting [8]. The scale of data involved in advancing AI requires focused investment to further develop data management, curation, publication, standardization, and streaming software and services—with an emphasis on exploring the use of AI methods for these tasks (Chapters 13 and 14). A variety of independent activities along these lines is already taking place at every DOE facility. However, the progress needed in this area will require a tightly coordinated effort across the DOE complex. For example, user facilities need edge computing ecosystems that can integrate data preprocessing for AI and connect to the next-generation ESnet. These edge ecosystems will automatically process (compress, label, reformat, restack, tokenize, etc.) raw data for in situ experiment control or for migration to other facilities via ESnet to train or update AI surrogates or foundation models [9, 10].

17.3.2 SMART CYBERINFRASTRUCTURES THROUGH AI AT THE EDGE

DOE has invested in high-performance, national-scale cyberinfrastructure, such as the ESnet, to support large-scale scientific research [11]. ESnet interconnects the entire national laboratory complex, including its HPC and user facilities, allowing scientists to access data independent of time and location through fast connections to the facilities at speeds up to 100 gigabits per second. Today, ESnet carries approximately 20 petabytes of data each month, and DOE's

Basic Energy Science (DOE-BES) Program predicts that its use alone will increase by an order of magnitude in the coming decade. We envision using a large portion of the data handled by ESnet to train AI models. To enable AI-ready ESnet to make intelligent decisions and coordinate actions across the globe, cyberinfrastructure operation—from local instrument to facility to laboratory scale—requires embedded edge AI capabilities throughout the entire ESnet system [12, 13] (Chapter 04).

In addition, DOE's Biological and Environmental Research (BER) Program recently began building large-scale urban integrated field laboratories, establishing the first instances in Chicago, Baltimore, and on the Texas Gulf Coast. These and additional laboratories planned for 2023 and beyond will integrate new field measurement infrastructure with climate and environmental models, relying on AI approaches, from edge-AI-enabled sensors to foundation and surrogate models, to explore climate change and its impact on urban communities.

17.3.3 AI-ASSISTED SELF-DRIVING EXPERIMENTATION ACROSS FACILITIES

An exciting scientific discovery possibility for interconnected instruments across facilities lies in going beyond today's human-in-the-loop experiments to enable large-scale AI models to evaluate results and steer experiments. One example is material imaging. Studying atomic structure and properties of a new material usually requires multiple types of imaging experiments, such as x-ray, electron microscopy, or neutron scattering, each of which unveils certain properties of the material [14]. New, scalable AI and data infrastructures (Chapters 13 and 19) and foundation models (Chapter 02) will enable new scenarios unheard of with today's infrastructure. For instance, X-ray crystallography work at DOE's Advanced Light Source (ALS) at LBL will generate data to train models using HPC at Argonne's ALCF. There, inverse analysis using AI-based foundation models will refine the atomic structure of the material and pass the suggested neutron experiment setup to SNS at ORNL to perform a neutron scattering experiment [15, 16]. The interaction among those facilities—located in three different regions of the country—could repeat several iterations until experimenters obtain the desired results. The implementation of this process as a self-driving experimentation at scale will further require not only sufficiently fast networking and computing powers, but also large-scale, reliable AI-based controllers to coordinate, optimize, and operate experiments (Chapter 05). This type of AI-enabled self-driving material design, synthesis, and evaluation will increase the pace of scientific discovery by orders of magnitude.

Additionally, integrating AI and supporting instrumentation, such as traditional and new edge-AI-enabled sensors (Chapter 15), into experimental facilities must be considered during the design phase of new facilities or upgrades of existing facilities. For example, DOE has approved the design and construction of the Second Target Station (STS) at the SNS to address emerging science challenges [17]. The target station and associated instrument designs must readily accommodate current and future AI technologies by installing additional sensors and controllers for AI agents to perceive, learn, and control entire STS operations.

17.4 References

[1] Gil, Y., Selman, B., chairs, 2019. *A 20-year community roadmap for Artificial Intelligence research in the U.S., A report from the Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI)*, Washington, D.C.

[2] Exascale Computing Project (ECP). https://www.exascaleproject.org/, accessed September 25, 2022.

[3] Top500. https://www.top500.org/. Updated semi-annually.

[4] Liu, S., Zhang, P., Lu, D., et al., 2022. PI3NN: Out-of-distribution-aware prediction intervals from three neural networks. In: *Proceedings of 10th International Conference on Learning Representations (ICLR),* April 25–April 29, virtual conference, arXiv:2108.02327

[5] U.S. Department of Energy, Office of Electricity, 2019. *Smart Grid System Report: 2018 Report to Congress,* Washington, D.C.

[6] Bonawitz, K., Ivanov, V., Kreuter, B., et al., 2017. Practical secure aggregation for privacy-preserving machine learning. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.* October 30–November 3, Dallas, TX, pp. 1175–1191.

[7] LeCun, Y., 2019. Deep learning hardware: past, present, and future. In: *Proceedings of the 2019 IEEE International Solid-State Circuits Conference (ISSCC).* February 17–21, San Francisco, CA, pp. 12–19. DOI: 10.1109/ISSCC.2019.8662396.

[8] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. DOI: 10.1038/SDATA.2016.18.

[9] Blaiszik, B., Ward, L., Schwarting, M., et al., 2019. A data ecosystem to support machine learning in materials science. *MRS Communications,* 9, pp. 1125–1133, arXiv:1904.10423.

[10] National Research Council, 2013. *Frontiers in Massive Data Analysis, a* report by the National Research Council of the National Academies, Washington, D.C., National Academies Press. DOI: 10.17226/18374.

[11] Energy Sciences Network (ESnet). https://www.es.net/, accessed September 14, 2022.

[12] Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L., 2016. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5), pp. 637–646. DOI: 10.1109/JIOT.2016.2579198, accessed October 11, 2019.

[13] Lopez, P., Montresor, A., Epema, D., et al., 2015. Edge-centric computing: Vision and challenges. *ACM SIGCOMM Computer Communication Review,* 45(5), pp. 37–42. DOI: 10.1145/2831347.2831354, accessed September 11, 2022.

[14] Kalyan, A., Mohta, A., Polozov, O., et al., 2018. Neural-guided deductive search for real-time program synthesis from examples. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR),* April 30–May 3, Vancouver, BC, Canada.

[15] Chang, M.-C., Wei, Y., Chen, W.-R., et al., 2019. Accelerating neutron scattering data collection and experiments using AI deep super-resolution learning. arXiv:1904.08450.

[16] U.S. Department of Energy, Basic Energy Sciences Advisory Committee, 2015. *Challenges at the Frontiers of Matter and Energy: Transformative Opportunities for Discovery Science,* a report from the Basic Energy Sciences Advisory Committee, Washington, D.C.

[17] Adams, P., Ankner, J.F., Anovitz, L., et al., 2019. *First Experiments: New Science Opportunities at the Spallation Neutron Source Second Target Station,* Oak Ridge National Laboratory (ORNL) Technical Report ORNL/SPR-2019/1407, DOI: 10.2172/1784183.

18. COMPUTATIONAL RESOURCES

Computational sciences have in the past several decades become integral to virtually all science and technological breakthroughs, and sustained progress is intimately tied to the capacities, usability, and capabilities of available computational resources. However, with the diversity of application classes that have emerged, including the critical operations underpinning artificial intelligence (AI) model training and execution, data preparation, and other functions, computational resources are no longer characterized solely in terms of traditional measures such as floating-point operations per second (flops). In the increasingly diverse hardware ecosystem, having the right type of resources at the right place and with the right connectivity is at least as important as the nominal flop count. There now exists a wide range of computational tasks and challenges, from running massively parallel simulations to training machine learning (ML) models and from controlling experiments using edge computing to mobile sensor swarms. Each of these applications relies on different mixes of hardware, software, storage, communications, and other technical requirements, each calling for specialized solutions. Consequently, each application space represents both a new opportunity for AI and ML to have a significant impact and also unique challenges to developing and deploying effective AI-driven solutions. It is therefore important to understand what computational resources exist today; how they match to the current and predicted workflows, applications, and challenges, especially in relation to AI; and what type of computing must be provided or, if necessary, designed and developed to address future needs and drive new innovation.

18.1 Current State

The U.S. Department of Energy (DOE) has a long tradition of designing and deploying some of the largest supercomputers in the world, regularly fielding multiple systems ranked in the top 10 fastest supercomputers worldwide as tracked by the Top 500 list [1]. In addition to such flagship systems as Frontier [2], Summit [3], Sierra [4], Perlmutter [5], and Polaris [6], DOE facilities also deploy a wide variety of smaller systems from a wide range of vendors and with different architectures and performance characteristics.

Until recently, the fundamental drivers for deployment of these systems have been simulation codes that enable scientific discovery and support national security missions, such as stockpile stewardship. Lately, however, the explosion of AI in commercial applications, among other trends, has caused a shift toward the use of accelerators rather than general-purpose central processing units. Following this trend, most of the recent DOE systems also rely heavily on accelerators, which, from the perspective of modeling and simulation, has caused significant challenges in porting applications to new machines and achieving even a fraction of the code’s theoretical peak performance.

However, in the context of AI-based research at DOE, this architecture choice has turned out to be fortuitous, as many of the current systems are well suited to address the computing needs for much of the initial, small-scale AI workloads. At the same time, at larger scales it has become apparent that current supercomputers are materially different from the computing environments that industry is using to drive the impressive recent advances in AI, such as the creation of foundation models and models with powerful natural language processing capabilities (Chapter 02). Furthermore, the AI challenges in science and engineering also differ substantially from those addressed by industry AI systems: DOE applications typically require much larger per-sample data, have more stringent constraints related to robustness and uncertainty quantification (UQ), and involve far more diverse data types. These needs have led to the development of custom toolchains, such as the Livermore Big Artificial Neural Network (LBANN) toolkit [7], both to exploit DOE-specific hardware and to solve challenges specific to science, energy, and security applications.

The DOE is also continuing to explore next-generation computing concepts and architectures, working with industry on various testbeds to define future directions. This work includes significant investments in AI-specific hardware from partners, such as Cerebras [8], SambaNova [9], Graphcore [10], Groq [11], or Habana [12]—deployed for example in the Argonne Leadership Computing Facility AI testbed [13] or the Livermore Computing machines [14]. Conceptually, all of these architectures refine the notion of a general-purpose accelerator based on graphics processing units, such as those available from NVIDIA [15] and AMD [16], to focus nearly exclusively on the deep-learning use case. However, each of these architectures has its own peculiarities, specialized interfaces, and unique integration mechanisms, which makes even testing capabilities—let alone deploying code in production—challenging. There also exist forays into cloud-based computing environments [17], Arm-based systems [18], and continuous efforts to define future generations of supercomputing systems [19]. Finally, DOE’s large-scale experimental facilities have started to deploy more general-purpose edge computing resources such as those that embed AI hardware within sensor devices [20], in addition to the traditional streaming processing necessary to manage live data streams from particle colliders or light sources.

18.2 Grand Challenges

Although many of the computational resources discussed above are now used to support the development and deployment of AI-driven solutions, few were designed explicitly with this goal in mind. We highlight four grand challenges here.

As AI becomes more central to DOE's mission and the techniques become more specialized and sophisticated, the dual purpose of computing resources (i.e., supporting both traditional simulation and AI) will require careful consideration. For example, as discussed in Chapter 02, there will be a need to develop large-scale foundation models that integrate massive volumes of diverse, multimodal data across many subdomains to support large science communities. Given that the initial industry-developed foundation-style models, trained primarily with text and images, are reaching trillions of parameters, training foundations for science domains, using much more diverse and voluminous data, will certainly require leadership-class computing and push the boundary of what is feasible. For example, training the Megatron-Turing NLG 530B [21] model in 2021 already reached a sustained throughput of roughly 380 petaflops, a number on par with the 102 petaflops and 4.4 exaflops quoted for the 2021 Gordon Bell Prize [22] or the 171-petaflop to 1.1-exaflop range for the 2020 Gordon Bell Prize [23]. However, the architectures used were very different.

The Megatron project used 480 of NVIDIA's DGX nodes [24], while the 2021 Gordon Bell Prize [21] used almost 108K Sunway nodes [25] (with a total of 42M compute cores). This discrepancy raises a question about how future resources will be structured to address all future needs. On one end of the spectrum, the current model may prevail, in which fundamentally scientific computing resources are adapted to AI workloads through custom tool chains. The other extreme would be dedicated AI hardware, potentially built with specialized compute cores. Independent of the specific solution, the overarching grand challenge is to **provide computing resources dedicated to AI-focused workloads on par with—and potentially surpassing—the large simulation workloads.** Just as for current cutting-edge simulations, we expect the most complex possible AI model to be determined by the size of the largest supercomputer available and how effective that machine is solving the respective problem.

The second grand challenge is **the need for ubiquitous access to AI-ready computing resources through the complex**. As highlighted in the technical sections, there exist a plethora of potential use cases for AI in virtually all aspects of DOE's mission, and we expect this number to only increase as AI technologies mature. Exploring these possibilities will require access to relevant resources for all stakeholders.

The third grand challenge will be **the unique need within DOE applications to couple AI-driven solutions to existing simulations, experiments, or sensors**. This need will favor a compromise in which the various stakeholders co-design future systems to be as broadly applicable as possible and a tight coupling between resources both within and across facilities.

The final grand challenge is **the need to support the many DOE-specific edge cases, whether these are operating in hard radiation, on the edge of large-scale experiments, or in autonomous sensors and drones**. In the absence of commercial drivers for such use cases, it will be up to DOE scientists to adapt, extend, and develop the necessary solutions.

18.3 Path Forward

To become a leader in AI for mission-relevant problems, DOE will need to field substantial, AI-focused compute resources on par with current supercomputing systems. Depending on the direction of the commercial systems and in collaboration with vendor partners, DOE will need to determine whether to pursue a hybrid approach in which flagship systems are suitable for both scientific computing and AI workloads or whether there is a need to purse independent, specialized solutions. Given the need for integration (discussed later in this section), the first approach appears more suitable, but the unavoidable price in peak performance of a hybrid system will need to be evaluated.

Beyond the raw compute power necessary for the largest and most complex models, DOE faces a number of additional challenges equally important to the overall success of leveraging emerging AI approaches to drive scientific discovery. To support an ever-growing number of use cases, DOE will need to provide easy access to small- and mid-range, AI-ready computing resources for any application scientist and engineer within the complex. However, as mentioned above, AI applications require different hardware, software stacks, and data infrastructure than most other computing-based problems. Addressing this need will require all DOE sites to house—and support the user communities of—AI-focused systems of sufficient capacity and to provide easy access for model training and development. Furthermore, this access must include support for a variety of AI software stacks, as well as resource management tools adapted to AI needs (e.g., on-the-fly access to compute nodes during training). Unlike the flagship resources and problems which are typically driven by specialized tools and dedicated efforts, the commodity use of AI within DOE will have to rely on commercially driven software stacks. These stacks are typically less well adapted to DOE's security needs, science requirements (e.g., UQ, explainability), and hardware and thus may require additional effort for deployment. Alternatively, cloud-based resources could supplement local shortfalls and provide additional flexibility. Commercial cloud resources could provide elasticity, though at the cost of additional challenges in data movement and information security. A DOE-wide cloud could provide an interesting alternative that rather than moving data to the compute resource, as in commercial systems, could instead move compute resources (in the form of cloud allocations) to the data, that is, to whatever facility is currently housing it. Finally, on-premises clouds are already used to mirror commercial development practices and provide a seamless integration with commodity software stacks. As a final challenge, these resources must be easily accessible to outside partners at universities or other agencies (e.g., Department of Defense, Department of Homeland Security) to facilitate collaboration and reduce barriers.

Another hallmark of DOE applications is the need for a tight coupling between the resources and components involved in a particular problem (as also discussed in Chapter 13). A typical example is the notion of a self-driving facility for autonomous discovery (Chapter 05). Here, data from sensors are collected and processed on edge-based computing resources and integrated into AI-based models, which themselves are trained at and integrated with leadership-class facilities. These high-performance computing (HPC) resources in turn execute high-fidelity simulations (e.g., a “digital twin” of the laboratory process) to provide real-time control to robotic systems operating and optimizing the experiment. Such an approach not only requires AI-enabled systems throughout the entire chain from experiment to supercomputer but also relies on tight connections between these components to achieve the overarching goal. A similar coupling exists in many other situations, for example, when integrating fast inference into a simulation or when guiding an ensemble for optimization. Such use cases fundamentally require a mix of different computing solutions either in the form of hybrid systems not optimized for either simulations or AI but able to adequately serve both or through a tight coupling of specialized resources. Hybrid systems likely will require more concerted efforts in co-designing hardware with various vendors, as DOE will have particular needs that will differ from most other customers, whereas tight coupling will put more pressure on networks, resource management, and workflows.

Finally, within the DOE mission, there exist many important niche applications with unusual requirements. An example is already emerging in DOE research and development in edge computing and sensing hardware that must operate in extreme or hazardous environments; such hardware might include ultrafast edge devices or low-power sensor swarms. These devices may require dedicated hardware designs developed in collaboration with industry partners. Furthermore, the notion of dedicated or at least highly specialized hardware may not be limited to the deployment environment. For example, DOE has a strong need for robust models and reliable uncertainty quantification. Most of the existing techniques that begin to address these concerns rely on ensemble-style training and inference, random variations in the training data, or similar approaches. These strategies might drastically increase the already large computing resources required to train models. One potential solution is to express the necessary replications directly *in silico* using specialized UQ-enabled chips, which could alleviate such problems. Similar concerns might apply for model and data provenance or adversarial defense. In summary, DOE will require a large range of AI-ready compute resources—from leadership-class systems to dedicated hardware—to drive the next generation of scientific discovery and technological progress.

18.4 References

[1] TOP500: The List, undated. https://www.top500.org/, accessed May 12, 2023.

[2] Frontier [supercomputer]. Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory, Oak Ridge, TN. https://www.olcf.ornl.gov/frontier/, accessed May 12, 2023.

[3] Summit [supercomputer]. Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory, Oak Ridge, TN. https://www.olcf.ornl.gov/summit/, accessed May 12, 2023.

[4] Sierra [supercomputer]. Lawrence Livermore National Laboratory, Livermore, CA. https://asc.llnl.gov/sites/asc/files/sierra-fact-sheet.pdf, accessed May 12, 2023.

[5] Perlmutter: High Performance Computing Optimized for Science. National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA. https://perlmutter.carrd.co/, accessed May 12, 2023.

[6] Polaris [supercomputer]. Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL. https://www.alcf.anl.gov/polaris, accessed May 12, 2023.

[7] Van Essen, B., et al., 2015. LBANN: Livermore Big Artificial Neural Network HPC toolkit. In: *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments* (MLHPC ’15), Article 5, pp. 1–6. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2834892.2834897.

[8] Cerebras [company website]. https://www.cerebras.net/, accessed May 12, 2023.

[9] SambaNova Systems [company website]. https://sambanova.ai/, accessed May 12, 2023.

[10] Graphcore [company website]. https://www.graphcore.ai/, accessed May 12, 2023.

[11] Groq [company website]. https://groq.com/, accessed May 12, 2023.

[12] Habana [company website]. https://habana.ai/, accessed May 12, 2023.

[13] ALCF AI Testbed. Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL. https://www.alcf.anl.gov/alcf-ai-testbed, accessed May 12, 2023.

[14] About Livermore Computing. Lawrence Livermore National Laboratory, Livermore, CA. https://hpc.llnl.gov/about-us, accessed May 12, 2023.

[15] NVIDIA Data Center GPUs: The Heart of the Modern Data Center. NVIDIA, Santa Clara, CA. https://www.nvidia.com/en-us/data-center/data-center-gpus/, accessed October 14, 2022.

[16] AMD Accelerators. Advanced Micro Devices, Santa Clara, CA. https://www.amd.com/en/accelerators, accessed May 12, 2023.

[17] LC Cloud Services. Livermore Computing, Lawrence Livermore National Laboratory, Livermore, CA. https://hpc.llnl.gov/cloud/, accessed May 12, 2023.

[18] Evaluation Testbeds. Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL. https://www.alcf.anl.gov/alcf-resources/evaluation-testbeds, accessed May 12, 2023.

[19] Trader, T., 2021. Berkeley Lab debuts Perlmutter, world's fastest AI supercomputer. *HPCwire*, May 21. https://www.hpcwire.com/2021/05/27/nersc-debuts-perlmutter-worlds-fastest-ai-supercomputer/, accessed October 14, 2022.

[20] Beckman, P., et al., 2016. Waggle: An open sensor platform for edge computing. In: *2016 IEEE SENSORS*, pp. 1–3. https://doi.org/10.1109/SENSORS34402.2016

[21] Smith, S., et al., 2022. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. arXiv:2201.11990 [cs.CL], v1 submitted January 28. https://doi.org/10.48550/arxiv.2201.11990

[22] Liu, Y.(A.), Liu, X.(L.), Li, F.(N.), et al., 2021. Closing the "quantum supremacy" gap: Achieving real-time simulation of a random quantum circuit using a new Sunway supercomputer. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (SC '21), Article 3, pp. 1–12. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3458817.3487399

[23] Jia, W., Wang, H., Chen, M., et al., 2020. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. arXiv:2005.00223 [physics.comp-ph], v1 submitted May 1. https://doi.org/10.48550/arxiv.2005.00223

[24] NVIDIA DGX Systems, NVIDIA, Santa Clara, CA. https://www.nvidia.com/en-us/data-center/dgx-systems/, accessed May 12, 2023.

[25] Sunway TaihuLight [supercomputer operated by National Supercomputing Center in Wuxi, Jiangsu, China]. 2022. Wikipedia. https://en.wikipedia.org/wiki/Sunway_TaihuLight, accessed October 14.

19. DATA INFRASTRUCTURE

The modern deep learning revolution has been driven in large part by access to enormous quantities of labeled data [1]. However, most of these datasets have been assembled for reasons other than science, and few have any relevance to the scientific, energy, or security applications that are of interest to the U.S. Department of Energy (DOE) as outlined in Section 02 of this report. The advancements in those areas as envisioned in this report will require a sustained effort on the acquisition, curation, preparation, and management of large quantities of new and extant scientific data and on the infrastructure required to support those activities.

19.1 Current State

DOE has made major investments to ensure that world-class scientific data are produced at its computational, observational, and experimental facilities. These facilities (e.g., light sources, supercomputers, environmental observatories and field campaigns) and those of DOE's partners (e.g., Large Hadron Collider [LHC], Rubin Observatory) generate enormous and exponentially growing quantities of data at scales that dwarf those seen at most other federal agencies. DOE laboratories have also invested in data infrastructure, particularly at DOE's high-performance computing (HPC) centers (including Leadership Computing Facilities, NERSC, and Trilab facilities), some of which have 100's of petabytes of storage. DOE researchers are pursuing a growing volume and variety of projects that employ artificial intelligence (AI) methods to analyze, manage, and otherwise make use of DOE data. DOE efforts to foster a broadly deployed yet integrated research infrastructure are also bearing fruit. The new (sixth generation) of the Energy Sciences Network (ESnet6) connects DOE laboratories and facilities at up to 400+ Gbps. High-speed "Science DMZs" [2] and data transfer nodes, along with the near-ubiquitous Globus software, today allow scientific applications with diverse requirements to use this bandwidth effectively, so that transferring petabytes is now routine [3].

While DOE data clearly have high scientific value, much of DOE's data infrastructure lags that of adversarial nations and even other U.S. domestic agencies (e.g., National Institutes of Health data infrastructure), as well as industry standards and best practices. While DOE data clearly should be curated and preserved for use by current and future generations, many DOE facilities that generate large quantities of data lack sufficient long-term storage capacity and thus are forced to roll-off data to avoid astronomical growth and to prevent untenable data inertia.

Beyond managing the data lifecycle and moving bulk data among facilities, many science disciplines supported by DOE require that data be shared within large and diverse communities [4]. Yet the current state of infrastructure makes this far from routine. Today's fragmented DOE data storage ecosystem, with equally fragmented and often inflexible retention and access policies, leads to repetition of effort and dilution of capabilities that weaken the return on DOE's investment in data production and storage. In almost all cases, the requirements (e.g., formats, metadata, access methods) of large-scale AI model training are not contemplated, confounding our ability to leverage these data resources as necessary to unlock the potential of the AI approaches—requiring unprecedented volumes of multi-modal training data to be prepared, evaluated, and used—as detailed in Section 01 of this report. Additionally, DOE has unique security concerns related to classified and other sensitive data that often prevent effective data integration and use. A unifying system for integrated but distributed data storage with robust and secure AI at all levels in the infrastructure would optimize the ability of researchers to extract scientific knowledge from the data that is produced across the DOE complex.

Imminent developments exacerbate a number of these problems. Next-generation instruments, as discussed in more detail in the next section, will increase data volumes by orders of magnitude. An increased push to federate instruments (e.g., microscopy, light sources, neutron sources linked with HPC) introduces additional new data challenges. So too will more automated generation of data via automated facilities as discussed in Chapter 05. Today, data are rarely generated systematically at scale but rather to address specific user questions. A transition to systematic and automated data generation will be transformative in many fields [5]. Other AI approaches such as foundation models, as discussed next and in Chapter 02, also introduce new requirements.

19.2 Grand Challenges

We envision a future in which the DOE laboratory system has created an AI-enabled and AI-enabling data infrastructure such that all data and models produced within DOE are organized and connected to permit effective discovery, adaptation, curation, and reuse, subject to security concerns that ensure that confidential information is not revealed inappropriately. This infrastructure will allow new data to be produced and enhanced via a co-design process that maximizes the value of collected data for AI-driven discovery. Additionally, powerful integrating models (e.g., foundation models or surrogates underpinning digital twins) are created and updated automatically over time as new data are generated—and all of these capabilities are available for use to guide discovery and innovation. We coin the term ***active collective memory*** to denote this integrated and integrating data infrastructure, with the aspiration that, similar to human memory, it will permit retrieval of relevant information in many different settings, at different levels of detail and abstraction depending on context, while evolving and adapting to maintain dynamic consistency with evolving experience and knowledge.

Realizing this vision of a DOE-wide *active collective memory* will require overcoming important challenges as detailed in the following.

1. **Creation of an AI-driven data and model observatory.** Large quantities of data and trained machine learning (ML) / active learning (AL) models are of little value if data are not structured, discoverable, and accessible in ways suitable for AI applications. The challenge here is to enable AI agents—for example, engaged in prediction and control of complex engineered systems—to rapidly locate/integrate similar data and models, from across and indeed beyond the DOE complex (Figure 19-1). In the case of digital twins (Chapter 04), this approach would entail selecting optimal models for the overall system and for each subsystem. The rich complexity of large datasets renders human-supplied metadata insufficient for fully capturing relevant characteristics as necessary to enable discovery. Here, AI-driven indexing and search methods are likely required, particularly those that can discover and characterize patterns and internal relationships. For example, in the case of an additive manufacturing system tasked with generating a new design, this system should be able to locate data from dozens of similar runs in high-dimensional embedding space, select and fine-tune the associated model(s)—themselves data objects that can be discovered—and with those models produce an optimal manufacturing schedule.

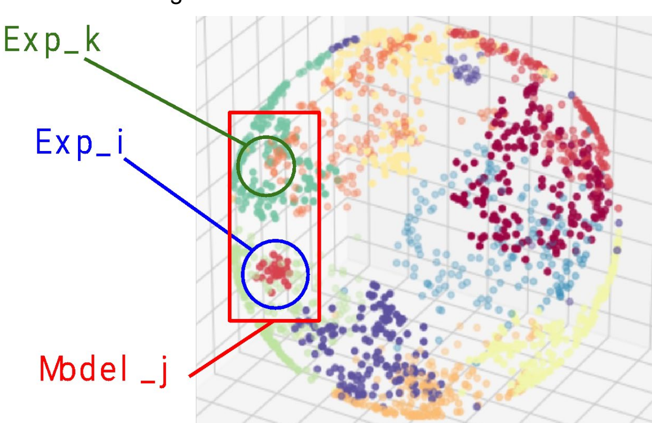

Figure 19-1. An AI-driven data and model observatory should allow for discovery of datasets (e.g., Exp_i, Exp_k) and associated trained models (e.g., Model_j) within a high-dimensional embedding space.

2. **Data infrastructure for creating foundation models.** As described in Chapter 02, foundation models [6]—large AI models trained on large quantities of unlabeled data, usually by self-supervised learning—have emerged as an important technology exploiting deep learning due to the wide variety of downstream tasks to which a trained foundation model can be applied. Technology companies, such as Google and Meta, have demonstrated the ability to train foundation models on large quantities of text, which to date are focused on natural language processing and other applications atypical of DOE science and energy research. However, these models and the techniques for creating them can be leveraged within DOE to build and train foundation models on the large collections of scientific text and/or on large bodies of experimental, observational, and simulation data. In pursuit of the *active collective memory* concept introduced above, we may imagine a malleable, tiered set of AI foundation models with high bandwidth connections. A compact representation of the data will allow usage of this information across a hierarchy of computational infrastructure. The data representations, tailored for AI foundation models (Chapter 11 discusses data representation research challenges), would evolve over time as new measurements are made at DOE facilities, field laboratories, and scientific instruments. These varying foundation models would also connect and coalesce, as relationships are discovered between the different data sources, either by the growing understanding of domain scientists or through connections made by computational analysis at scale. This integrated endeavor could be made to act similar to a *collaboration suggestion engine* for complementary but unwittingly isolated efforts across the DOE complex.

   Realizing these goals will require substantial effort and investments at many levels. One immediate challenge is that DOE data (e.g., documents preserved by DOE's Office of Scientific and Technical Information; data produced at experimental facilities) are not organized in forms suitable for use by foundation models. The current push to make data more easily findable, accessible, interoperable, and reusable (FAIR) [7] can be a step in the right direction but is typically oriented toward the needs of human consumers rather than AI training applications. Overall, data management for foundation models (as well as surrogate models, discussed in Chapter 01) is a multidisciplinary problem that will require sustained effort from data, AI, disciplinary science, and other experts.

3. **Hierarchical federated learning across sources and scales.** DOE scientists produce and use data in many locations and many sources. Certain important data are sensitive: for example, data relating to national security applications (e.g., NNSA) and from commercial partners (e.g., data from power grid operators). Methods are needed to allow AI agents to learn effectively from distributed data: for example, from multiple sensors at a

single facility, in which case rapid coordination is the primary concern; from a single sensor type at multiple facilities, in which case confidentiality is the primary concern; and in more complex cases involving sensor-to-sensor, facility-to-facility, and region-to-region federations. Such methods would enable, for example, an AI agent to learn to predict brownouts based on pooled power grid operator data without the need to reveal sensitive operations data. Significant innovation will be required to enable such hierarchical federated learning across scales, among facilities, and encompassing both observation and simulations. The ability to perform active learning over non-IID (independent, identically distributed) data will be important, as will the ability to integrate data from external sources (e.g., literature) and to encode negative results. These tasks can be tackled in collaboration with industrial partners (e.g., grid operators, battery manufacturers [8], materials companies) who want to learn how to improve processes without revealing sensitive data.

Security and federated learning mechanisms also need to be built into the architecture of federation for the AI foundation models just discussed. This capability will allow access and connections to the information through access control processes that are developed explicitly for multilevel privacy in distributed data-parallel training, also known as federated learning. This system would be much like how humans evolve and adapt or explicitly edit their own memories as their worldview changes, keeping dynamic consistency with our evolving experience and knowledge. Cross-agency security is itself akin to the human ability to control our own release of information depending on social context.

4. **Co-designing massive datasets.** Many methodological innovations in AI over the past decade have been driven by access to large, labeled datasets that were generated essentially at no or very little cost as a side effect of business processes (e.g., storing consumer photographs, consumers clicking on ads). In scientific AI, on the other hand, data collection and data labeling tend to be expensive. Furthermore, accuracy is far more important for scientific applications, given that the goal is fundamental understanding of nature rather than targeted advertising. These considerations suggest a need to develop AI models, science applications, and datasets together via a co-design approach, thereby maximizing the value of each experiment, observation, simulation, and human expert. Data collection processes need to be (1) designed with the end in mind; (2) automated and subject to quality control (QC) processes to ensure that proper contextual metadata are provided and that data meet quality standards; and (3) guided by applications and AI models to identify important sampling directions and to target data of maximum relevance to the scientific problem(s) at hand. For larger datasets, consideration needs to be given to downstream uses and to opportunities to combine with other datasets. These considerations all suggest a *need for* data curation and management to be considered as scientific and engineering skills in their own rights.
5. **Data infrastructure = compute infrastructure.** Current DOE computational facilities are designed primarily to support specialized use cases: primarily large-scale simulation (at leadership computing facilities) and in some cases (e.g., high-energy physics facilities) massively parallel data analysis. Data science applications require new capabilities, such as fast, smart response to new data (e.g., from a new experiment); rapid, random access reads (e.g., when training a foundation model); edge or in-transit processing capabilities (e.g., to filter out interesting events from a high-rate experimental data stream); and continual update of data and knowledge bases as new data appears—for example to perform automatic metadata inference or to update foundation models. These new capabilities may be deployed at existing centers (where they would benefit from proximity to current high-end simulation capabilities) or in other locations (e.g., near experimental facilities). In either case, they require new thinking about data and software architectures. Solutions to these problems will allow DOE facilities and scientists to process and respond in a timely manner to massive data streams coming from many sources and allow for effective integration of ML models into ongoing simulations and experiments.
6. **Online prediction and control of high-data-rate facilities.** DOE experimental and computational facilities face the unique problem of needing to identify interesting events and anomalies in multi-MHz, multi-Tbit/s data streams at decision-relevant speeds. As an example, in the case of a MHz ptychographic imaging facility used to scan a microprocessor or an optical fiber for defects (i.e., to enable rapid imaging of large devices), an AI agent needs to be able to combine historical and online data to detect interesting regions of the chip or fiber and to then to "zoom in" on those regions for more detailed investigation. To provide this capability, new methods are needed for online and continuous learning from high-data-rate sources; ultra-high-data-rate inference; integration of historical and online data and models; uncertainty quantification; and identification of important information for preservation. This work should be supported by pilot projects designed to demonstrate use of online control to achieve a factor of 10 or more improvement in scan speed for several different imaging processes.
7. **Low latency between data and decision.** Making the most of AI's ability to learn quickly from new data requires infrastructure that will respond rapidly to new data being generated. Approaches that rely on updating AI models given new data, such as self-driving laboratories or AI-enhanced simulation codes, will have significantly different

requirements for active memory systems. First, data streams vary greatly in the quantities and dates of data involved, making one-size-fits-all solutions unrealistic. Second, some systems will generate data from many different levels of fidelity (e.g., fast checks performed before big investments), meaning that databases must mix data from many sources though coherent interfaces. Third, there is a strong need for autonomous quality assurance, as “garbage” inputs could lead to garbage decisions. Data management systems must know when to trust and when not to trust data. Examples of success include designing manufactured/synthesized material microstructures with optimized properties; engineering nuclear deterrent systems that are survivable in radiation environments; and optimizing operation of electrical grid under evolving demand environments.

8. **Pervasive data collaboration.** The challenges above have emphasized the requirements of large, big data projects. But the many DOE projects with smaller datasets (projects that, in aggregate, comprise the vast majority of DOE scientists and facility users) also face profound challenges as they seek to leverage the entirety of DOE’s expertise and resources in advancing their own scientific goals while contributing their own products to the DOE knowledge base without compromising their own research. A DOE-wide knowledge base and secure, federated learning capabilities are both needed. Researchers need to be able to determine quickly and easily what data and models already exist relevant to their research problems. Methods are needed to allow data produced by small research teams to be captured, described, and published in ways that place manageable demands on research teams while maximizing value to others. Trustworthy and confidentiality-preserving federated learning will be essential if researchers are to make use of these capabilities. These are profoundly challenging requirements for which no solutions are currently known; extensive research and experimentation will be required to make progress. However, the benefits can be large, as evidenced by examples such as past DOE lab uses of grid storage field data from vendors, and polymer property prediction models trained across data from multiple teams [9].

19.3 Path Forward

19.3.1 ADVANCES IN THE NEXT DECADE

Both the opportunities and challenges associated with the use of DOE data for data-driven discovery are poised to increase substantially soon. New exascale and post-exascale computers will increase scientists’ ability to generate simulation data for AI model training by orders of magnitude. Major enhancements to experimental and observational facilities are also underway. For example, from 120 pulses per second to 1 million pulses per second at LCLS-II (2022); the brighter and more focused beam at the upgraded Advanced Photon Source (APS-U: 2024) will increase data rates by up to a factor of 1000; and the high-luminosity LHC (2029) will increase data rates by an order of magnitude. Other instruments are starting up as well, such as the Rubin Observatory, DUNE neutrino observatory, and high-throughput materials science and biological laboratories. These and other developments will demand major advances in data collection, analysis, and storage capabilities.

Work toward an integrated research infrastructure is also expected to advance quickly. Ultra-fast and reliable ESnet connectivity, broadly deployed data and computing connections, and extensive task automation [10] will make it trivial to implement and run flows that link experiments and simulations with AI agents, data repositories, and other elements of an AI-enabled and AI-enabling DOE science infrastructure. Continued work on policy will be required to avoid bureaucratic barriers to effective resource sharing and collaborative work.

19.3.2 ACCELERATING DEVELOPMENT

Research, infrastructure, and pilot projects are required to accelerate progress on the challenges articulated above: *research* to identify new approaches to known problems, *infrastructure* to support increasingly ambitious experiments, and *pilot projects* to build experience with solutions and to identify the as-yet-unknown challenges that will otherwise emerge, with perhaps fatal effect, only at much later stages.

Research needs include new methods for producing embedded databases; encoding high-dimensional data; capturing and navigating hierarchical relationships; identifying and exploiting redundancy in data; and exposing and supplementing sparse information.

New approaches are required for data storage, curation, and preservation: while current approaches to archival storage may provide cost-effective storage for large volumes of data, their contents are often hard to identify, access, digest, and process. AI advancements are needed to maximize the investments DOE makes in acquiring the best scientific data and to track the derived value of that data—information that can be used to inform dynamic retention policies. This effort should generate AI that acts as a foundation model for the instruments used at the DOE, encapsulating the behavior and properties of a palette of complementary instruments. We intend to generate AI that detects structure, functional relationships, and knowledge representations from large, diverse, and distributed datasets.

Pilot projects are required in a range of data modalities and application needs, and these could be structured similarly to past programs, recognizing that the complexity and volume contemplated in this report far exceed those associated with prior “Big Data” pilot projects. To give just two examples: one promising area will be to demonstrate use of aggregated data and models from 10 or more DOE cross-facility light source beamlines for advanced online prediction and control. A similar opportunity exists for additive manufacturing systems across labs and industry partners.

19.3.3 EXPECTED OUTCOMES

The work articulated in this chapter is intended to foster realization of an AI-enabled and AI-enabling *active collective memory* encompassing all information produced or used by DOE laboratories. Success in this endeavor will include an integrated data infrastructure spanning multiple DOE facilities, greatly increasing the quality and speed of the science performed within the labs, and the impact of DOE facilities on their external users.

At a more granular level, the following are examples of specific advances that we expect to be enabled by such an infrastructure:

- An AI-driven data and model observatory will allow AI agents to call upon the collective knowledge of thousands of experiments at dozens of facilities.
- Online prediction and control methods permit AI agents to make timely decisions based on MHz and Tb/s data streams.
- Hierarchical federated learning across data sources and scales enables AI agents to learn effectively from large, distributed data without centralization.

19.4 References

[1] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

[2] Dart, E., Rotman, L., Tierney, B., Hester, M., and Zurawski, J., 2013. The science DMZ: A network design pattern for data-intensive science. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pp. 1–10.

[3] Chard, K., Dart, E., Foster, I., Shifflett, D., Tuecke, S., and Williams, J., 2018. The modern research data portal: A design pattern for networked, data-intensive science. *PeerJ Computer Science,* 4, e144.

[4] Byna, S., Idreos, S., Jones, T., Mohror, K., Ross, R., and Rusu, F., 2022. *Management and storage of scientific data*, United States. https://doi.org/10.2172/1845705 and https://www.osti.gov/servlets/purl/1845705, accessed January 10, 2023.

[5] Stach, E., DeCost, B., Kusne, A.G., Hattrick-Simpers, J., Brown, K.A., Reyes, K.G., Schrier, J., et al., 2021. Autonomous experimentation systems for materials development: A community perspective. *Matter,* 4(9), pp. 2702–2726.

[6] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, A., von Arx, S., Bernstein, M.S., et al., 2021. On the opportunities and risks of foundation models. *arXiv preprint*, arXiv:2108.07258.

[7] Wilkinson, M.D., Dumontier, M., Jan Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3(1), pp. 1–9.

[8] Ward, L., Babinec, S., Dufek, E.J., Howey, D.A., Viswanathan, V., Aykol, M., Beck, D.A.C., et al., 2022. Principles of the battery data genome. *Joule* 6(10), pp. 2253–2271.

[9] Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R., 2018. Polymer genome: A data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C,* 122(31), pp. 17575–17585.

[10] Vescovi, R., Chard, R., Saint, N., Blaiszik, B., Pruyne, J., Bicer, T., Lavens, A., et al., 2022. Linking scientific instruments and computation: Patterns, technologies, experiences. *Patterns*.

APPENDIXES

AA. AGENDAS

Workshops 1, 2, and 3

WORKSHOP 1: TUESDAY, JUNE 14, 2022

Tennessee State University, Elliot Hall

8:00 a.m. **Registration and Breakfast**

8:45 a.m. **Workshop Welcome**

- *Doug Kothe, Associate Laboratory Director, ORNL, and Jason Pruett, ASC Program Director, LANL*

9:00 a.m. **Tennessee State University Welcome**

- *Dr. Quincy Quick, Interim Vice President for Research and Sponsored Programs*

9:15 a.m. **Workshop Opening**

- *David Womble, AI Program Director, ORNL, and Russell Bent, ASC Deputy Program Director, LANL*

9:20 a.m. **AI4SES Overview**

- *Rick Stevens, Associate Laboratory Director, ANL*

9:50 a.m. **Morning Break**

10:00 a.m. **Plenary Talks**

- *Karen Wilcox, Director, Oden Institute for Computational Engineering and Sciences, University of Texas at Austin*
- *Mike Grosskopf, Scientist, LANL*

11:30 a.m. **Breakout Charge**

- *David Womble, AI Program Director, ORNL, and Russell Bent, ASC Deputy Program Director, LANL*

12:00 p.m. **Lunch**

1:00 p.m. **Domain Breakout Sessions Running Concurrently**

- T1D1: Large, Engineered Networks; location: HSB 103A
- T1D2: Energy Generation and Scheduling; location: HSB 103B
- T1D3: Physics, including High-Energy Physics, Fission, and Fusion; location: HSB 110
- T1D4: Advanced Manufacturing; location HSB 114
- T1D5: Facilities Operations; location HSB 205
- T2D1: Energy Systems and Storage; location HSB 206
- T2D2: Bio and Health Science; location HSB 210
- T2D3: Materials Science; location HSB 243
- T2D4: Climate Science and Earth Systems Predictivity; location HUM 313
- T2D5: Multiscale Physics; location HUM 323

5:00 p.m. **Day One Concludes**

WEDNESDAY, JUNE 15, 2022

Tennessee State University, Elliot Hall

8:00 a.m. **Registration and Breakfast**

9:00 a.m. **Domain Breakout Reports**

- T1D1: Large, Engineered Networks
- T1D2: Energy Generation and Scheduling
- T1D3: Physics, including High-Energy Physics, Fission, and Fusion
- T1D4: Advanced Manufacturing

- T1D5: Facilities Operations

10:15 a.m. **Morning Break**

- T2D1: Energy Systems and Storage
- T2D2: Bio and Health Science
- T2D3: Materials Science
- T2D4: Climate Science and Earth Systems Predictivity
- T2D5: Multiscale Physics

12:00 p.m. **Lunch**

1:00 p.m. **Technology Breakout Sessions Running Concurrently**

- T1T1: AI Foundations and Mathematics; location: HSB 103A
- T1T2: AI Software and Frameworks; location: HSB 103B
- T1T3: Large-scale AI Workflows; location: HSB 110
- T1T4: Data Capabilities and Management for AI; location HSB 114
- T1T5: AI Hardware Architectures; location HSB 205
- T2T1: AI Software and Frameworks; location HSB 206
- T2T2: Bio and Health Science; location HSB 210
- T2T3: Large-scale AI Workflows; location HSB 243
- T2T4: Data Capabilities and Management; location HUM 313
- T2T5: AI Hardware Architectures; location HUM 323

5:00 p.m. **Day Two Concludes**

THURSDAY, JUNE 16, 2022

Tennessee State University, Elliot Hall

8:00 a.m. **Registration and Breakfast**

9:00 a.m. **Technology Breakout Reports**

- T1T1: AI Foundations and Mathematics
- T1T2: AI Software and Frameworks
- T1T3: Large-scale AI Workflows
- T1T4: Data Capabilities and Management for AI
- T1T5: AI Hardware Architectures

10:15 a.m. **Morning Break**

- T2T1: AI Software and Frameworks
- T2T2: Bio and Health Science
- T2T3: Large-scale AI Workflows
- T2T4: Data Capabilities and Management
- T2T5: AI Hardware Architectures

12:00 p.m. **Lunch**

1:00 p.m. **Leadership / Writing Team Convenes for Writing**

3:00 p.m. **Workshop 1: Adjourn**

WORKSHOP 2: TUESDAY, JULY 26, 2022

University of California – Davis, California Hall

8:00 a.m.	**Registration and Breakfast**
8:45 a.m.	**Workshop Welcome** • *J. Rob Neely, Program Coordinator for Computing Environments and CASC ADL, LLNL*
9:00 a.m.	**UC Davis Welcome** • *Cristina Davis, Association Vice Chancellor, UC Davis*
9:15 a.m.	**Workshop Opening** • *Bert de Jong, Group Lead, LBNL, and Brian Spears, Principal Investigator, LLNL*
9:20 a.m.	**AI4SES Overview** • *Rick Stevens, Associate Laboratory Director, ANL*
9:50 a.m.	**Q&A re: AI4SES** • *Laboratory Leadership*
10:00 a.m.	**Morning Break**
10:15 a.m.	**Plenary Talk: AI Applications in Next Generation Food Systems** • *Xin Liu, Professor, Computer Science, UC Davis*
10:45 a.m.	**Plenary Talk: AI for Scientific Computing at Scale – Opportunities and Open Challenges** • *Paris Perdikaris, Assistant Professor of Mechanical Engineering and Applied Mechanics, University of Pennsylvania*
11:15 a.m.	**Confronting (Un)Conscious Bias in AI** • *Tina Park, Head of Inclusive Research and Design, Partnership on AI*
11:45 a.m.	**Breakout Charge** • *Bert de Jong, Group Lead, LBNL, and Brian Spears, Principal Investigator, LLNL*
12:00 p.m.	**Lunch**
1:00 p.m.	**Domain Breakout Sessions Running Concurrently** Topic #1: AI for Advanced Properties Inference and Inverse Design • T1D1: Rational Design in Biochemistry, Chemistry, and Materials; location: California Hall • T1D2: Design and Operation of Multiscale and Multiphysics Systems; location: SCC-Meeting Room A 2nd Floor • T1D3: Automated Design and Optimization of Engineered and Manufacturable Systems; location: SCC-Meeting Room B 2nd Floor • T1D4: Resilient Water and Agriculture Resources; location: SCC-Meeting Room B Room 2nd Floor • T1D5: AI for Energy Resilient Infrastructure; location: SCC-Meeting Room E 2nd Floor Topic #2: Foundation AI for Scientific Knowledge Discovery, Integration, and Synthesis • T2D1: Biomedicine and Healthcare; location: SCC-Multi-Purpose Room 2nd Floor • T2D2: Synthesis of Diverse Data in the Physical Sciences; location: MU-De Carli Room 2nd Floor • T2D3: Emerging Threats in the AI Era; location: MU-Fielder Room 2nd Floor • T2D4: New Approaches to AI Enabled Scientific Discovery; location: MU- Garrison Room 2nd Floor • T2D5: Foundation Models for Decision Support, and Risk and Policy Analysis; location: MU-Smith Room 4th Floor
5:00 p.m.	**Day One Concludes**

WEDNESDAY, JULY 27, 2022

University of California – Davis, California Hall

8:00 a.m. **Registration and Breakfast**

9:00 a.m. **Domain Breakout Reports**

Topic #1: AI for Advanced Properties Inference and Inverse Design

- T1D1: Rational Design in Biochemistry, Chemistry, and Materials
- T1D2: Design and Operation of Multiscale and Multiphysics Systems
- T1D3: Automated Design and Optimization of Engineered and Manufacturable Systems
- T1D4: Resilient Water and Agriculture Resources
- T1D5: AI for Energy Resilient Infrastructure

10:15 a.m. **Morning Break**

10:30 a.m. **Domain Breakout Reports**

Topic #2: Foundation AI for Scientific Knowledge Discovery, Integration, and Synthesis

- T2D1: Biomedicine and Healthcare
- T2D2: Synthesis of Diverse Data in the Physical Sciences
- T2D3: Emerging Threats in the AI Era
- T2D4: New Approaches to AI Enabled Scientific Discovery
- T2D5: Foundation Models for Decision Support, and Risk and Policy Analysis

12:00 p.m. **Lunch**

1:00 p.m. **Technology Breakout Sessions Running Concurrently**

Topic #1: AI for Advanced Properties Inference and Inverse Design

- T1T1: AI Foundations and Mathematics; location: California Hall
- T1T2: AI Software Frameworks, Libraries, and Tools; location: SCC-Meeting Room A 2nd Floor
- T1T3: Large-scale AI Workflows; location: SCC-Meeting Room B 2nd Floor
- T1T4: Data Capabilities and Management for AI; location: SCC-Meeting Room B Room 2nd Floor
- T1T5: AI Hardware Architecture; location: SCC-Meeting Room E 2nd Floor

Topic #2: Foundation AI for Scientific Knowledge Discovery, Integration, and Synthesis

- T2T1: AI Foundations and Mathematics; location: SCC-Multi-Purpose Room 2nd Floor
- T2T2: AI Software Frameworks, Libraries, and Tools; location: MU-De Carli Room 2nd Floor
- T2T3: Large-scale AI Workflows; location: MU-Fielder Room 2nd Floor
- T2T4: Data Capabilities and Management for AI; location: MU-Garrison Room 2nd Floor
- T2D5: AI Hardware Architectures; location: MU-Smith Room 4th Floor

5:00 p.m. **Day Two Concludes**

THURSDAY, JULY 28, 2022

University of California – Davis, California Hall

8:00 a.m. **Registration and Breakfast**

9:00 a.m. **Technology Breakout Reports**

Topic #1: AI for Advanced Properties Inference and Inverse Design

- T1T1: AI Foundation and Mathematics
- T1T2: AI Software and Frameworks
- T1T3: Large-scale AI Workflows
- T1T4: Data Capabilities and Management for AI
- T1T5: AI Hardware Architectures

10:15 a.m. **Morning Break**

10:30 a.m. **Technology Breakout Reports**

Topic #2: Foundation AI for Scientific Knowledge Discovery, Integration, and Synthesis

- T2T1: AI Foundation and Mathematics
- T2T2: AI Software and Frameworks
- T2T3: Large-scale AI Workflows
- T2T4: Data Capabilities and Management for AI
- T2T5: AI Hardware Architectures

12:00 p.m. **Lunch**

1:00 p.m. **Building a Diverse, Equitable, and Inclusive AI Research Community**

2:00 p.m. **Leadership / Writing Team Convenes for Writing**

3:00 p.m. **Workshop 2: Adjourn**

WORKSHOP 3: TUESDAY, AUGUST 16, 2022

Bowie State University, Bldg. 20 - Student Center: Wiseman Ballroom

8:00 a.m. **Registration and Breakfast**

9:00 a.m. **Workshop Welcome**
- *Dr. Carl B. Goodman, Provost and VP for Academic Affairs, BSU*

9:15 a.m. **AI for Science: Overview**
- *Rick Stevens, Associate Laboratory Director, ANL*

9:50 a.m. **AI for Energy: Introduction**
- *Sydni Credle, Technology Manager, NETL*

10:10 a.m. **AI for Security: An NNSA Perspective**
- *Ron Oldfield, Manager, SNL*

10:30 a.m. **Morning Break**

10:45 a.m. **Plenary Talk**
- *Dr. Rosemary Shumba, Chair and Professor, Department of Computer Science, BSU*

11:15 a.m. **Breakout Charge**
- *Ian Foster, Division Director, ANL*

12:00 p.m. **Collect Box Lunch and Proceed to Domain Breakout Sessions**

Domain Breakout #1: AI and Robotics for Autonomous Discovery – Autonomous (robotic) laboratories, e.g., in biology, chemistry, materials, choosing the next experiments
BSU Bldg. 20 - Student Center: Wiseman Ballroom
- *Chair: Arvind Ramanathan, ANL*
- *Co-chair: Joshua Schrier, Fordham University*
- *Scribe: Dinali Jawardana, BSU*

Domain Breakout #2: AI and Robotics for Autonomous Discovery – Analysis of data from large instruments, e.g., in HEP and astronomy
BSU Bldg. 20 - Student Center: Baltimore/Columbia
- *Chair: Luc Peterson, LLNL*
- *Co-chair: Tom Peterka, ANL*
- *Scribe: Tia Dean, BSU*

Domain Breakout #3: AI and Robotics for Autonomous Discovery – Scenarios in which AI is used to steer experimental apparatus, e.g., light sources, Z-machine
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1220
- *Chair: Marcus Noack, LBNL*
- *Co-chair: Christine Sweeney, LANL*
- *Scribe: Mariam Kiran, LBNL*

Domain Breakout #4: AI and Robotics for Autonomous Discovery – Additive and advanced manufacturing with autonomous control, e.g., materials, pits, microelectronics
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1221
- *Chair: David Stevens, LLNL*
- *Co-chair: John Feddema, SNL*
- *Scribe: Vivia Lewis, BSU*

Domain Breakout #5: AI and Robotics for Autonomous Discovery – Automation in field and inhospitable environments, e.g., NNSA materials
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1222
- *Chair: Philip Bingham, ORNL*
- *Co-chair: Steve Buerger, SNL*
- *Scribe: Joed Ngangmeni, SNL*

Domain Breakout #6: AI for Programming and Software Engineering – HPC modeling and simulation, e.g., performance, productivity, using Transformer models to move data from one accelerator to another, using AI for hardware design
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1223

- *Chair: Damian Rouson, LBNL*
- *Co-chair: Feiyi Wang, ORNL*
- *Scribe: Nick Winovich, SNL*

Domain Breakout #7: AI for Programming and Software Engineering – AI hardware and edge devices, e.g., experimental systems, data architectures, neuromorphic computing, co-design
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1224

- *Chair: Siva Rajamanickam, SNL*
- *Co-chair: Valerie Taylor, ANL*
- *Scribe: Zack Morrow, SNL*

Domain Breakout #8: AI for Programming and Software Engineering – Data intensive science, e.g., AI methods for data analysis on HPC systems, deploying AI on HPC systems
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1225

- *Chair: Danny Dunlavy, SNL*
- *Co-chair: Guojing Cong, ORNL*
- *Scribe: Nathaniel Hudson, UChicago*

Domain Breakout #9: AI for Programming and Software Engineering – Real-time control systems, e.g., nuclear reactors, critical infrastructure, grid, etc.
BSU Bldg. 19 – Thurgood Marshall Library: Library Special Collections

- *Chair: Draguna Vrabie, PNNL*
- *Co-chair: David Womble, ORNL*
- *Scribe: Valerie Hayot-Sasson, UChicago*

Domain Breakout #10: AI for Programming and Software Engineering – AI-assisted software development, e.g., vulnerability analysis of software, using AI to identify flaws / vulnerabilities in software, programming systems, transformation, modernization, performance analysis, optimization
BSU Bldg. 19 – Thurgood Marshall Library: Library Auditorium

- *Chair: Rajeev Thakur, ANL*
- *Co-chair: Chunhua Leo Liao, LLNL*
- *Scribe: Qian Gong, ORNL*

5:00 p.m. **Day One Concludes**

WEDNESDAY, AUGUST 17, 2022

Bowie State University, Bldg. 20 - Student Center: Wiseman Ballroom

8:00 a.m. **Registration and Breakfast**

9:00 a.m. **Domain Breakouts Report Out (10 min. each)**

- Domain Breakout #1: AI and Robotics for Autonomous Discovery – Autonomous (robotic) laboratories
- Domain Breakout #2: AI and Robotics for Autonomous Discovery – Analysis of data from large instruments
- Domain Breakout #3: AI and Robotics for Autonomous Discovery – Scenarios in which AI is used to steer experimental apparatus

- Domain Breakout #4: AI and Robotics for Autonomous Discovery – Additive and advanced manufacturing with autonomous control
- Domain Breakout #5: AI and Robotics for Autonomous Discovery – Automation in field and inhospitable environments
- Domain Breakout #6: AI for Programming and Software Engineering – HPC modeling and simulation
- Domain Breakout #7: AI for Programming and Software Engineering – AI hardware and edge devices
- Domain Breakout #8: AI for Programming and Software Engineering – Data intensive science
- Domain Breakout #9: AI for Programming and Software Engineering – Real-time control systems
- Domain Breakout #10: AI for Programming and Software Engineering – AI-assisted software development

11:45 a.m. **Crosscut Breakout Charge**
Ron Oldfield, Manager, SNL

12:00 p.m. **Collect Lunch and Proceed to Crosscut Breakout Sessions**

Crosscut Breakout #1: AI and Robotics for Autonomous Discovery – Software and Frameworks
BSU Bldg. 20 - Student Center: Wiseman Ballroom
- *Chair: Brian Van Essen, LLNL*
- *Co-chair: Mike Grosskopf, LANL*
- *Scribe: Dinali Jawardana, BSU*

Crosscut Breakout #2: AI for Programming and Software Engineering – Software and Frameworks
BSU Bldg. 20 - Student Center: Baltimore/Columbia
- *Chair: Prasanna Balaprakash, ANL*
- *Co-chair: Aleksandra Ciprijanovic, FermiLab*
- *Scribe: Pamela Moses, BSU*

Crosscut Breakout #3: AI and Robotics for Autonomous Discovery – Mathematics and Foundations
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1220
- *Chair: Tommie Catanach, SNL*
- *Co-chair: Sven Leyffer, ANL*
- *Scribe: Amina Ayodeji-Ogundiran, BSU*

Crosscut Breakout #4: AI for Programming and Software Engineering – Mathematics and Foundations
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1221
- *Chair: Rick Archibald, ORNL*
- *Co-chair: Silvia Crivelli, LBNL*
- *Scribe: Aditya Kashi, ORNL*

Crosscut Breakout #5: AI and Robotics for Autonomous Discovery – Workflows (Edge to Center)
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1222
- *Chair: Shantenu Jha, BNL*
- *Co-chair: Peer-Timo Bremer, LLNL*
- *Scribe: Hao Lu, ORNL*

Crosscut Breakout #6: AI for Programming and Software Engineering – Workflows (Edge to Center)
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1223
- *Chair: Arjun Shankar, ORNL*
- *Co-chair: Nicola Ferrier, ANL*
- *Scribe: Tom Uram, ANL*

Crosscut Breakout #7: AI and Robotics for Autonomous Discovery – Data Management and Data Infrastructure
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1224

- *Chair: Rosalyn Rael, LANL*
- *Co-chair: Deb Agarwal, LBNL*
- *Scribe: Kadir Amasyali, ORNL*

Crosscut Breakout #8: AI for Programming and Software Engineering – Data Management and Data Infrastructure
BSU Bldg. 18 – Center for Natural Sciences, Mathematics and Nursing (CSMN): Classroom 1225

- *Chair: Michela Tauffer, University of Tennessee Knoxville*
- *Co-chair: Franck Cappello, ANL*
- *Scribe: Casey Stone, ANL*

Crosscut Breakout #9: AI and Robotics for Autonomous Discovery – AI Hardware Architectures
BSU Bldg. 19 – Thurgood Marshall Library: Library Special Collections

- *Chair: Frank Liu, ORNL*
- *Co-chair: Jim Ang, PNNL*
- *Scribe: Ana Gainaru, ORNL*

Crosscut Breakout #10: AI for Programming and Software Engineering – AI Hardware Architectures
BSU Bldg. 19 – Thurgood Marshall Library: Library Auditorium

- *Chair: Galen Shipman, LANL*
- *Co-chair: Clayton Hughes, SNL*
- *Scribe: Khaled Ibrahim, LBNL*

5:00 p.m. **Day Two Concludes**

THURSDAY, AUGUST 18, 2022

Bowie State University, Bldg. 20 - Student Center: Wiseman Ballroom

8:00 a.m. **Registration and Breakfast**

9:00 a.m. **Crosscut Breakouts Report Out (10 min. each)**

- Crosscut Breakout #1: AI and Robotics for Autonomous Discovery – Software and Frameworks
- Crosscut Breakout #2: AI for Programming and Software Engineering – Software and Frameworks
- Crosscut Breakout #3: AI and Robotics for Autonomous Discovery – Mathematics and Foundations
- Crosscut Breakout #4: AI for Programming and Software Engineering – Mathematics and Foundations
- Crosscut Breakout #5: AI and Robotics for Autonomous Discovery – Workflows (Edge to Center)
- Crosscut Breakout #6: AI for Programming and Software Engineering – Workflows (Edge to Center)
- Crosscut Breakout #7: AI and Robotics for Autonomous Discovery – Data Management and Data Infrastructure
- Crosscut Breakout #8: AI for Programming and Software Engineering – Data Management and Data Infrastructure
- Crosscut Breakout #9: AI and Robotics for Autonomous Discovery – AI Hardware Architectures
- Crosscut Breakout #10: AI for Programming and Software Engineering – AI Hardware Architectures

11:45 a.m. **Concluding Remarks**
Rick Stevens, Associate Laboratory Director, ANL

12:00 p.m. **Collect Lunch and Writing Group Convenes in the Ballroom**

3:00 p.m. **Workshop 3 Adjourns**

AB. COMBINED WORKSHOP REGISTRANTS

FIRST NAME	LAST NAME	INSTITUTION
Jonas	Actor	Sandia National Laboratories
Omotoyosi	Adams	National Nuclear Security Administration
Deb	Agarwal	Lawrence Berkeley National Laboratory
James	Ahrens	Los Alamos National Laboratory
Ahmad	Al Rashdan	Idaho National Laboratory
Francis	Alexander	Brookhaven National Laboratory
Boian	Alexandrov	Los Alamos National Laboratory
Jonathan	Allen	Lawrence Livermore National Laboratory
Kadir	Amasyali	Oak Ridge National Laboratory
Oluwamayowa	Amusat	Lawrence Berkeley National Laboratory
Gemma	Anderson	Lawrence Livermore National Laboratory
James	Ang	Pacific Northwest National Laboratory
Katie	Antypas	Lawrence Berkeley National Laboratory
Rick	Archibald	Oak Ridge National Laboratory
Daniel	Arnold	Lawrence Berkeley National Laboratory
Pedro	Arrechea	IBM
Lloyd	Arrowood	CNS (Y-12)
Halima	Audu	Bowie State University
Amina	Ayodeji-Ogundiran	Bowie State University
Tyler	Backman	Lawrence Berkeley National Laboratory
Prasanna	Balaprakash	Argonne National Laboratory
Feng	Bao	Florida State University
Jennifer	Bauer	National Energy Technology Laboratory
Tom	Beck	Oak Ridge National Laboratory
Pete	Beckman	Argonne National Laboratory
Kristian	Beckwith	Sandia National Laboratories
Carolyn	Begeman	Los Alamos National Laboratory
Mehmet	Belviranli	Colorado School of Mines
Russell	Bent	Los Alamos National Laboratory
Debasis	Bera	Samsung
Wahid	Bhimji	Lawrence Berkeley National Laboratory
Philip	Bingham	Oak Ridge National Laboratory
Aron	Bishop	Bowie State University
Jonathan	Bisila	Sandia National Laboratories
Anika	Bissahoyo	Bowie State University
Johannes	Blaschke	Lawrence Berkeley National Laboratory
Patrick	Blonigan	Sandia National Laboratories
Harry	Bonilla-Alvarado	Ames National Laboratory
Peter	Bosler	Sandia National Laboratories
Kristofer	Bouchard	Lawrence Berkeley National Laboratory
Peer-Timo	Bremer	Lawrence Livermore National Laboratory
Thomas	Brettin	Argonne National Laboratory
Ben	Brown	Lawrence Berkeley National Laboratory
Stephen	Buerger	Sandia National Laboratories
Tan	Bui-Thanh	University of Texas at Austin

FIRST NAME	LAST NAME	INSTITUTION
Aydin	Buluc	Lawrence Berkeley National Laboratory
Josh	Burby	Los Alamos National Laboratory
Shawn	Burns	National Nuclear Security Administration
Paolo	Calafiura	Lawrence Berkeley National Laboratory
Andrea	Calloway	Bowie State University
Eden	Canlilar	Google
Yanzhao	Cao	Auburn University
Franck	Cappello	Argonne National Laboratory
Matthew	Carbone	Brookhaven National Laboratory
Janine	Carney	National Energy Technology Laboratory
Austin	Carson	SeedAI
Jonathan	Carter	Lawrence Berkeley National Laboratory
Tommie	Catanach	Sandia National Laboratories
Charlie	Catlett	Argonne National Laboratory
Mayanka	Chandra Shekar	Oak Ridge National Laboratory
Barry	Chen	Lawrence Livermore National Laboratory
Junhong	Chen	Argonne National Laboratory
Matthew	Cherukara	Argonne National Laboratory
Taylor	Childers	Argonne National Laboratory
Seonho	Choi	Bowie State University
Youngsoo	Choi	Lawrence Livermore National Laboratory
Alok	Choudhary	Northwestern University
Sutanay	Choudhury	Pacific Northwest National Laboratory
Giridhar	Chukkapalli	NVIDIA
Michael	Churchill	Princeton Plasma Physics Laboratory
Randy	Churchill	Princeton Plasma Physics Laboratory
Aleksandra	Ciprijanovic	Fermilab
Mary Ann	Clarke	National Energy Technology Laboratory
Austin	Clyde	Argonne National Laboratory
Ryan	Coffee	Stanford Linear Accelerator Center
William (Bill)	Collins	Lawrence Berkeley National Laboratory
Guojing	Cong	Oak Ridge National Laboratory
Dylan	Copeland	Lawrence Livermore National Laboratory
Sydni	Credle	National Energy Technology Laboratory
Silvia	Crivelli	Lawrence Berkeley National Laboratory
Sajal	Dash	Oak Ridge National Laboratory
Warren	Davis	Sandia National Laboratories
Wibe	de Jong	Lawrence Berkeley National Laboratory
Tia	Dean	Bowie State University
Diego	Del-Castillo-Negrete	Oak Ridge National Laboratory
Thomas	Desautels	Lawrence Livermore National Laboratory
Chris	DeYoung	Penguin
Gautham	Dharuman	Argonne National Laboratory
Sayera	Dhaubhadel	Los Alamos National Laboratory
Emily	Dietrich	Argonne National Laboratory
William	Dorland	Princeton Plasma Physics Laboratory
Eamon	Duede	University of Chicago

FIRST NAME	LAST NAME	INSTITUTION
Vincent	Dumont	Lawrence Berkeley National Laboratory
Danny	Dunlavy	Sandia National Laboratories
Mary	Dzielski	Argonne National Laboratory
Christopher	Earls	Cornell University
Auralee	Edelen	Stanford Linear Accelerator Center
Hilary	Egan	National Renewable Energy Laboratory
Hoda	El-Sayed	Bowie State University
Austin	Ellis	Oak Ridge National Laboratory
Patrick	Emami, Patrick	National Renewable Energy Laboratory
Tegan	Emerson	Pacific Northwest National Laboratory
Keith	Erickson	Princeton Plasma Physics Laboratory
David	Etim	National Nuclear Security Administration
Katherine	Evans	Oak Ridge National Laboratory
Sam	Evans	Harvard University
John	Feddema	Sandia National Laboratories
Kyle	Felker	Argonne National Laboratory
Nicola	Ferrier	Argonne National Laboratory
Hal	Finkel	Department of Energy
Garrison	Flynn	Los Alamos National Laboratory
Sam	Foreman	Argonne National Laboratory
Ian	Foster	Argonne National Laboratory
Geoffrey	Fox	University of Virginia
Devin	Francom	Los Alamos National Laboratory
Joshi	Fullop	Los Alamos National Laboratory
Ana	Gainaru	Oak Ridge National Laboratory
Baskar	Ganapathysubramanian	Iowa State University
Hector	Garcia Martin	Lawrence Berkeley National Laboratory
Anthony	Garland	Sandia National Laboratories
Tim	Germann	Los Alamos National Laboratory
Dipak	Ghosal	University of California, Davis
Ayana	Ghosh	Oak Ridge National Laboratory
Brian	Giera	Lawrence Livermore National Laboratory
Andrew	Gillette	Lawrence Livermore National Laboratory
Jens	Glaser	Oak Ridge National Laboratory
Peter	Glaskowsky	Esperanto Technologies, Inc.
Sonja	Glavaski-Radovanovic	Pacific Northwest National Laboratory
Ylicia	Godinez	National Nuclear Security Administration
Michael	Goldman	Lawrence Livermore National Laboratory
Qian	Gong	Oak Ridge National Laboratory
Aldair	Gongora	Lawrence Livermore National Laboratory
Renee	Gooding	Sandia National Laboratories
Carl	Goodman	Bowie State University
Wyatt	Gorman	Google
Alex	Gorodetsky	University of Michigan
John	Gounley	Oak Ridge National Laboratory
David	Graham	Oak Ridge National Laboratory
Michael	Grosskopf	Los Alamos National Laboratory

FIRST NAME	LAST NAME	INSTITUTION
Salman	Habib	Argonne National Laboratory
Simon	Hammond	National Nuclear Security Administration
Peter	Harrington	Lawrence Berkeley National Laboratory
Valerie	Hayot-Sasson	University of Chicago
Bruce	Hendrickson	Lawrence Livermore National Laboratory
Tae Wook	Heo	Lawrence Livermore National Laboratory
Michael	Heroux	Sandia National Laboratories
Kyle	Hickmann	Los Alamos National Laboratory
Jeffrey	Hittinger	Lawrence Livermore National Laboratory
Justin	Hnilo	U.S. Department of Energy
Thuc	Hoang	National Nuclear Security Administration
Eric	Hoar	Savannah River National Laboratory
Andy	Hock	Cerebras Systems
Sameera	Horawalavithana	Pacific Northwest National Laboratory
Jason	Hou	Pacific Northwest National Laboratory
Paul	Hovland	Argonne National Laboratory
Yu-Ting (Tim)	Hsu	Lawrence Livermore National Laboratory
Xun	Huan	University of Michigan
Andy	Huang	Sandia National Laboratories
Xiaobiao	Huang	Stanford Linear Accelerator Center
Nathaniel	Hudson	University of Chicago
Eliu	Huerta	Argonne National Laboratory
Clay	Hughes	Sandia National Laboratories
Kelli	Humbird	Lawrence Livermore National Laboratory
Lisa	Hundley	Argonne National Laboratory
Wade	Hunter	NextSilicon
Khaled	Ibrahim	Lawrence Berkeley National Laboratory
Michael	Irvin	Argonne National Laboratory
Toby	Isaac	Argonne National Laboratory
Olexandr	Isayev	Carnegie Mellon University
Dan	Jacobson	Oak Ridge National Laboratory
Dinali	Jayawardana	Bowie State University
Shantenu	Jha	Brookhaven National Laboratory
Grant	Johnson	Ames National Laboratory
Earl	Joseph	Hyperion Research
Amy	Justice	VA Connecticut Healthcare System West Haven
Aditya	Kashi	Oak Ridge National Laboratory
Karthik	Kashinath	NVIDIA
Beth	Kaspar	Los Alamos National Laboratory
Jennifer	King	National Renewable Energy Laboratory
Ryan	King	National Renewable Energy Laboratory
Mariam	Kiran	Esnet
Kerstin	Kleese van Dam	Brookhaven National Laboratory
Natalie	Klein	Los Alamos National Laboratory
Elena	Klimova	Bowie State University
Risi	Kondor	University of Chicago
John	Korbin	Bowie State University

FIRST NAME	LAST NAME	INSTITUTION
Ron	Koshita	Pacific Northwest National Laboratory
Olivera	Kotevska	Oak Ridge National Laboratory
Douglas	Kothe	Oak Ridge National Laboratory
Sharlotte	Kramer	Sandia National Laboratories
Aditi	Krishnapriyan	Lawrence Berkeley National Laboratory
Michael	Kruse	Argonne National Laboratory
Ralph	Kube	Princeton Plasma Physics Laboratory
Neeraj	Kumar	Pacific Northwest National Laboratory
Suhas	Kumar	Rain AI
Ana	Kupresanin	Lawrence Livermore National Laboratory
Kuldeep	Kurte	Oak Ridge National Laboratory
Paul	Laiu	Oak Ridge National Laboratory
Michael	Lang	National Nuclear Security Administration
Earl	Lawrence	Los Alamos National Laboratory
Patricia	Lee	U.S. Department of Energy
Steven	Lee	U.S. Department of Energy
Margaret	Lentz	U.S. Department of Energy
Edgar	Leon	Lawrence Livermore National Laboratory
Katie	Lewis	Lawrence Livermore National Laboratory
Vivia	Lewis	Bowie State University
Sven	Leyffer	Argonne National Laboratory
Frankie	Li	HPE
Chunhua	Liao	Lawrence Livermore National Laboratory
Linyu	Lin	Idaho National Laboratory
Frank	Liu	Oak Ridge National Laboratory
Xin	Liu	University of California, Davis
Yajun	Liu	SuperMicro
Zhengchun	Liu	Argonne National Laboratory
Andrey	Lokhov	Los Alamos National Laboratory
Manuel	Lopez Roldan	Graphcore
Vaness	Lopez-Marrero	Brookhaven National Laboratory
Hao	Lu	Oak Ridge National Laboratory
Zarija	Lukic	Lawrence Berkeley National Laboratory
Dalton	Lunga	Oak Ridge National Laboratory
Massimiliano	Lupo Pasini	Oak Ridge National Laboratory
Darlyn	Lutes	Argonne National Laboratory
Nancy	Lybeck	Idaho National Laboratory
Isaac	Lyngaas	Oak Ridge National Laboratory
Heng	Ma	Argonne National Laboratory
Ravi	Madduri	Argonne National Laboratory
Ramana	Madupu	U.S. Department of Energy
Alister	Maguire	Lawrence Livermore National Laboratory
Michael	Mahoney	Lawrence Berkeley National Laboratory
Thomas	Maier	Oak Ridge National Laboratory
Linda	Malone	Oak Ridge National Laboratory
Carla	Mann	Argonne National Laboratory
Manohar	Mareboyana	Bowie State University

FIRST NAME	LAST NAME	INSTITUTION
Vasileios	Maroulas	University of Tennessee Knoxville
Cari	Martinez	Sandia National Laboratories
Chris	Mayes	Stanford Linear Accelerator Center
Benjamin	McMahon	Los Alamos National Laboratory
Diana	McSpadden	Jefferson Laboratory
Murali (Gopalakrishnan)	Meena	Oak Ridge National Laboratory
Daniel	Merl	Lawrence Livermore National Laboratory
Wayne	Mitchell	Lawrence Livermore National Laboratory
Bashir	Mohammed	Lawrence Berkeley National Laboratory
Kelly	Moran	Los Alamos National Laboratory
Dmitriy	Morozov	Lawrence Berkeley National Laboratory
Zachary	Morrow	Sandia National Laboratories
Pamela	Moses	Bowie State University
Silvia	Mulligan	Argonne National Laboratory
Albert	Musaelian	Harvard University
Jamie	Myers	Oak Ridge National Laboratory
Kary	Myers	Los Alamos National Laboratory
Kristian	Myhre	U.S. Department of Energy
Habib	Najm	Sandia National Laboratories
Hai Ah	Nam	Lawrence Berkeley National Laboratory
Ben	Nebgen	Los Alamos National Laboratory
Rob	Neely	Lawrence Livermore National Laboratory
Joed	Ngangmeni	Pacific Northwest National Laboratory
Nga	Nguyen-Fotiadis	Los Alamos National Laboratory
Jeff	Nichols	Oak Ridge National Laboratory
Jonathan	Nistor	BlueWave AI Labs
Marcus	Noack	Lawrence Berkeley National Laboratory
Andrew	Norman	Fermilab
Alex	Norton	Hyperion Research
Peter	Nugent	Lawrence Berkeley National Laboratory
Dan	O'Malley	Los Alamos National Laboratory
Adetunji	Oduduwa	Bowie State University
Aderonke	Oduwole	Bowie State University
Ayodeji	Ogundiran	Bowie State University
Ron	Oldfield	Sandia National Laboratories
Diane	Oyen	Los Alamos National Laboratory
Pinaki	Pal	Argonne National Laboratory
Michael	Papka	Argonne National Laboratory
Vincent	Paquit	Oak Ridge National Laboratory
Julie	Parente	Argonne National Laboratory
Tina	Park	Partnership on AI
Lekha	Patel	Sandia National Laboratories
Ravi	Patel	Sandia National Laboratories
Sean	Peisert	Lawrence Berkeley National Laboratory
Slaven	Peles	Oak Ridge National Laboratory
Swann	Perarnau	Argonne National Laboratory

FIRST NAME	LAST NAME	INSTITUTION
Talita	Perciano	Lawrence Berkeley National Laboratory
Paris	Perdikaris	University of Pensylvania
Tom	Peterka	Argonne National Laboratory
Luc	Peterson	Lawrence Livermore National Laboratory
Yarom	Polsky	Oak Ridge National Laboratory
Stanley	Posey	NVIDIA
Thomas	Potok	Oak Ridge National Laboratory
Line	Pouchard	Brookhaven National Laboratory
Zach	Prince	Idaho National Laboratory
Jason	Pruet	Los Alamos National Laboratory
Irene	Qualters	Los Alamos National Laboratory
Rosalyn	Rael	Los Alamos National Laboratory
Siva	Rajamanickam	Sandia National Laboratories
Kishansingh	Rajput	Jefferson Laboratory
Robert	Rallo	Pacific Northwest National Laboratory
Lavanya	Ramakrishnan	Lawrence Berkeley National Laboratory
Sreenivasan	Ramamurthy	University Of Maryland, Baltimore County
Arvind	Ramanathan	Argonne National Laboratory
Monsuru	Ramnoi	Navajo Tech
Pradeep	Ramuhalli	Oak Ridge National Laboratory
Timothy	Randles	Los Alamos National Laboratory
Nageswara	Rao	Oak Ridge National Laboratory
Jaideep	Ray	Sandia National Laboratories
Yihui	Ren	Brookhaven National Laboratory
Matthew	Reno	Sandia National Laboratories
Juan	Restrepo	Oak Ridge National Laboratory
Ryan	Richard	Ames National Laboratory
Rob	Rieben	Lawrence Livermore National Laboratory
Joshua	Romero	NVIDIA
Damian	Rouson	Lawrence Berkeley National Laboratory
Wissam	Saidi	National Energy Technology Laboratory
Brian	Sammuli	General Atomics
Nandakishore	Santhi	Los Alamos National Laboratory
Soumalya	Sarkar	Raytheon Technologies
Kento	Sato	Riekn
Nobuo	Sato	Jefferson Laboratory
Markus	Schordan	Lawrence Livermore National Laboratory
Mark	Schraad	Los Alamos National Laboratory
Malachi	Schram	Thomas Jefferson National Accelerator Facility
Joshua	Schrier	Fordham University
Sudip	Seal	Oak Ridge National Laboratory
William	Severa	Sandia National Laboratories
Zubair	Shafiq	University of California, Davis
Vivek	Shandilya	Bowie State University
Arjun	Shankar	Oak Ridge National Laboratory
Millikarjun	Shankar	Oak Ridge National Laboratory
Yongning	Sheng	SambaNova Systems Inc.

FIRST NAME	LAST NAME	INSTITUTION
Chung	Shih	National Energy Technology Laboratory
Galen	Shipman	Los Alamos National Laboratory
Amir	Shirkhodaie	Tennessee State University
Rose	Shumba	Bowie State University
Horst	Simon	U.S. Department of Energy
Prashant	Singh	Ames National Laboratory
Mike	Sohn	Lawrence Berkeley National Laboratory
Carlos	Soto	Brookhaven National Laboratory
Brian	Spears	Lawrence Livermore National Laboratory
Claudia	Spiro	NNSA-NA-22
Michael	Sprague	National Renewable Energy Laboratory
Suhas	Sreehari	Oak Ridge National Laboratory
George	Stelle	Los Alamos National Laboratory
Rick	Stevens	Argonne National Laboratory
David	Stevens	Lawrence Livermore National Laboratory
Panos	Stinis	Pacific Northwest National Laboratory
Jennifer	Stokes-Draut	Lawrence Berkeley National Laboratory
Casey	Stone	Argonne National Laboratory
Otto Erik	Strack	Sandia National Laboratories
Thomas	Strohmer	University of California, Davis
Shashank	Subramanian	Lawrence Berkeley National Laboratory
Fred	Sudler	Oak Ridge National Laboratory
Sreenivas	Sukumar	HPE
WaiChing	Sun	Columbia University
Rajeev	Surendran Assary	Argonne National Laboratory
Samantika	Sury	Samsung
Ceren	Susut	U.S. Department of Energy
Sriram	Swaminarayan	Los Alamos National Laboratory
Christine	Sweeney	Los Alamos National Laboratory
Anika	Tabassum	Oak Ridge National Laboratory
Bill	Tang	Princeton Plasma Physics Laboratory
Michela	Taufer	University of Tennessee, Knoxville
Valerie	Taylor	Argonne National Laboratory
Kazuhiro	Terao	Stanford Linear Accelerator Center
Rajeev	Thakur	Argonne National Laboratory
Peter	Thornton	Oak Ridge National Laboratory
Peyton	Ticknor	Oak Ridge National Laboratory
Zoe	Tosi	Lawrence Livermore National Laboratory
Gina	Tourassi	Oak Ridge National Laboratory
Nathaniel	Trask	Sandia National Laboratories
Thomas	Uram	Argonne National Laboratory
Daniela	Ushizima	Lawrence Berkeley National Laboratory
Bart	van Bloemen Waanders	Sandia National Laboratories
Hubertus	Van Dam	Brookhaven National Laboratory
Brian	Van Essen	Lawrence Livermore National Laboratory
Bart	vanBloemenWaanders	Sandia National Laboratories
Dirk	VanEssendelft	National Energy Technology Laboratory

FIRST NAME	LAST NAME	INSTITUTION
Lav	Varshney	Brookhaven National Laboratory
Rama	Vasudevan	Oak Ridge National Laboratory
Stephen	Verzi	Sandia National Laboratories
Richard	Vilim	Argonne National Laboratory
Svitlana	Volkova	Pacific Northwest National Laboratory
Draguna	Vrabie	Pacific Northwest National Laboratory
Adam	Wachtor	Los Alamos National Laboratory
Feiyi	Wang	Oak Ridge National Laboratory
Logan	Ward	Argonne National Laboratory
Jean-Paul	Watson	Lawrence Livermore National Laboratory
Bobbie-Jo	Webb-Robertson	Pacific Northwest National Laboratory
Justin	Weber	National Energy Technology Laboratory
Jack	Wells	NVIDIA
Madison	Wenzlick	National Energy Technology Laboratory
Daniel	White	Lawrence Livermore National Laboratory
Lauret	White	Advanced Micro Devices, Inc.
Rebekah	White	Sandia National Laboratories
Andrew	Wiedlea	Lawrence Berkeley National Laboratory
Stefan	Wild	Argonne National Laboratory
Timothy	Wildey	Sandia National Laboratories
Karen	Willcox	University of Texas, Austin
Nolan	Wilson, Nolan	National Renewable Energy Laboratory
Theresa	Windus	Ames National Laboratory
Nickolas	Winovich	Sandia National Laboratories
Kristen	Winther	Stanford Linear Accelerator Center
Robert	Wisniewski	Samsung
Lora	Wolfe	Oak Ridge National Laboratory
David	Womble	Oak Ridge National Laboratory
John	Wu	Lawrence Berkeley National Laboratory
Dongbin	Xiu	Ohio State University
Angel	Yanguas-Gil	Argonne National Laboratory
Fenghui	Yao	Tennessee State University
Shinjae	Yoo	Brookhaven National Laboratory
Guannan	Zhang	Oak Ridge National Laboratory
Pei	Zhang	Oak Ridge National Laboratory
Amanda	Ziemann	Los Alamos National Laboratory
Steve	Zitney	National Energy Technology Laboratory

AC. ACRONYMS AND ABBREVIATIONS

ACRONYMS	ABBREVIATIONS
3D	three-dimensional
5G, 6G	fifth-generation, sixth-generation [networks]
AAR	Annual Assessment Report (NNSA)
ADAPD	Advanced Data Analytics for Proliferation Detection
AI	artificial intelligence
AI4SES	AI for Science, Energy, and Security
ALCF	Argonne Leadership Computing Facility
ALS	Advanced Light Source (LBNL)
AML	Advanced Machine Learning
AMO	Advanced Manufacturing Office (DOE)
ARD	advanced research direction
ARPA-E	Advanced Research Projects Agency–Energy (DOE)
ASC	Advanced Simulation and Computing (LLNL)
ASCR	Advanced Scientific Computing Research (DOE)
BER	Biological and Environmental Research (DOE-SC)
BES	Basic Energy Sciences (DOE-SC)
CANDLE	Cancer Distributed Learning Environment
CFD	computational fluid dynamics
CMOS	complementary metal oxide semiconductor
DAE	differential-algebraic equation
DARPA	Defense Advanced Research Projects Agency
DDMD	Discovery, Design Optimization, Manufacturing and Certification, and Deployment and Surveillance
DFT	density functional theory
DNN	deep neural network
DNN R&D	Defense Nuclear Nonproliferation Research and Development (NNSA, also NA-22)
DOE	U.S. Department of Energy
DP	Office of Defense Programs (DOE)
DT	digital twin
ECP	Exascale Computing Project
EDA	electronic design automation
EERE	Office of Energy Efficiency and Renewable Energy (DOE)
ESnet	Energy Sciences Network
ESnet6	sixth generation of ESnet
FAIR	Findable, Accessible, Interoperable, Reusable
FECM	Office of Fossil Energy and Carbon Management (DOE)
Flop, flops	floating point operations
FPU	first production unit
FSM	finite state machine
FY	fiscal year
GPU	graphical processing unit
HED	high-energy-density
HEDP	high-energy-density physics
HIL	hardware-in-the-loop
HPC	high-performance computing
I/O	input/output

ACRONYMS	ABBREVIATIONS
IC	integrated circuit
IP	intellectual property
IRI	integrated research infrastructure
ITER	International Thermonuclear Experimental Reactor
LANSC	Los Alamos Neutron Science Center
LBANN	Livermore Big Artificial Neural Network
LEP	life extension programs
LLNL	Lawrence Livermore National Laboratory
LYNM	Low Yield Nuclear Monitoring
MINOS	Multi-Informatics for Nuclear Operations Scenarios
ML	machine learning
NA-10	Office of Defense Programs (NNSA)
NA-20	Office of Defense Nuclear Nonproliferation (NNSA)
NA-22	Defense Nuclear Nonproliferation Research and Development (NNSA)
NA-50	Office of Infrastructure (NNSA)
NA-114	Reference in Ch. 10
NAERM	North American Energy Resilience Model
ND	nuclear deterrent
NDES	nuclear deterrence electrical system
NDTE	Non-Destructive Test and Evaluation
NE	Office of Nuclear Energy (DOE)
NERSC	National Energy Research Scientific Computing Center (DOE-ASCR)
NIH	National Institutes of Health
NNSA	National Nuclear Security Administration
NSF	National Science Foundation
OOD	out-of-distribution
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
PDE	partial differential equation
PF-4	Reference in Ch. 10
PI	Principal Investigator
PPL	probabilistic programming language
QoI	quite OK image (format)
R&D	research and development
RadMHD	radiation-magnetohydrodynamics
RKE	resilient knowledge ecosystem
RL	reinforcement learning
ROM	reduced-order model
SAW	Sandia Analysis Workbench
SC	Office of Science (DOE)
SME	subject matter expert
SNS	Spallation Neutron Source (ORNL)
SSP	scientific seed prompt
STEM	science, technology, engineering, and medicine
STS	Second Target Station (ORNL-SNS)
TB	terabyte
TSRH	Trusted strategically rad-hard

ACRONYMS	ABBREVIATIONS
UQ	**uncertainty quantification**
V&V	**validation and verification**
WCI-ICF	**Weapons and Complex Integration-Inertial Confinement Fusion**
XAI	**Explainable AI**

AD. REFERENCES BY CHAPTER

Executive Summary

[1] Association for the Advancement of AI, 2023. Working together on our future with AI, April 5. https://aaai.org/working-together-on-our-future-with-ai/, accessed May 12, 2023.

[2] National Security Commission on Artificial Intelligence, 2021. *Final Report*, October. https://www.nscai.gov/2021-final-report, accessed December 16, 2022.

[3] Grout, R., Rose, K., Taylor, V., and Essen, B., 2022. *AI@DOE Interim Executive Report*, United States. https://doi.org/10.2172/1872103, https://www.osti.gov/servlets/purl/1872103, accessed May 9, 2023.

Introduction

[1] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., and Chi, E.H., 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682.*

[2] Bender, E.M., Gebru, T., McMillan-Major, A., and Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March, pp. 610–623.

[3] Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., et al. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* 2 (4), 100179. DOI: https://doi.org/10.1016/j.xinn.2021.100179

[4] U.S. Department of Energy-Office of Science, 2020. *AI For Science: Report on the Department of Energy (DOE) town halls on artificial intelligence (AI) for science*, Stevens, R., Taylor, V., Nichols, J., Maccabe, A. B., Yelick, K., and Brown, D. (eds.), Feb. https://publications.anl.gov/anlpubs/2020/03/158802.pdf and https://doi.org/10.2172/1604756, accessed November 30, 2022.

[5] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., and Brynjolfsson, E., 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[6] Kapteyn, M., Pretorius, J., and Willcox, K., 2022. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, Jan. 31 (special one-year anniversary collection).

[7] National Security Commission on Artificial Intelligence, 2021. *Final Report*, October. https://www.nscai.gov/2021-final-report, accessed December 16, 2022.

[8] National Academies, 2022. Machine Learning and Artificial Intelligence to Advance Earth System Science: Opportunities and Challenges – A Workshop, February. https://www.nationalacademies.org/our-work/machine-learning-and-artificial-intelligence-to-advance-earth-system-science-opportunities-and-challenges---a-workshop, accessed December 16, 2022.

[9] Artificial Intelligence and Business Strategy, 2022. *MIT Sloan Management Review*. https://sloanreview.mit.edu/tag/artificial-intelligence-business-strategy/, accessed December 16, 2022.

[10] Schmidt, E., Schawlow, N., Work, R.O., Thornberry III, W., and Flournoy, M., 2022. Mid-decade challenges to national competitiveness. *Special Competitive Studies Project (SCSP)*, September.

[11] Park, Y.J., Kaplan, D., Ren, Z., Hsu, C.-W., Li, C., Xu, H., Li, S., and Li, J., 2023. Can ChatGPT be used to generate scientific hypotheses?, arXiv:2304.12208.

[12] OpenAI, 2023. GPT-4 System Card, March 23. https://cdn.openai.com/papers/gpt-4-system-card.pdf, accessed May 12, 2023.

[13] Dong, G., et al., 2021. Deep learning-based surrogate Model for first-principles global simulations of fusion plasmas. *Nuclear Fusion* 61, 126061.

[14] Scheinker, A., Cropp, F., Paiagua, S., et al., 2021. An adaptive approach to machine learning for compact particle accelerators. *Sci. Rep.* 11, 19187. https://doi.org/10.1038/s41598-021-98785-0

[15] Leeman, S., et al., 2019. Demonstration of machine learning-based model-independent stabilization of source properties in synchrotron light sources. *Phys. Rev. Lett.* 123, 194801. https://doi.org/10.1103/PhysRevLett.123.194801

[16] Wang, D., Du, Q., et al., 2021. Stabilization of the 81-channel coherent beam combination using machine learning. *Optics Express* 29 (4), pp. 5694–5709.

[17] Tactician, undated. A seamless, interactive tactic learner and prover for Coq. https://coq-tactician.github.io/, accessed February 13, 2023.

[18] Crouse, M., 2021. A deep reinforcement learning approach to first-order logic theorem proving. *AAAI*.

[19] Loos, S., Irving, G., Szegedy, C., and Kaliszyk, C., 2017. Deep network guided proof search. https://arxiv.org/abs/1701.06972.

[20] Bansal, K., Loos, S.M., Rabe, M.N., Szegedy, C., and Wilcox, S., 2019. HOList: An Environment for Machine Learning of Higher-Order Theorem Proving. https://arxiv.org/abs/1904.03241.

01. AI and Surrogate Models for Scientific Computing

[1] Choi, Y., Arrighi, W.J., Copeland, D.M., Anderson, R.W. and Oxberry, G.M., 2019. *libROM*. Lawrence Livermore National Laboratory, Livermore, CA (United States).

[2] Kennedy, M.C., and O'Hagan, A., 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3), pp. 425–464.

[3] Heitmann, K., Bingham, D., Lawrence, E., Bergner, S., Habib S., Higdon D., Pope, A., et al., 2016. The Mira–Titan Universe: precision predictions for dark energy surveys. *The Astrophysical Journal* 820(2).

[4] Schunck, N., McDonnell, J.D., Higdon, D., Sarich, J., and Wild, S.M., 2015. Uncertainty quantification and propagation in nuclear density functional theory. *The European Physical Journal A*, 51(12), pp. 1–14.

[5] Schunck, N., O'Neal, J., Grosskopf, M., Lawrence, E., and Wild, S.M., 2020. Calibration of energy density functionals with deformed nuclei. *Journal of Physics G: Nuclear and Particle Physics,* 47(7).

[6] Tapia, G., Khairallah, S., Matthews, M., King, W.E., and Elwany, A., 2018. Gaussian process-based surrogate modeling framework for process planning in laser powder-bed fusion additive manufacturing of 316L stainless steel. *The International Journal of Advanced Manufacturing Technology*, 94(9), pp. 3591–3603.

[7] Daniel Marjavaara, B., Staffan Lundström, T., Goel, T., Mack, Y. and Shyy, W., 2007. Hydraulic turbine diffuser shape optimization by multiple surrogate model approximations of Pareto fronts. *Journal of Fluids Engineering,*129(9), pp. 1228–1240.

[8] Huang, F., Wang, L., and Yang, C., 2015. Hull form optimization for reduced drag and improved seakeeping using a surrogate-based method. In: *The Twenty-fifth International Ocean and Polar Engineering Conference,* OnePetro, June.

[9] Han, Z.H., Görtz, S., and Zimmermann, R., 2013. Improving variable-fidelity surrogate modeling via gradient-enhanced kriging and a generalized hybrid bridge function. *Aerospace Science and technology*, *25*(1), pp. 177–189.

[10] Han, Z.H. and Görtz, S., 2012. Hierarchical kriging model for variable-fidelity surrogate modeling. *AIAA Journal*, *50*(9), pp. 1885–1896.

[11] Guo, X., Li, W., and Iorio, F., 2016. Convolutional neural networks for steady flow approximation. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 481–490. August.

[12] Zhang, Y., Sung, W.J., and Mavris, D.N., 2018. Application of convolutional neural network to predict airfoil lift coefficient. In: *2018 AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference,* p. 1903.

[13] Kadeethum, T., O'Malley, D., Fuhg, J.N., Choi, Y., Lee, J., Viswanathan, H.S., and Bouklas, N., 2021. A framework for data-driven solution and parameter estimation of PDEs using conditional generative adversarial networks. *Nature Computational Science,* 1, pp. 819–829.

[14] Kadeethum, T., Ballarin, F., Choi, Y., O'Malley, D., Yoon, H., and Bouklas, N., 2022. Non-intrusive reduced-order modeling of natural convection in porous media using convolutional autoencoders: Comparison with linear subspace techniques. *Advances in Water Resources*, 160, p. 104098.

[15] Kadeethum, T., O'Malley, D., Choi, Y., Viswanathan, H.S., Bouklas, N., and Yoon, H., 2021. Continuous conditional generative adversarial networks for data-driven solutions of poroelasticity with heterogeneous material properties. *arXiv preprint,* arXiv:2111.14984.

[16] Heitmann, K., Bingham, D., Lawrence, E., Bergner, S., Habib, S., Higdon, D., Pope, A., Biswas, R., Finkel, H., Frontiere, N., and Bhattacharya, S., 2016. The Mira–Titan universe: Precision predictions for dark energy surveys. *The Astrophysical Journal*, 820(2), p. 108.

[17] Ray, J., DeChant, L., Lefantzi, S., Ling J., and Arunajatesan, S., 2018. Robust Bayesian calibration of a k-e model for compressible jet-in-crossflow simulations. *AIAA Journal*, 56(12), pp. 4893–4909, December.

[18] Huang, M., Ray, J., Hou, Z., Ren, H., Liu, Y., and Swiler, L., 2016. On the applicability of surrogate-based MCMC-Bayesian inversion to the Community Land Model: Case studies at flux tower sites. *Journal of Geophysical Research – Atmospheres*, 121(13).

[19] Smith, R.C., 2014. *Uncertainty Quantification*, SIAM Computational Science and Engineering Series.

[20] Brunton, S.L., Proctor, J.L., and Kutz, J.N., 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. In: *Proceedings of the National Academy of Sciences*, 113(15), pp. 3932–3937.

[21] Fries, W.D., He, X., and Choi, Y., 2022. LaSDI: Parametric latent space dynamics identification. *arXiv preprint,* arXiv:2203.02076.

[22] He, X., Choi, Y., Fries, W.D., Belof, J., and Chen, J.S., 2022. gLaSDI: Parametric physics-informed greedy latent space dynamics identification. *arXiv preprint*, arXiv:2204.12005.

[23] Qian, E., Kramer, B., Peherstorfer, B., and Willcox, K., 2020. Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems. *Physica D: Nonlinear Phenomena*, 406, p. 132401.

[24] Schmidt, M., and Lipson, H., 2009. Distilling free-form natural laws from experimental data. *Science*, 324(5923), pp. 81–85.

[25] Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S., 2020. Discovering symbolic models from deep learning with inductive biases. *arXiv preprint,* arXiv:2006*.11287*.

[26] Mezić, I., 2013. Analysis of fluid flows via spectral properties of the Koopman operator. *Annual Review of Fluid Mechanics*, 45, pp. 357–378.

[27] Huhn, Q., Tano, M.E., Ragusa, C.R., and Choi, Y., 2022. Parametric dynamic mode decomposition for reduced order modeling. *arXiv preprint,* arXiv:2204.12006.

[28] Koch, J., 2021. Data-driven surrogates of rotating detonation engine physics with neural ordinary differential equations and high-speed camera footage. *Physics of Fluids,* 33, p. 091703. https://doi.org/10.1063/5.0063624

[29] Raissi, M., Perdikaris, P., and Karniadakis, G.E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, pp. 686–707.

[30] Wang, S., Yu, X., and Perdikaris, P., 2022. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449, p.110768.

[31] Wang, S., Teng, Y., and Perdikaris, P., 2021. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5), pp. A3055–A3081.

[32] Shin, Y., Darbon, J., and Karniadakis, G.E., 2020. On the convergence of physics-informed neural networks for linear second-order elliptic and parabolic type PDEs. *arXiv preprint,* arXiv:*2004.01806*.

[33] Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G.E., 2021. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3), pp. 218–229.

[34] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A., 2020. Fourier neural operator for parametric partial differential equations. *arXiv preprint,* arXiv:*2010.08895*.

[35] Ling, J., Kurzawski, A., and Templeton, J., 2016. Reynolds-averaged turbulence modeling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics,* 807, pp. 155–166. doi:10.1017/jfm.2016.615

[36] Singh, A.P., Medida, S., and Duraisamy, K., 2017. Machine-learning-augmented predictive modeling of turbulent separated flows over airfoils. *AIAA Journal,* 55 (7), pp. 2215-2227.

[37] Frankel, A.L., Safta, C., Alleman, C., and Jones, R., 2022. Mesh-based graph convolutional neural networks for modeling materials with microstructure. *Journal of Machine Learning for Modeling and Computing*, 3(1).

[38] Frankel, A.L., Jones, R.E., and Swiler, L.P., 2020. Tensor basis Gaussian process models of hyperelastic materials. *Journal of Machine Learning for Modeling and Computing*, 1(1).

[39] Kim, H., Kim, J., Won, S., and Lee, C., 2021. Unsupervised deep learning for super-resolution reconstruction of turbulence. *Journal of Fluid Mechanics,* 910, p. A29. doi:10.1017/jfm.2020.1028

[40] Copeland, D.M., Cheung, S.W., Huynh, K., and Choi, Y., 2022. Reduced-order models for Lagrangian hydrodynamics. *Computer Methods in Applied Mechanics and Engineering*, 388, p. 114259.

[41] Amsallem, D., and Farhat, C., 2008. Interpolation method for adapting reduced-order models and application to aeroelasticity. *AIAA journal*, 46(7), pp. 1803–1813.

[42] Cheung, S.W., Choi, Y., Copeland, D.M., and Huynh, K., 2022. Local Lagrangian reduced-order modeling for Rayleigh-Taylor instability by solution manifold decomposition. *arXiv preprint,* arXiv:*2201.07335*.

[43] Lauzon, J.T., Cheung, S.W., Shin, Y., Choi, Y., Copeland, D.M., and Huynh, K., 2022. S-OPT: A points selection algorithm for hyper-reduction in reduced-order models. *arXiv preprint,* arXiv:2203.16494.

[44] Xiao, D., Fang, F., Buchan, A.G., Pain, C.C., Navon, I.M., Du, J., and Hu, G., 2014. Non-linear model reduction for the Navier–Stokes equations using residual DEIM method. *Journal of Computational Physics*, 263, pp. 1–18.

[45] Stabile, G., and Rozza, G., 2018. Finite volume POD-Galerkin stabilized reduced-order methods for the parametrized incompressible Navier–Stokes equations. *Computers & Fluids*, 173, pp. 273–284.

[46] Veroy, K., and Patera, A.T., 2005. Certified real-time solution of the parametrized steady incompressible Navier–Stokes equations: Rigorous reduced-basis a posteriori error bounds. *International Journal for Numerical Methods in Fluids*, 47(8-9), pp. 773–788.

[47] Choi, Y., Brown, P., Arrighi, W., Anderson, R., and Huynh, K., 2021. Space–time reduced-order model for large-scale linear dynamical systems with application to Boltzmann transport problems. *Journal of Computational Physics*, 424, p. 109845.

[48] McBane, S., and Choi, Y., 2021. Component-wise reduced-order model lattice-type structure design. *Computer Methods in Applied Mechanics and Engineering*, 381, p. 113813.

[49] McBane, S., Choi, Y., and Willcox, K., 2022. Stress-constrained topology optimization of lattice-like structures using component-wise reduced-order models. *arXiv preprint*, arXiv:2205.09629.

[50] Kapteyn, M.G., Knezevic, D.J., Huynh, D.B.P., Tran, M., and Willcox, K.E., 2022. Data-driven, physics-based digital twins via a library of component-based reduced-order models. *International Journal for Numerical Methods in Engineering*, 123(13), pp. 2986–3003.

[51] Choi, Y., Oxberry, G., White, D., and Kirchdoerfer, T., 2019. Accelerating design optimization using reduced-order models. *arXiv preprint,* arXiv:1909.11320.

[52] Choi, Y., Boncoraglio, G., Anderson, S., Amsallem, D., and Farhat, C., 2020. Gradient-based constrained optimization using a database of linear reduced-order models. *Journal of Computational Physics*, 423, p. 109787.

[53] Choi, Y., Oxberry, G., White, D., and Kirchdoerfer, T., 2019. *Accelerating topology optimization using reduced-order models*. LLNL-CONF-771564. Lawrence Livermore National Laboratory, Livermore, CA (United States).

[54] Amsallem, D., Zahr, M., Choi, Y., and Farhat, C., 2015. Design optimization using hyper-reduced-order models. *Structural and Multidisciplinary Optimization*, 51(4), pp. 919–940.

[55] Kim, Y., Choi, Y., Widemann, D., and Zohdi, T., 2021. A fast and accurate physics-informed neural network reduced-order model with shallow-masked autoencoder. *Journal of Computational Physics*, p. 110841.

[56] Kim, Y., Choi, Y., Widemann, D., and Zohdi, T., 2020. Efficient nonlinear manifold reduced-order model. *arXiv preprint,* arXiv:*2011.07727*.

[57] Sandia Analysis Workbench, 2022. National Technology and Engineering Solutions of Sandia, LLC. https://www.sandia.gov/saw/, accessed May 12, 2023.

[58] Allan, B.A., Armstrong, R.C., Wolfe, A.P., Ray, J., Bernholdt, D.E., and Kohl, J.A., 2002. The CCA core specification in a distributed memory SPMD framework. *Concurrency Computat.: Pract. Exper.,* 14, pp. 323–345. https://doi.org/10.1002/cpe.651

[59] Schmelzer, M., Dwight, R.P., and Cinnella, P., 2020. Discovery of algebraic Reynolds-stress models using sparse symbolic regression. *Flow Turbulence Combust* 104, pp. 579–603. https://doi.org/10.1007/s10494-019-00089-x

[60] Barone, M., Ray, J., and Domino, S., 2022. Feature selection, clustering, and prototype placement for turbulence data sets. *AIAA Journal*, 60(3), pp.1332–1246.

[61] Dakota Web page, 2021. National Technology and Engineering Solutions of Sandia, LLC. https://dakota.sandia.gov, accessed May 12, 2023.

[62] DeVore, R., Hanin, B., and Petrova, G., 2021. Neural network approximation. *Acta Numerica*, 30, pp. 327–444. doi:10.1017/S0962492921000052

[63] Petersen, P., 2022. *Neural Network Theory*. University of Vienna. http://pc-petersen.eu/Neural_Network_Theory.pdf, accessed May 12, 2023.

[64] Schmelzer, M., Dwight, R.P., and Cinnella, P., 2020. Discovery of algebraic Reynolds-stress models using sparse symbolic regression. *Flow Turbulence Combust* 104**,** pp. 579–603. https://doi.org/10.1007/s10494-019-00089-x

[65] Boullé, N., Earls, C.J., and Townsend, A., 2022. Data-driven discovery of Green's functions with human-understandable deep learning. *Sci Rep.* 12**,** p. 4824. https://doi.org/10.1038/s41598-022-08745-5

[66] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D., 2019. A survey of methods for explaining black box models. *ACM Comput. Surv.,* 51, 5, Article 93, September, 42 pp. https://doi.org/10.1145/3236009

[67] Burkart, N., and Huber, M.F., 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, pp. 245–317.

02. AI Foundation Models for Scientific Knowledge Discovery, Integration, and Synthesis

[1] Bommasani, R., Hudson, D.A., and Adeli, E., et al., 2021. On the opportunities and risks of foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2108.07258.

[2] Steinhardt, J., 2021. On the risks of emergent behavior in foundation models, Stanford University Human-Centered Artificial Intelligence. https://crfm.stanford.edu/commentary/2021/10/18/steinhardt.html, accessed December 8, 2022.

[3] Bommasani, R., and Liang, P., 2021. Reflections on foundation models, Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/news/reflections-foundation-models, accessed December 8, 2022.

[4] Jia, Y., 2019. Attention mechanism in machine translation. *Journal of Physics: Conference Series*, Vol. 1314, pp. 012186. DOI:10.1088/1742-6596/1314/1/012186

[5] Thomas, J., 2022. A shot for the ages: Fusion ignition breakthrough hailed as 'one of the most impressive scientific feats of the 21st century.' LLNL News, Dec. 14. https://www.llnl.gov/news/shot-ages-fusion-ignition-breakthrough-hailed-one-most-impressive-scientific-feats-21st, accessed Jan. 10, 2023.

[6] Chen, M., et al., 2020. Generative pretraining from pixels. In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 1691–1703. http://proceedings.mlr.press/v119/chen20s.html, accessed May 12, 2023.

[7] Rives, A., et al., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, Vol. 118, pp. e2016239118. https://doi.org/10.1073/pnas.2016239118

[8] Rothchild, D., et al., 2021. C5T5: Controllable generation of organic molecules with transformers. *arXiv*. https://doi.org/10.48550/arxiv.2108.10307

[9] Lee, J., et al., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Vol. 36, pp. 1234–1240.

[10] Yanrong J., et al., 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37, pp. 2112–2120. https://doi.org/10.1093/bioinformatics/btab083

[11] Hugging Face, undated. Transformers. https://huggingface.co/docs/transformers/index, accessed May 12, 2023.

[12] AI21 Labs, 2021. Announcing AI21 Studio and Jurassic-1 language models. https://www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1, accessed May 12, 2023.

[13] Sullivan, M., 2021. Ex-Googlers raise $40 million to democratize natural-language AI. *Fast Company*. https://www.fastcompany.com/90670635/ex-googlers-raise-40-million-to-democratize-natural-language-ai, accessed May 12, 2023.

[14] Ricadela, A., 2021. Powered by cloud, self-learning AI models are turning programming on its head. *Fast Company*. https://www.fastcompany.com/90683767/powered-by-cloud-self-learning-ai-models-are-turning-programming-on-its-head, accessed May 12, 2023.

[15] Nayak, P., 2019. Understanding searches better than ever before. *The Keyword*. https://blog.google/products/search/search-language-understanding-bert/, accessed May 12, 2023.

[16] Meta AI, 2020, AI advances to better detect hate speech. https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/, accessed May 12, 2023.

[17] Rosset, C., 2020. Turing-NLG: A 17-billion-parameter language model by Microsoft. *Microsoft Research Blog.* https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/, accessed May 12, 2023.

[18] Hatakeyama-Sato, K., and Oyaizu, K., 2020. Integrating multiple materials science projects in a single neural network. *Communications Materials*, 1, pp. 1–10. https://doi.org/10.1038/s43246-020-00052-8

[19] Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., and Grover, A., 2023. ClimaX: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343.

03. AI for Advanced Property Inference and Inverse Design

[1] Carleo, G., et al., 2019. Machine learning and the physical sciences. *Rev. Modern Phys.* 91(4), 045002. DOI 10.1103/RevModPhys.91.045002

[2] Hayat, M.A., Stein, G., Harrington, P., Lukić, Z., Mustafa, M., 2021. Self-supervised representation learning for astronomical images. *ApJL* 911(2), L33. https://doi.org/10.3847/2041-8213/abf2c7

[3] Charnock, T., Perreault-Levasseur, L., Lanusse, F., 2022. Bayesian neural networks. In *Artificial Intelligence for High Energy Physics,* pp. 663–713. https://doi.org/10.1142/9789811234033_0018

[4] Cranmer, K., Brehmer, J., Louppe, G., 2020. The frontier of simulation-based inference. *PNAS*, 117(48), 30055-30062.

[5] Jumper, J., Evans, R., Pritzel, A. et al., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature,* 596, pp. 583–589. https://doi.org/10.1038/s41586-021-03819-2

[6] Blay, V., Radivojevic, T., Allen, J.E., Hudson, C.M., Garcia Martin, H., 2022. MACAW: An accessible tool for molecular embedding and inverse molecular design. *J. Chem. Inf. Mod,* 62(15), pp. 3551-3564. DOI: 10.1021/acs.jcim.2c00229

[7] Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos Jr, J.L., Xiong, C., Sun, Z.Z., Socher, R., and Fraser, J.S., 2023. Large language

models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp.1–8.

[8] Volk, M.J., Lourentzou, I., Mishra, S., Tung Vo, L., Zhai, C., Zhao, H., 2020. Biosystems design by machine learning. *ACS Synth. Bio.* 9(7), pp. 1514–1533. DOI: 10.1021/acssynbio.0c00129

[9] Alberi, K,. et al., 2019. The 2019 materials by design roadmap. *J. Phys. D: Appl. Phys.* 52, 013001. https://doi.org/10.1088/1361-6463/aad926

[10] Choudhary, K., DeCost, B., Chen, C., et al., 2022. Recent advances and applications of deep learning methods in materials science. *npj Comput Mater,* 8, 59. https://doi.org/10.1038/s41524-022-00734-6

[11] DOE ASCR Report, 2018. *Basic Research Needs for Microelectronics*, Oct. 23–25, https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN_Microelectronics_rpt.pdf, accessed May 12, 2023.

[12] Max Mowbray, M., Vallerio, M., Perez-Galvan, C., Zhang, D., Del Rio Chanona, A., Navarro-Brull, F.J., 2022. Industrial data science – a review of machine learning applications for chemical and process industries, *React. Chem. Eng,* 7, pp. 1471–1509. DOI: 10.1039/D1RE00541C

[13] DOE Office of Electricity Report, 2019. *North American Energy Resilience Model,* July https://www.energy.gov/sites/prod/files/2019/07/f65/NAERM_Report_public_version_072219_508.pdf, accessed May 12, 2023.

[14] IAW Report, undated. *Digital Water: Artificial Intelligence Solutions for the Water Sector*, https://iwa-network.org/wp-content/uploads/2020/08/IWA_2020_Artificial_Intelligence_SCREEN.pdf, accessed May 12, 2023.

[15] Freiesleben, T., König, G., Molnar, C., Tejero-Cantero, A., 2022. Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. arXiv:2206.05487 [stat.ML]. https://doi.org/10.48550/arXiv.2206.05487

[16] DOE ASCR Report, 2019. *Data and Models: A Framework for Advancing AI in Science*, Dec. 16. https://www.osti.gov/biblio/1579323, accessed May 12, 2023.

[17] Gunning, D., Vorm, E., Yunyan Wang, J., Turek, M., 2021. DARPA's explainable AI (XAI) program: A retrospective. *Appl. AI Lett.* (2)e61. https://doi.org/10.1002/ail2.61

[18] McGovern, A., 2021. NSF AI institute for research on trustworthy AI in weather, climate, and coastal oceanography. *AI Matters*, *6*(3), pp. 14–16.

[19] Bates, J., 2021. Expanding the geography of innovation: NSF AI Research Institutes 2021. *NSF Science Matters*. https://beta.nsf.gov/science-matters/expanding-geography-innovation-nsf-ai-research, accessed May 12, 2023.

[20] DOE ASCR, 2019. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*, PRD #2, https://www.osti.gov/biblio/1478744, accessed May 12, 2023.

[21] DOE ASCR Report, 2021. *Toward a Seamless Integration of Computing, Experimental, and Observational Science Facilities: A Blueprint to Accelerate Discovery*, March 8. https://www.osti.gov/biblio/1863562, accessed May 12, 2023.

04. AI-Based Design, Prediction, and Control of Complex Engineered Systems

[1] Thurner, S., Klimek, P., and Hanel, R., 2018. *Introduction to the Theory of Complex Systems*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780198821939.001.0001

[2] Chai, T., Qin, J., and Wang, H., 2014. Optimal operational control for complex industrial processes. *Annual Reviews in Control*, 38, pp. 81–92.

[3] Gao, R.X., et al., 2020. Big data analytics for smart factories of the future. *CIRP Annals*, 69(2), pp. 668–692. https://doi.org/10.1016/j.cirp.2020.05.002

[4] Ran, Y., et al., 2019, “A survey of predictive maintenance: Systems, purposes and approaches. arXiv:1912.07383v1. Submitted December 12. https://doi.org/10.48550/arXiv.1912.07383

[5] Aymar, R., Barabaschi, P., and Shimomura. Y., 2002. The ITER design. *Plasma physics and controlled fusion* 44 (5), p. 519. (See also https://www.iter.org.)

[6] Dong, G., et al., 2021. Deep learning based surrogate model for first-principles global simulations of fusion plasmas. *Nuclear Fusion*, 61, p. 126061–126071.

[7] Jones, D., et al., 2020. Characterizing the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29(A), pp. 36–52. https://doi.org/10.1016/j.cirpj.2020.02.002

[8] Niederer, S., et al., 2021. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science*, 1, pp. 313–320. https://doi.org/10.1038/s43588-021-00072-5

[9] He, Y., Guo, J., and Zheng, X., 2018. From surveillance to digital twin: Challenges and recent advances of signal processing for industrial Internet of Things. *IEEE Signal Processing Magazine*, 35(5), pp. 120–129. http://doi.org/10.1109/MSP.2018.2842228

[10] Moyne, J., et al., 2020. A requirements driven digital twin framework: Specification and opportunities. *IEEE*

Access, 8, pp. 107781–107801. https://doi.org/10.1109/ACCESS.2020.3000437

[11] Womble, D., and Hembree, C., eds., 2022. *The Second Artificial Intelligence for Robust Engineering and Science Workshop Report*. Technical Report ORNL/LTR-2022/399. Oak Ridge, TN: Oak Ridge National Laboratory, March.

[12] Silva, S.H., and Najafirad, P., 2020. Opportunities and challenges in deep learning adversarial robustness: A survey. arXiv:2007.00753. https://doi.org/10.48550/arXiv.2007.00753

[13] Stracuzzi, D.J., et al, 2017. *Uncertainty Quantification for Machine Learning*. Sandia Report SAND2017-6776. Albuquerque, NM, and Livermore, CA: Sandia National Laboratories.

[14] Sutton, R.S., and Barto, A.G., 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

[15] Capra, M., et al., 2019. Edge computing: A survey on the hardware requirements in the Internet of Things world. *Future Internet*, 11(4), pp. 100–124. https://doi.org/10.3390/fi11040100

[16] Baracaldo, N., et al., 2017. November. Mitigating poisoning attacks on machine learning models: A data provenance based approach. In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 103–110.

[17] Kates-Harbeck, J., Svyatkovskiy, A., and Tang, W., 2019. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568, pp. 526–531.

05. AI and Robotics for Autonomous Discovery

[1] Tansley, S., and Tolle, K.M., 2009. *The fourth paradigm: Data-intensive scientific discovery*. Hey, A. J. G. (ed.). Vol. 1. Redmond, WA: Microsoft research.

[2] King, R.D., et al., 2009. The automation of science. *Science*, 324(5923), pp. 85–89. Available at https://www.science.org/doi/10.1126/science.1165620, accessed December 7, 2022.

[3] Baker, M., 2016. Reproducibility crisis. *Nature* 533 (26): pp. 353–366.

[4] Al Saadi, A., Alfe, D., Babuji, Y., Bhati, A., Blaiszik, B., Brace, A., Brettin, T., Chard, K., Chard, R., Clyde, A., Coveney, P., Foster, I., Gibbs, T., Jha, S., Keipert, K., Kranzlmüller, D., Kurth, T., Lee, H., Li, Z., Ma, H., Mathias, G., Merzky, A., Partin, A., Ramanathan, A., Shah, A., Stern, A., Stevens, R., Tan, L., Titov, M., Trifan, A., Tsaris, A., Turilli, M., Van Dam, H., Wan, S., Wifling, D., and Yin, J., 2021. IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better Leads. In: *50th International Conference on Parallel Processing (ICPP 2021)*, Association for Computing Machinery, New York, NY, USA, Article 40, pp. 1–12. https://doi.org/10.1145/3472456.3473524.

[5] Clyde, A., Galanie, S., Kneller, D.W., Ma, H., Babuji, Y., Blaiszik, B., Brace, A., Brettin, T., Chard, K., Chard, R., Coates, L., Foster, I., Hauner, D., Kertesz, V., Kumar, N., Lee, H., Li, Z., Merzky, A., Schmidt, J. G., Tan, L., Titov, M., Trifan, A., Turilli, M., Van Dam, H., Chennubhotla, S.C., Jha, S., Kovalevsky, A., Ramanathan, A., Head, M.S., and Stevens, R., 2021. High-throughput virtual screening and validation of a SARS-CoV-2 main protease noncovalent inhibitor. *J. Chem. Inf. Model.*, 62(1), pp. 116–128. https://doi.org/10.1021/acs.jcim.1c00851

[6] Trifan, A., Gorgun, D., Salim, M., Li, Z., Brace, A., Zvyagin, M., Ma, H., Clyde, A., Clark, D., Hardy, D.J., Burnley, T., Huang, L., McCalpin, J., Emani, M., Yoo, H., Yin, J., Tsaris, A., Subbiah, V., Raza, T., Liu, J., Trebesch, N., Wells, G., Mysore, V., Gibbs, T., Phillips, J., Chennubhotla, S.C., Foster, I., Stevens, R., Anandkumar, A., Vishwanath, V., Stone, J.E., Tajkhorshid, E., Harris, S.A., and Ramanathan, A., 2022. Intelligent resolution: Integrating Cryo-EM with AI-driven multi-resolution simulations to observe the severe acute respiratory syndrome coronavirus-2 replication-transcription machinery in action. *The International Journal of High Performance Computing Applications,* 36(5-6). https://doi.org/10.1177/10943420221113513

[7] Bhati, A.P., Wan, S., Alfè, D., Clyde, A.R., Bode, M., Tan, L., Titov, M., Merzky, A., Turilli, M., Jha, S., Highfield, R.R., Rocchia, W., Scafuri, N., Succi, S., Kranzlmüller, D., Mathias, G., Wifling, D., Donon, Y., Di Meglio, A., Vallecorsa, S., Ma, H., Trifan, A., Ramanathan, A., Brettin, T., Partin, A., Xia, F., Duan, X., Stevens, R., and Coveney, P V., 2021. Pandemic drugs at pandemic speed: infrastructure for accelerating COVID-19 drug discovery with hybrid machine learning- and physics-based simulations on high-performance computers. *The Royal Society: Interface Focus* 11(6). https://doi.org/10.1098/rsfs.2021.0018

[8] Gasparetto, A., and Scalera, L., 2019. From the Unimate to the Delta robot: The early decades of industrial robotics. In: *Explorations in the History and Heritage of Machines and Mechanisms*, pp. 284–295, Springer, Cham.

[9] Kuipers, B., Feigenbaum, E.A., Hart, P.E., and Nilsson, N.J., 2017. Shakey: From conception to history. *Ai Magazine*, 38(1), pp. 88–103.

[10] King, R.D., et al., 2009. The robot scientist Adam. *Computer* 42(8). https://ieeexplore.ieee.org/document/5197424, accessed December 7, 2022.

[11] Williams, K., et al., 2015. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal*

Society: Interface, 12 (104). https://royalsocietypublishing.org/doi/10.1098/rsif.2014.1289, accessed December 7, 2022.

[12] Zhang, B., Merker, L., Sanin, A., and Stein, H.S., 2022. Robotic cell assembly to accelerate battery research. *Digital Discovery*, 1, pp. 755–762. https://doi.org/10.1039/D2DD00046F

[13] van der Westhuizen, C.J., du Toit, J., Neyt, N., Riley, D., and Panayides, J.-L., 2022. Use of open-source software platform to develop dashboards for control and automation of flow chemistry equipment. *Digital Discovery* 1, pp. 596–604. https://doi.org/10.1039/D2DD00036A

[14] Gongora, A.E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K.G., … and Brown, K.A., 2020. A Bayesian experimental autonomous researcher for mechanical design. *Science Advances*, 6(15), eaaz1708.

[15] Szymanski, N.J., Zeng, Y., Huo, H., Bartel, C.J., Kim, H., and Ceder, G., 2021. Toward autonomous design and synthesis of novel inorganic materials. *Materials Horizons* 8(8), pp. 2169–2198.

[16] Masubuchi, S., Watanabe, E., Seo, Y., Okazaki, S., Sasagawa, T., Watanabe, K., ... and Machida, T., 2020. Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials. *npj 2D Materials and Applications*, 4(1), pp. 1–9.

[17] Harnden, K.A., Wang, Y., Vo, L., Zhao, H., Lu, Y., 2021. Engineering artificial metalloenzymes. In: *Protein Engineering: Tools and Applications*, First Edition. Zhao, H., et al. (eds.), Wiley-VCH GmbH. https://doi.org/10.1002/9783527815128.ch8

[18] Grisoni, F., and Schneider, G., 2022. De novo molecular design with chemical language models. In: *Artificial Intelligence in Drug Design,* pp. 207–232, Humana, New York, NY.

[19] Thakkar, A., Johansson, S., Jorner, K., Buttar, D., Reymond, J.L., and Engkvist, O., 2021. Artificial intelligence and automation in computer aided synthesis planning. *Reaction chemistry & engineering* 6(1), pp. 27–51.

[20] Häse, F., Roch, L.M., Aspuru-Guzik, A., 2019. Next-generation experimentation with self-driving laboratories. *Trends in Chemistry* 1(3), pp. 282–291. https://doi.org/10.1016/j.trechm.2019.02.007

[21] Janssen, M., Falcke, H., Kadler, M., Ros, E., Wielgus, M., Akiyama, K., Baloković, M., Blackburn, L., Bouman, K.L., Chael, A., and Chan, C.K., 2021. Event horizon telescope observations of the jet launching and collimation in Centaurus A. *Nature Astronomy* 5(10), pp. 1017–1028.

[22] Gach, P.C., et al., 2016. A droplet microfluidic platform for automating genetic engineering. *ACS Synthetic Biology* 5 (5), pp. 426–433. https://doi.org/10.1021/acssynbio.6b00011

[23] Iwai, K., et al., 2018. Automated flow-based/digital microfluidic platform integrated with onsite electroporation process for multiplex genetic engineering applications. *IEEE Micro Electro Mechanical Systems (MEMS)*, pp. 1229–1232. doi: 10.1109/MEMSYS.2018.8346785

[24] Fuller, C.W., et al., 2022. Molecular electronics sensors on a scalable semiconductor chip: A platform for single-molecule measurement of binding kinetics and enzyme activity. *PNAS,* 119 (5). https://doi.org/10.1073/pnas.2112812119

[25] Cortese, A.J., et al., 2020. Microscopic sensors using optical wireless integrated circuits. *PNAS*, 117(17), pp. 9173–9179. https://doi.org/10.1073/pnas.1919677117

[26] Nie., L., et al., 2021. Quantum monitoring of cellular metabolic activities in single mitochondria. *Science Advances* 7(21). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8133708/, accessed December 7, 2022.

[27] Rienzo, M., et al., 2021. High-throughput optofluidic screening for improved microbial cell factories via real-time micron-scale productivity monitoring. *Lab on a Chip* 15. https://pubs.rsc.org/en/content/articlelanding/2021/LC/D1LC00389E, accessed December 7, 2022.

[28] Wegner, S.A., et al., 2022. The bright frontiers of microbial metabolic optogenetics. *Current Opinion in Chemical Biology* 17: 102207. https://www.sciencedirect.com/science/article/pii/S1367593122000928?via%3Dihub, accessed December 7, 2022.

[29] Wikipedia, “Standardization in Lab Automation,” last edited November 21, 2022. https://en.wikipedia.org/wiki/Standardization_in_Lab_Automation, accessed December 8, 2022.

[30] Beckman, P., Sankaran, R., Catlett, C., Ferrier, N., Jacob, R., and Papka, M., 2016. Waggle: An open sensor platform for edge computing. In: *2016 IEEE SENSORS*, pp. 1–3, Oct.

[31] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., and Ng, A.Y., 2009. ROS: An open-source Robot Operating System. In: *ICRA Workshop on Open Source Software*, 3(3.2), p. 5, May.

06. AI for Programming and Software Engineering

[1] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I., et al, 2018. Improving language understanding by generative pre-training. *OpenAI*. https://cdn.openai.com/research-covers/language-

unsupervised/language_understanding_paper.pdf, accessed November 8, 2022.

[2] Chen, M., Tworek, J., Jun, H., Yuan, Q., Ponde de Oliveira Pinto, H., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Petroski Such, F., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Hebgen Guss, W., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W., 2021. Evaluating large language models trained on code. https://arxiv.org/abs/2107.03374, accessed on November 8, 2022.

[3] Stevens, R., Taylor, V., Nichols, J., Maccabe, A.B., Yelick, K., and Brown, D., 2020. *AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science*. https://doi.org/10.2172/1604756, accessed January 10, 2023.

[4] Finkel, H., and Laguna, I., 2020. *Report of the Workshop on Program Synthesis for Scientific Computing*, August.: https://www.anl.gov/cels/program-synthesis-for-scientific-computing-report, accessed January 10, 2023.

[5] Bernholdt, D.E., Cary, J., Heroux, M., and McInnes, L.C., 2022. *The Science of Scientific-Software Development and Use*, United States. https://www.osti.gov/servlets/purl/1846008 and https://doi.org/10.2172/1846008, accessed January 10, 2023.

[6] Github, 2022. Github Copilot: Your AI pair programmer, October. https://github.com/features/copilot, accessed November 9, 2022.

[7] AWS, 2022. Amazon CodeWhisperer, October. https://aws.amazon.com/codewhisperer/, accessed November 9, 2022.

[8] Mendis, C., Renda, A., Amarasinghe, S., and Carbin, M., 2019. Ithemal: Accurate, portable and fast basic block throughput estimation using deep neural networks. In: *International Conference on Machine Learning*, pp. 4505–4515, PMLR.

[9] Bruch, M., Monperrus, M., and Mezini, M., 2009. Learning from examples to improve code completion systems. In: *Proceedings of the 7th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering*, ESEC/FSE '09, pp. 213–222, New York, New York, Association for Computing Machinery.

[10] Roziere, B., Lachaux, M.A., Chanussot, L., and Lample, G., 2020. Unsupervised translation of programming languages. In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (editors), *Advances in Neural Information Processing Systems* 33, pp. 20601–20611, Curran Associates, Inc.

[11] Chernis, B., and Verma, R., 2018. Machine learning methods for software vulnerability detection. In: *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, IWSPA '18, pp. 31–39, New York, New York, Association for Computing Machinery.

[12] Chakraborty, S., Krishna, R., Ding, Y., and Ray, B., 2022. Deep learning based vulnerability detection: Are we there yet? *IEEE Transactions on Software Engineering*, 48 (9), pp. 3280–3296.

[13] Ghofrani, J., Mohseni, M., and Bozorgmehr, A., 2017. A conceptual framework for clone detection using machine learning. In: *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, pp. 0810–0817.

[14] Allamanis, M., Brockschmidt, M., and Khademi, M., 2018. Learning to represent programs with graphs. In: *International Conference on Learning Representations*.

[15] Sachdev, S., Li, H., Luan, S., Kim, S., Sen, K., and Chandra, S., 2018. Retrieval on source code: A neural code search. In: *Proceedings of the 2nd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, MAPL 2018, pp. 31–41, New York, New York, Association for Computing Machinery.

[16] Edwards, H.C., Trott, C.R., and Sunderland, D., 2014. Kokkos: Enabling manycore performance portability through polymorphic memory access patterns. *Journal of Parallel and Distributed Computing*, *74* (12), pp. 3202–3216.

[17] Hornung, R.D., and Keasler, J.A., 2014. The RAJA portability layer: Overview and status. Lawrence Livermore National Laboratory, LLNL-TR-661403, September 24.

[18] Huerta, E.A., Khan, A., Davis, E., Bushell, C., Gropp, W.D., Katz, D.A., Kindratenko, V., Koric, S., Kramer, W.T.C., McGinty, B., McHenry, K., and Saxton, A., 2020. Convergence of artificial intelligence and high performance computing on NSF-supported cyberinfrastructure. *Journal of Big Data*, 7 (1), p. 88. https://doi.org/10.1186/s40537-020-00361-2.

[19] Koteska, B., Mishev, A., and Pejov, L., 2018. Quantitative measurement of scientific software quality:

Definition of a novel quality model. *International Journal of Software Engineering and Knowledge Engineering*, 28 (03), pp. 407–425.

[20] Storer, T., 2017. Bridging the chasm: A survey of software engineering practice in scientific programming. *ACM Comput. Surv.* 50 (4), August.

[21] Vetter, J.S., Brightwell, R., Gokhale, M., McCormick, P., Ross, R., Shalf, J., Antypas, K., Donofrio, D., Humble, T., and Schuman, C., et al., 2018. Extreme heterogeneity 2018-productive computational science in the era of extreme heterogeneity: Report for DOE ASCR Workshop on Extreme Heterogeneity, technical report, U.S. DOE Office of Science (SC), Washington, DC, (United States). https://www.osti.gov/biblio/1473756, accessed November 8, 2022.

[22] Delgado-Frias, J., Ahmed, A., and Payne, R., 1991. A dataflow architecture for AI. In: *VLSI for Artificial Intelligence and Neural Networks*, pp. 23–32, Springer.

[23] Emani, M., Vishwanath, V., Adams, C., Papka, M.E., Stevens, R., Florescu, L., Jairath, S., Liu, W., Nama, T., and Sujeeth, A., 2021. Accelerating scientific applications with sambanova reconfigurable dataflow architecture. *Computing in Science & Engineering*, 23 (2), pp. 114–119.

[24] Schuman, C.D., Kulkarni, S.R., Parsa, M., Parker Mitchell, J., and Kay, B., et al., 2022. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2 (1), pp. 10–19.

[25] Britt, K.A., Mohiyaddin, F.A., and Humble, T.S., 2017. Quantum accelerators for high-performance computing systems. In: *2017 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1–7.

[26] Möller, M., and Vuik, C., 2017. On the impact of quantum computing technology on future developments in high-performance scientific computing. *Ethics and information technology*, 19 (4), pp. 253–269.

07. SC (BER, BES, HEP, NP, FES, and Scientific User Facilities)

[1] Rosner, J., et al., 2013. *Planning the Future of U.S. Particle Physics: Chapter 1: Summary*, arXiv:1401.6075.

[2] FESAC (Fusion Energy Sciences Advisory Committee), 2022. *Powering the Future: Fusion & Plasmas*. https://science.osti.gov/-/media/fes/fesac/pdf/2020/202012/FESAC_Report_2020_Powering_the_Future.pdf, accessed December 2, 2022.

[3] Office of Science, 2016, *Exascale Requirements Review: Nuclear Physics*. https://exascaleage.org/wp-content/uploads/sites/67/2017/05/DOE-ExascaleReport_NP_R27.pdf, accessed December 2, 2022.

[4] Nachman, B., 2020. Anomaly detection for physics analysis and less than supervised learning. arXiv:2010.14554.

[5] Butter, A., and Plehn, T., 2020. Generative networks for LHC events. arXiv:2008.08558.

[6] Chen, T., et al., 2022. Interpretable uncertainty quantification in AI for HEP. arXiv: 2208.03284.

[7] Shanahan, P., et al., 2021. CompF3: Machine learning, Snowmass 2022 Report. arXiv:2209.07559.

[8] HEPML-Living Review, undated. *A living review of machine learning for particle physics*. https://iml-wg.github.io/HEPML-LivingReview/, accessed December 2, 2022.

[9] Boehnlein, A., et al., 2022. Colloquium: Machine learning in nuclear physics. *Reviews of Modern Physics* 94 (3). 10.1103/revmodphys.94.031003

[10] Hatakeyama-Sato, K., and Oyaizu, K. 2020. Integrating multiple materials science projects in a single neural network, *Communications Materials* 1(1): pp. 1–10.

[11] Dueben, P.D., et al., 2022. Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook. *Artificial Intelligence for the Earth Systems* 1(3), e210002. https://journals.ametsoc.org/view/journals/aies/1/3/AIES-D-21-0002.1.xml, accessed December 2, 2022.

[12] Carbonell, et al., 2019.

[13] Martin, H.G., et al., 2022. Perspectives for self-driving labs in synthetic biology, arXiv:2210.09085.

08. Energy (EERE, OE, FECM, NE)

[1] Bhattacharyya, A., and Hastak, M., 2022. Indirect cost estimation of winter storm–induced power outage in Texas. *Journal of Management in Engineering* 38(6), 04022057.

[2] NERC (North American Electric Reliability Corporation), 2014. Hurricane Sandy Event Analysis Report, January. https://www.nerc.com/pa/rrm/ea/Oct2012HurricanSandyEvntAnlyssRprtDL/Hurricane_Sandy_EAR_20140312_Final.pdf, accessed November 9, 2022.

[3] Argonne National Laboratory, Berkeley Lab, National Renewable Energy Laboratory, Oak Ridge National Laboratory, and Pacific Northwest National Laboratory, 2021. *Designing for Deep Decarbonization: Accelerating the U.S. Bioeconomy*. https://biosciences.lbl.gov/wp-content/uploads/2021/12/21-BAO-3054-Designing-the-Bioeconomy-for-Deep-Decarbonization-Report_v5.pdf, accessed November 9, 2022.

[4] Carbonell, P., Radivojevic, T., and García Martín, H., 2019. Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synth. Biol*. 8(7), pp. 1474–1477. https://pubs.acs.org/doi/full/

10.1021/acssynbio.8b00540, accessed November 9, 2022.

[5] DOE-SC (U.S. Department of Energy Office of Science), 2018. *Basic Research Needs for Microelectronics: Report of the Office of Science Workshop on Basic Research Needs for Microelectronics*, October 23–25. https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN_Microelectronics_rpt.pdf, accessed November 9, 2022.

[6] EPRI (Electric Power Research Institute), 2019. *An Introduction to AI, its Use Cases, and Requirements for the Electric Power Industry*, August. https://www.epri.com/research/products/000000003002017143, accessed November 9, 2022.

[7] DOE-OE, 2022. *NOTICE of INTENT: Building a Better Grid Initiative to Upgrade and Expand the Nation's Electric Transmission Grid to Support Resilience, Reliability, and Decarbonization*, 6450-01-P. https://www.energy.gov/sites/default/files/2022-01/Transmission%20NOI%20final%20for%20web_1.pdf, accessed November 9, 2022.

[8] Descour, M., Tsao, J., Stracuzzi, D., Wakeland, A., Schultz, D., Smith, W., and Weeks, J., 2019. *Workshop Report: AI-Enhanced Co-Design for Next-Generation Microelectronics: Innovating Innovation*, Sandia National Laboratories, SAND2021-16012R. https://www.osti.gov/servlets/purl/1845383, accessed November 9, 2022.

[9] Natural Resources Canada and DOE, 2006, *Final Report on the Implementation of the Task Force Recommendations*, U.S.-Canada Power System Outage Task Force, September. https://www.energy.gov/sites/default/files/oeprod/DocumentsandMedia/BlackoutFinalImplementationReport%282%29.pdf, accessed November 9, 2022.

[10] FERC (Federal Energy Regulatory Commission), NERC, and Regional Entities, 2021. *The February 2021 Cold Weather Outages in Texas and the South Central United States*, November. https://www.ferc.gov/media/february-2021-cold-weather-outages-texas-and-south-central-united-states-ferc-nerc-and, accessed November 9, 2022.

[11] Tang, L., and Ferris, M., 2015. *Collection of power flow models: Mathematical formulations*, University of Wisconsin: Madison, WI, USA.

09. Earthshots

[1] U.S. Department of Energy, 2021. *Energy Earthshots Initiative*, June. https://www.energy.gov/policy/energy-earthshots-initiative, accessed September 21, 2022.

[2] Choen, Z., 2015. The F-35: Is the world's most expensive weapons program worth it?, *CNN*, July 16.

[3] China Power Team. 2017. What do we know (so far) about China's second aircraft carrier?, China Power, Center for Strategic and International Studies, April 22, updated June 15, 2021. https://chinapower.csis.org/china-aircraft-carrier-type-001a/, accessed September 21, 2022.

[4] Kaufman, E. and Liebermann, O., 2022. US Navy's latest and most advanced aircraft carrier deploys for the first time. *CNN*, October 4.

[5] Clark, J.O., 2009. System of systems engineering and family of systems engineering from a standards, V-model, and dual-V model perspective. In: *2009 3rd Annual IEEE Systems Conference*, pp. 381–387.

[6] Barricelli, B.R., Casiraghi, E., and Fogli, D., 2019. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE Access*, 7, pp. 167653–167671.

[7] Nguyen, T., Ponciroli, R., Bruck, P., Esselman, T.C., Rigatti, J., and Vilim, R., 2022. A digital twin approach to system-level fault detection and diagnosis for improved equipment health monitoring. *Annals of Nuclear Energy*, 170, June.

[8] Zhang, X., Xie, F., Ji, T., Zhu, Z. and Zheng, Y., 2021. Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization. *Computer Methods in Applied Mechanics and Engineering*, 373, p.113485.

[9] Birchfield, A.B., Xu, T., Gegner, K.M., Shetye, K.S., and Overbye, T.J., 2017. Grid structural characteristics as validation criteria for synthetic networks. *IEEE Transactions on Power Systems*, 32(4), pp. 3258–3265.

[10] Sandu, A., and Gunther, M., 2015. A generalized-structure approach to additive Runge-Kutta methods. *SIAM Journal on Numerical Analysis*, 53(1): pp. 17–42.

[11] Mou, C., Koc, B., San, O., Rebholz, L.G., and Iliescu, T., 2021. Data-driven variational multiscale reduced order models. *Computer Methods in Applied Mechanics and Engineering*, 373:113470.

[12] Xie, X., Mohebujjaman, M., Rebholz, L.G., and Iliescu, T., 2018. Data-driven filtered reduced order modeling of fluid flows. *SIAM J. Sci. Comput.*, 40(3): B834–B857.

[13] Mou, C., Merzari, E., San, O., and Iliescu. T., 2022. A numerical investigation of the lengthscale in the mixing-length reduced order model of the turbulent channel flow. In: *Proceedings of 19th International Topical Meeting on Nuclear Reactor Thermal Hydraulics (NURETH-19)*, Brussels, Belgium, March 6–11.

10. National Nuclear Security Administration (NNSA)

[1] U.S. Department of Energy, 2021. *U.S. Department of Energy FY 2022 Congressional Budget Request*,

National Nuclear Security Administration, Office of Chief Financial Officer, Vol. 1, DOE/CF-0171, May.

[2] National Nuclear Security Administration, undated. *National Nuclear Security Administration FY 2023 Congressional Budget Justification*. https://www.energy.gov/sites/default/files/2022-04/doe-fy2023-budget-volume-1-nnsa.pdf, accessed October 18, 2022.

[3] National Nuclear Security Administration, 2022. *Accelerating Product Realization: Aligning the NNSA Nuclear Security Enterprise with Industry Best Practices*, Office of Defense Programs, Science Council, April.

[4] Ellis, J.A., Fiedler, L., Popoola, G.A., Modine, N.A., Stephens, J.A., Thompson, A.P., Cangi, A., and Rajamanickam, S., 2021. Accelerating finite-temperature Kohn-Sham density functional theory with deep neural networks. *Physical Review B* 104(3): 035120.

[5] Zuo, Y., Chen, C., Li, X., Deng, Z., Chen, Y., Behler, J., Csányi, G., et al.,2020. Performance and cost assessment of machine learning interatomic potentials. *The Journal of Physical Chemistry A* 124 (4), pp. 731–745.

[6] National Nuclear Security Administration, 2022. https://www.energy.gov/nnsa/nonproliferation, accessed Nov. 22, 2022.

[7] CNBC Technology Executive Council, 2022. How using analytics and AI can help companies manage the semiconductor supply chain. https://www.cnbc.com/2022/10/19/how-ai-can-help-companies-manage-the-semiconductor-supply-chain.html, accessed Oct. 19, 2022.

11. Software and Frameworks

[1] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32.

[2] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. {TensorFlow}: A system for {Large-Scale} machine learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283.

[3] Baydin, A.G., Pearlmutter B.A., Radul, A.A., and Siskind, J.M., 2018. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research,* 18(153), pp. 1–43. https://www.jmlr.org/papers/volume18/17-468/17-468.pdf, accessed May 12, 2023.

[4] Goodrich, C.P., King, E.M., Schoenholz, S.S., Cubuk E.D., and Brenner, M.P., 2021. Designing self-assembling kinetics with differentiable statistical physics models. In: *Proceedings of the National Academy of Sciences* 118(10), e2024083118. https://doi.org/10.1073/pnas.2024083118

[5] Krammer, M., Schuch, K., Kater, C., Alekeish, K., Blochwitz, T., Materne, S., Soppa, A., Benedikt, M., 2019. Standardized integration of real-time and non-real-time systems: The distributed co-simulation protocol. In: *Proceedings of the 13th International Modelica Conference*, pp. 87–96. https://doi.org/10.3384/ecp1915787

[6] Blochwitz, T., Otter, M., Arnold, M., Bausch, C., Clauss, C., Elmqvist, H., Junghanns, A., Mauss, J., Monteiro, M., Neidhold, T., Neumerkel, D., Olsson, H., Peetz, J.-V., and Wolf, S., 2011. The functional mockup interface for tool independent exchange of simulation models. In: *Proceedings of the 8th International Modelica Conference*. https://doi.org/10.3384/ecp11063105

[7] Heroux, M.A., McInnes, L., Li, X.S., Ahrens, J., Munson, T., Mohror, K., Turton, T., Vetter, J., and Thakur, R., 2022. *ECP Software Technology Capability Assessment Report*. https://doi.org/10.2172/1888898

[8] JAX: Composable transformations of Python+NumPy programs, v. 0.3.13, 2018. http://github.com/google/jax

[9] Baydin, A.G., Shao, L., Bhimji, W., Heinrich, L., Meadows, L., Liu, J., Munk, A., et al., 2019. Etalumis: Bringing probabilistic programming to scientific simulators at scale. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19)*, Article 29, pp. 1–24. https://doi.org/10.1145/3295500.3356180

[10] Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N.D., 2019. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(1), pp. 973–978. https://doi.org/10.48550/arXiv.1810.09538

[11] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A., 2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), pp.1–32. https://doi.org/10.18637/jss.v076.i01

[12] Salvatier J., Wiecki T.V., Fonnesbeck C., 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science,* 2, pp. e55. https://doi.org/10.7717/peerj-cs.55

[13] Schoenholz, S.S., and Cubuk, E.D., 2020. JAX, M.D.: A framework for differentiable physics. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS'20)*, Article 959, pp. 11428–11441. https://doi.org/10.48550/arXiv.1912.04232

[14] Bernholdt, D.E., Cary, J., Heroux, M., and McInnes, L.C., 2022. *The Science of Scientific-Software Development and Use*. https://doi.org/10.2172/1846008

12. Mathematics and Foundations

[1] Pion-Tonachini, L., et al., 2021. Learning from learning machines: A new generation of AI technology to meet the needs of science. *Technical Report, Preprint*: arXiv:2111.13786.

[2] Coveney, P.V., Dougherty, E.R., Highfield, R.R., 2016. Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* 374, 2080, 20160153.

[3] Lee, E.A., and Sirjani, M., 2018. What good are models? *International Conference on Formal Aspects of Component Software*. Springer, Cham.

[4] Niederer, S.A., Sacks, M.S., Girolami, M., Willcox, K., 2021. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science*, 1(5), 313–320.

[5] Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J., 2021. Unsolved problems in ml safety. *arXiv preprint,* arXiv:2109.13916.

[6] Dulac-Arnold, G., et al., 2021. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9), pp. 2419–2468.

[7] LeCun, Y., 2022. *A Path Towards Autonomous Machine Intelligence*. Version 0.9, 2, 2022-06-27.

[8] Burt, D.R., Ober, S.W., Garriga-Alonso, A., van der Wilk, M., 2020. Understanding variational inference in function-space. *arXiv preprint,* arXiv:2011.09421.

[9] Cao, S., 2021. Choose a transformer: Fourier or galerkin. *Advances in Neural Information Processing Systems*, 34, pp. 24924–24940.

[10] Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., Catanzaro, B., 2021. Efficient token mixing for transformers via adaptive Fourier neural operators. In: *International Conference on Learning Representations*.

[11] Broderick, T., et al., 2021. Toward a taxonomy of trust for probabilistic machine learning. *arXiv preprint,* arXiv:2112.03270.

[12] Gal, Y., et al. 2022. Bayesian uncertainty quantification for machine-learned models in physics. *Nature Reviews Physics,* 4(9), pp. 573–577.

[13] Abdar, M., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, pp. 243–297.

[14] Zhang, S., et al., 2021. Bayesian attention belief networks. *International Conference on Machine Learning*, PMLR.

[15] Bommasani, R., Hudson, D.A., Adeli, E. et al., 2021. On the opportunities and risks of foundation models. *arXiv preprint,* arXiv:2108.07258.

[16] Chen, Z., Liu, Y., and Sun, H., 2021. Physics-informed learning of governing equations from scarce data. *Nature communications,* 12(1), pp. 1–13.

[17] Zhuang, F., et al., 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE,* 109(1), pp. 43–76.

[18] Cuomo, S., Di Cola, V.S., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F., 2022. Scientific machine learning through physics-informed neural networks: Where we are and What's next?, *arXiv preprint,* arXiv:2201*.05624*.

[19] Karniadakis, G.E., et al., 2021. Physics-informed machine learning. *Nature Reviews Physics,* 3(6), pp. 422–440.

[20] Chakraborty, S., 2021. Transfer learning based multi-fidelity physics informed deep neural network. *Journal of Computational Physics*, *426*, 109942.

[21] Brown, T., et al., 2020. Language models are few-shot learners. *Advances in neural information processing systems,* 33, pp. 1877–1901.

[22] Wang, S., Tuor, T., Salonidis, T., Leung, K.K., Makaya, C., He, T., Chan, K., 2018. When edge meets learning: Adaptive control for resource-constrained distributed machine learning. In: *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, April, pp. 63–71.

[23] Kairouz, P., McMahan, H.B., Avent, B. et al., 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2), pp. 1–210.

[24] Speed, A., and Stracuzzi, D.J., 2020. *Research Needs for Trusted Analytics in National Security Settings*. United States.

13. AI Workflows (Edge, Center, Cloud)

[1] Beckman, P., Sankaran, R., Catlett, C., Ferrier, N., Jacob, R., and Papka, M., 2016. Waggle: An open sensor platform for edge computing. *2016 IEEE Sensors*, pp. 1–3. https://doi.org/10.1109/ICSENS.2016.7808975

[2] LeCun, Y., 2022. A path towards autonomous machine intelligence. https://openreview.net/pdf?id=BZ5a1r-kVsf, accessed May 12, 2023.

[3] Ali, A., Sharma, H., Kettimuthu, R., Kenesei, P., Trujillo, D., Miceli, A., Foster, I., Coffee, R., Thayer, J., and Liu, Z., 2022. fairDMS: Rapid model training by data and model reuse (preprint). https://doi.org/10.48550/arXiv.2204.09805

[4] Yin, J., Wang, F., and Shankar, M.,2022. Strategies for integrating deep learning surrogate models with HPC

simulation applications. United States: N. p., Web. doi:10.1109/IPDPSW55747.2022.00222

[5] da Silva, R.F., et al., 2021. A community roadmap for scientific workflows research and development. *2021 IEEE Workshop on Workflows in Support of Large-Scale Science (WORKS)*, pp. 81–90. https://doi.org/10.1109/WORKS54523.2021.00016

[6] National Academies of Sciences, Engineering, and Medicine, 2022. *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop*. https://doi.org/10.17226/26532

[7] U.S. Department of Energy (DOE), n.d. *Advanced Scientific Computing Research*. Office of Science. https://science.osti.gov/ascr, accessed May 12, 2023.

[8] Stevens, R., Taylor, V., Nichols, J., Maccabe, A.B., Yelick, K., and Brown, D., 2020. *AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science*. https://doi.org/10.2172/1604756

[9] Churchill, R.M., et al., 2021. A framework for international collaboration on ITER using large-scale data transfer to enable near-real-time analysis. *Fusion Science and Technology* 77(2), pp. 98–108, Feb. doi: 10.1080/15361055.2020.1851073

[10] Vasudevan, R., 2022. Machine learning for materials characterization and visualization, *2022 Gordon Research Conference on Computational Materials Science and Engineering*, October 6, Newry, Maine, USA. Zenodo. https://doi.org/10.5281/zenodo.7153303

14. Data Ecosystem

[1] Baracaldo, N., Chen, B., Ludwig, H., and Safavi, J.A., 2017. Mitigating poisoning attacks on machine learning models: A data provenance-based approach. In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 103–110, November.

[2] Wilkinson; M.D., Dumontier, M., Aalbersberg, I.J., et al., 2016. The FAIR guiding principles for scientific data management and stewardship, *Scientific Data,* 3(1), 160018. DOI: 10.1038/SDATA.2016.18

[3] Dunning, A., De Smaele, M., and Böhmer, J., 2017. Are the FAIR data principles fair?, *International Journal of Digital Curation*, 12(2), pp. 177–195.

[4] Ravi, N., Chaturvedi, P., Huerta, E.A., Liu, Z., Chard, R., Scourtas, A., Schmidt, K.J., Chard, K., Blaiszik, B., and Foster, I., 2022. Fair principles for AI models, with a practical application for accelerated high energy diffraction microscopy, *arXiv preprint*, arXiv:2207.00611.

[5] Research Data Alliance (RDA). https://www.rd-alliance.org/, accessed October 5, 2022.

[6] Orr, L., Goel, K., Ré, C., 2022. Data management opportunities for foundation models. *12th Annual Conference on Innovative Data Systems Research (CIDR '22)*, January 9–12, Santa Cruz, Calif., USA.

[7] Thirumuruganathan, S., Tang, N., Ouzzani, M., Doan, A., 2020. Data curation with deep learning, In: *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT)*, March 30–April 2, Copenhagen, Denmark.

[8] Hiszpanski, A.M., Gallagher, B., Chellappan, K., Li, P., Liu, S., Kim, H., Han, J., Kailkhura, B., Buttler, D.J., and Han, T.-Y., 2020. Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge, *Journal of Chemical Information and Modeling,* 60(6), pp. 2876–2887. https://pubs.acs.org/doi/10.1021/acs.jcim.0c00199, accessed October 12, 2022.

[9] Yang, H., et al., 2019. Pipelines for procedural information extraction from scientific literature: Towards recipes using machine learning and data science, In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2., IEEE.

15. AI-Oriented Hardware Architectures

[1] Roberts, H., Cowls, J., Morley, J., et al., 2021. The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society*, 36, pp. 59–77. https://doi.org/10.1007/s00146-020-00992-2

[2] Top500., n.d. List Statistics. *Statistics*. https://top500.org/statistics/list/, accessed May 12, 2023.

[3] Special Competitive Studies Project, 2022. *Mid-Decade Challenges to National Competitiveness*. https://www.scsp.ai/wp-content/uploads/2022/09/SCSP-Mid-Decade-Challenges-to-National-Competitiveness.pdf, accessed May 12, 2023.

[4] Wang, B., et al., 2020. Multi-physics-resolved digital twin of proton exchange membrane fuel cells with a data-driven surrogate model. *Energy and AI 1,* p. 100004. https://doi.org/10.1016/j.egyai.2020.100004

[5] Yin, J., Wang, F., and Shankar M., 2022. Strategies for integrating deep learning surrogate models with HPC simulation applications. In: *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 01–10. https://doi.org/10.1109/IPDPSW55747.2022.00222

[6] Blanchard, A.E., et al., 2021. Language models for the prediction of SARS-CoV-2 inhibitors. *International Journal of High Performance Computing Applications*, preprint. https://doi.org/10.1101/2021.12.10.471928

[7] Tripathy, R.K., and Bilionis, I., 2018. Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of*

Computational Physics, 375, pp: 565–588. https://doi.org/10.1016/j.jcp.2018.08.036

[8] Tang, Y., et al., 2020. Uncertainty-aware score distribution learning for action quality assessment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp: 9836–9845. https://doi.org/10.1109/CVPR42600.2020.00986

[9] Tsoutsouras, V., et al., 2021. The LAPLACE microarchitecture for tracking data uncertainty and its implementation in a RISC-V processor. In: *Proceedings of 54th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 1254–1269. https://doi.org/10.1145/3466752.3480131

[10] Devlin, J., et al., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*: https://arxiv.org/abs/1810.04805v2

[11] Lian, X., et al., 2022. Persia: An open, hybrid system scaling deep learning-based recommenders up to 100 trillion parameters. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3288–3298. https://doi.org/10.1145/3534678.3539070

[12] Lake, B., Ullman, T., Tenenbaum, J., and Gershman, S., 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, E253. https://doi.org/10.1017/S0140525X16001837

[13] Rahimi, A., et al., 2017. High-dimensional computing as a nanoscalable paradigm. In: *IEEE Transactions on Circuits and Systems I: Regular Papers*, 64(9), pp. 2508–2521. https://doi.org/10.1109/TCSI.2017.2705051

[14] Messina, P., 2017. The exascale computing project. *Computing in Science & Engineering,* 19(3), pp. 63–67. https://doi.org/10.1109/MCSE.2017.57

[15] de Supinski, B.R., et al, 2022. *Non-Proprietary Companion to the Q1CY2021 Path Forward Final Assessment WBS 2.4.1, Milestone PM-HI-1040*. https://doi.org/10.2172/1845203

[16] AMD, 2022. AMD Details Strategy to Drive Next Phase of Growth Across $300 Billion Market for High-Performance and Adaptive Computing Solutions (news release), June 9. https://ir.amd.com/news-events/press-releases/detail/1078/amd-details-strategy-to-drive-next-phase-of-growth-across.

[17] Chiang, H.L., et al., 2020. Cold CMOS as a power-performance-reliability booster for advanced FinFETs. In: *Proceedings of 2020 IEEE Symposium on VLSI Technology*, pp. 1–2. https://doi.org/10.1109/VLSITechnology18217.2020.9265065

[18] Saligram, R., et al., 2021. Power performance analysis of digital standard cells for 28 nm bulk CMOS at cryogenic temperature using BSIM models. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits,* 7(2), pp. 193–200. https://doi.org/10.1109/JXCDC.2021.3131100

[19] Murray, C., et al., 2018. *Basic Research Needs for Microelectronics: Report of the Office of Science Workshop on Basic Research Needs for Microelectronics*, October 23–25. https://doi.org/10.2172/1616249

[20] Ang, J.A., Chien, A.A., Hammond, S.D., et al., 2021. *Reimagining Co-design for Advanced Scientific Computing: Report for the ASCR Workshop on Reimagining Co-design*. https://doi.org/10.2172/1822199

[21] The White House, 2022. Fact Sheet: CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China, Statements and Releases, August 9. https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china/, accessed May 12, 2023.

[22] Vetter, J.S., et al., 2018. *Extreme Heterogeneity 2018—Productive Computational Science in the Era of Extreme Heterogeneity: Report for DOE ASCR Workshop on Extreme Heterogeneity*. https://doi.org/10.2172/1473756

16. Workforce and Ethics

[1] National Laboratory Directors' Council, 2022, Demographic Data for the National Lab, October, https://nationallabs.org/staff/diversity/, accessed May 12, 2023.

[2] USA Facts, 2022. Our Changing Population: United States, October, https://usafacts.org/data/topics/people-society/population-and-demographics/our-changing-population, accessed May 12, 2023..

[3] Alexander, S., 2016. *The Jazz of Physics: The Secret Link Between Music and the Structure of the Universe*. New York, NY: Basic Books.

[4] Leung, M.A., Rouson, D., and Curfman-McInnes, L., 2020. Increasing productivity by broadening participation in scientific software communities. *2020 Collegeville Workshop on Scientific Software*, July 21–23. https://collegeville.github.io/CW20/WorkshopResources/WhitePapers/leung-broadening-participation-cse-hpc.pdf, accessed May 12, 2023.

[5] Hofstra, B., et al., 2020. The diversity-innovation paradox in science. In: *Proceedings of the National Academy of Science USA*, 117(17), pp. 9284–9291. http://doi.org/10.1073/pnas.1915378117

[6] National Laboratories Directors' Council, 2022. The National Laboratories STEM Resources.

https://nationallabs.org/our-labs/stem-resources/, accessed Nov. 28, 2022.

[7] Leung, M.A., 2020. Diversity and inclusion through leadership during challenging times. *Computing in Science and Engineering*, 22(6), pp. 92–96.

[8] Leung, M.A., Crivelli, S., and Brown, D., 2019. Sustainable research pathways: Building connections across communities to diversify the national laboratory workforce. Presented at: *Collaborative Network for Engineering and Computing Diversity (CoNECD)*, April 14–17, Crystal City, VA. https://monolith.asee.org/public/conferences/148/papers/24706/view, accessed May 12, 2023.

[9] The National GEM Consortium. https://www.gemfellowship.org/, accessed May 12, 2023.

[10] Whitney, T., and Taylor, V., 2018. Increasing women and underrepresented minorities in computing: The landscape and what you can do. *Computer*, 51, pp. 24–31. http://doi.org/10.1109/MC.2018.3971359

[11] Clyde, A., 2022. AI for science and global citizens. Patterns, 3(2), 100466. https://doi.org/10.1016/j.patter.2022.100446

[12] Sustainable Horizons Institute. https://shinstitute.org, accessed May 12, 2023.

[13] Level Playing Field Institute (LPFI) SMASH program. https://www.smash.org/about/our-story/, accessed May 12, 2023.

[14] CSforAll. https://www.csforall.org/, accessed May 12, 2023.

17. Scale

[1] Gil, Y., Selman, B., chairs, 2019. *A 20-year community roadmap for Artificial Intelligence research in the U.S., A report from the Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI)*, Washington, D.C.

[2] Exascale Computing Project (ECP). https://www.exascaleproject.org/, accessed September 25, 2022.

[3] Top500. https://www.top500.org/. Updated semi-annually.

[4] Liu, S., Zhang, P., Lu, D., et al., 2022. PI3NN: Out-of-distribution-aware prediction intervals from three neural networks. In: *Proceedings of 10th International Conference on Learning Representations (ICLR),* April 25–April 29, virtual conference, arXiv:2108.02327

[5] U.S. Department of Energy, Office of Electricity, 2019. *Smart Grid System Report: 2018 Report to Congress,* Washington, D.C.

[6] Bonawitz, K., Ivanov, V., Kreuter, B., et al., 2017. Practical secure aggregation for privacy-preserving machine learning. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.* October 30–November 3, Dallas, TX, pp. 1175–1191.

[7] LeCun, Y., 2019. Deep learning hardware: past, present, and future. In: *Proceedings of the 2019 IEEE International Solid-State Circuits Conference (ISSCC).* February 17–21, San Francisco, CA, pp. 12–19. DOI: 10.1109/ISSCC.2019.8662396.

[8] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. DOI: 10.1038/SDATA.2016.18.

[9] Blaiszik, B., Ward, L., Schwarting, M., et al., 2019. A data ecosystem to support machine learning in materials science. *MRS Communications,* 9, pp. 1125–1133, arXiv:1904.10423.

[10] National Research Council, 2013. *Frontiers in Massive Data Analysis, a* report by the National Research Council of the National Academies, Washington, D.C., National Academies Press. DOI: 10.17226/18374.

[11] Energy Sciences Network (ESnet). https://www.es.net/, accessed September 14, 2022.

[12] Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L., 2016. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5), pp. 637–646. DOI: 10.1109/JIOT.2016.2579198, accessed October 11, 2019.

[13] Lopez, P., Montresor, A., Epema, D., et al., 2015. Edge-centric computing: Vision and challenges. *ACM SIGCOMM Computer Communication Review,* 45(5), pp. 37–42. DOI: 10.1145/2831347.2831354, accessed September 11, 2022.

[14] Kalyan, A., Mohta, A., Polozov, O., et al., 2018. Neural-guided deductive search for real-time program synthesis from examples. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR),* April 30–May 3, Vancouver, BC, Canada.

[15] Chang, M.-C., Wei, Y., Chen, W.-R., et al., 2019. Accelerating neutron scattering data collection and experiments using AI deep super-resolution learning. arXiv:1904.08450.

[16] U.S. Department of Energy, Basic Energy Sciences Advisory Committee, 2015. *Challenges at the Frontiers of Matter and Energy: Transformative Opportunities for Discovery Science,* a report from the Basic Energy Sciences Advisory Committee, Washington, D.C.

[17] Adams, P., Ankner, J.F., Anovitz, L., et al., 2019. *First Experiments: New Science Opportunities at the*

Spallation Neutron Source Second Target Station, Oak Ridge National Laboratory (ORNL) Technical Report ORNL/SPR-2019/1407, DOI: 10.2172/1784183.

18. Computational Resources

[1] TOP500: The List, undated. https://www.top500.org/, accessed May 12, 2023.

[2] Frontier [supercomputer]. Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory, Oak Ridge, TN. https://www.olcf.ornl.gov/frontier/, accessed May 12, 2023.

[3] Summit [supercomputer]. Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory, Oak Ridge, TN. https://www.olcf.ornl.gov/summit/, accessed May 12, 2023.

[4] Sierra [supercomputer]. Lawrence Livermore National Laboratory, Livermore, CA. https://asc.llnl.gov/sites/asc/files/sierra-fact-sheet.pdf, accessed May 12, 2023.

[5] Perlmutter: High Performance Computing Optimized for Science. National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA. https://perlmutter.carrd.co/, accessed May 12, 2023.

[6] Polaris [supercomputer]. Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL. https://www.alcf.anl.gov/polaris, accessed May 12, 2023.

[7] Van Essen, B., et al., 2015. LBANN: Livermore Big Artificial Neural Network HPC toolkit. In: *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments* (MLHPC '15), Article 5, pp. 1–6. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2834892.2834897.

[8] Cerebras [company website]. https://www.cerebras.net/, accessed May 12, 2023.

[9] SambaNova Systems [company website]. https://sambanova.ai/, accessed May 12, 2023.

[10] Graphcore [company website]. https://www.graphcore.ai/, accessed May 12, 2023.

[11] Groq [company website]. https://groq.com/, accessed May 12, 2023.

[12] Habana [company website]. https://habana.ai/, accessed May 12, 2023.

[13] ALCF AI Testbed. Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL. https://www.alcf.anl.gov/alcf-ai-testbed, accessed May 12, 2023.

[14] About Livermore Computing. Lawrence Livermore National Laboratory, Livermore, CA. https://hpc.llnl.gov/about-us, accessed May 12, 2023.

[15] NVIDIA Data Center GPUs: The Heart of the Modern Data Center. NVIDIA, Santa Clara, CA. https://www.nvidia.com/en-us/data-center/data-center-gpus/, accessed October 14, 2022.

[16] AMD Accelerators. Advanced Micro Devices, Santa Clara, CA. https://www.amd.com/en/accelerators, accessed May 12, 2023.

[17] LC Cloud Services. Livermore Computing, Lawrence Livermore National Laboratory, Livermore, CA. https://hpc.llnl.gov/cloud/, accessed May 12, 2023.

[18] Evaluation Testbeds. Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL. https://www.alcf.anl.gov/alcf-resources/evaluation-testbeds, accessed May 12, 2023.

[19] Trader, T., 2021. Berkeley Lab debuts Perlmutter, world's fastest AI supercomputer. *HPCwire*, May 21. https://www.hpcwire.com/2021/05/27/nersc-debuts-perlmutter-worlds-fastest-ai-supercomputer/, accessed October 14, 2022.

[20] Beckman, P., et al., 2016. Waggle: An open sensor platform for edge computing. In: *2016 IEEE SENSORS*, pp. 1–3. https://doi.org/10.1109/SENSORS34402.2016

[21] Smith, S., et al., 2022. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. arXiv:2201.11990 [cs.CL], v1 submitted January 28. https://doi.org/10.48550/arxiv.2201.11990

[22] Liu, Y.(A.), Liu, X.(L.), Li, F.(N.), et al., 2021. Closing the "quantum supremacy" gap: Achieving real-time simulation of a random quantum circuit using a new Sunway supercomputer. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (SC '21), Article 3, pp. 1–12. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3458817.3487399

[23] Jia, W., Wang, H., Chen, M., et al., 2020. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. arXiv:2005.00223 [physics.comp-ph], v1 submitted May 1. https://doi.org/10.48550/arxiv.2005.00223

[24] NVIDIA DGX Systems, NVIDIA, Santa Clara, CA. https://www.nvidia.com/en-us/data-center/dgx-systems/, accessed May 12, 2023.

[25] Sunway TaihuLight [supercomputer operated by National Supercomputing Center in Wuxi, Jiangsu, China]. 2022. Wikipedia. https://en.wikipedia.org/wiki/Sunway_TaihuLight, accessed October 14.

19. Data Infrastructure

[1] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

[2] Dart, E., Rotman, L., Tierney, B., Hester, M., and Zurawski, J., 2013. The science DMZ: A network design pattern for data-intensive science. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pp. 1–10.

[3] Chard, K., Dart, E., Foster, I., Shifflett, D., Tuecke, S., and Williams, J., 2018. The modern research data portal: A design pattern for networked, data-intensive science. *PeerJ Computer Science,* 4, e144.

[4] Byna, S., Idreos, S., Jones, T., Mohror, K., Ross, R., and Rusu, F., 2022. *Management and storage of scientific data*, United States. https://doi.org/10.2172/1845705 and https://www.osti.gov/servlets/purl/1845705, accessed January 10, 2023.

[5] Stach, E., DeCost, B., Kusne, A.G., Hattrick-Simpers, J., Brown, K.A., Reyes, K.G., Schrier, J., et al., 2021. Autonomous experimentation systems for materials development: A community perspective. *Matter,* 4(9), pp. 2702–2726.

[6] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, A., von Arx, S., Bernstein, M.S., et al., 2021. On the opportunities and risks of foundation models. *arXiv preprint*, arXiv:2108.07258.

[7] Wilkinson, M.D., Dumontier, M., Jan Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3(1), pp. 1–9.

[8] Ward, L., Babinec, S., Dufek, E.J., Howey, D.A., Viswanathan, V., Aykol, M., Beck, D.A.C., et al., 2022. Principles of the battery data genome. *Joule* 6(10), pp. 2253–2271.

[9] Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R., 2018. Polymer genome: A data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C,* 122(31), pp. 17575–17585.

[10] Vescovi, R., Chard, R., Saint, N., Blaiszik, B., Pruyne, J., Bicer, T., Lavens, A., et al., 2022. Linking scientific instruments and computation: Patterns, technologies, experiences. *Patterns*.

www.energy.gov