Learning to Scale the Summit: AI for Science on a Leadership Supercomputer

Wayne Joubert, Bronson Messer, Philip C. Roth, Antigoni Georgiadou,
Justin Lietz, Markus Eisenbach and Junqi Yin
National Center for Computational Sciences
Oak Ridge National Laboratory
Oak Ridge, TN USA
{joubert, bronson, rothpc, georgiadoua, lietzjg, eisenbachm, yinj}@ornl.gov

***Abstract*—The Summit system at Oak Ridge National Laboratory (ORNL) has been the world's top AI for science supercomputer for several years, ranked world's fastest computer at its 2018 launch and currently top system in the US and #2 on the TOP500 list. Summit's purposeful design to handle both conventional modeling and simulation science and emerging AI workloads has made it a leading destination for AI-powered computational science. We report here on AI for science usage on Summit near the midpoint of its lifespan. We review AI usage across the many science projects that have used Summit. We then examine in detail a set of applications scaling AI to full system as well as projects implementing AI-coordinated science discovery workflows on Summit. Finally, we offer some observations regarding the future of advancing scientific knowledge and understanding via AI, especially in the context of leadership-class scientific computing.**

***Index Terms*—HPC, high performance computing, AI, artificial intelligence, machine learning**

I. Introduction

Artificial intelligence (AI) and machine learning (ML) methods are bringing revolutionary changes across many aspects of daily life, fueled by advancements in image classification, speech recognition, natural language processing and robotics, to name a few. The impacts are also being felt in scientific discovery, and computational scientists are incorporating AI methods and techniques throughout their workflows. Accordingly, AI is increasingly a focus of public-sector science funding [1]. The US Department of Energy (DOE) has hosted multiple white papers and town halls on AI for science, particularly in connection with high performance computing (HPC) [2]–[5].

Because of the remarkable success of AI and ML in many applications, there has been an explosion of optimism that these techniques will, in fact, supplant other methodologies in computational science. For example, much attention accrued to the recent use of AlphaFold to make a large step forward in the protein folding problem [6]. But, AlphaFold also serves as a striking example of how AI successes can be overstated. AlphaFold did not "solve" the problem of protein folding: roughly a third of the model's predictions were found to be not accurate enough and, more importantly, it does not provide insight into the relationship between protein structure and function. Nevertheless, the predictive power of AlphaFold and many other AI models is considerable. This fact alone makes their possible use very attractive for a wide variety of computationally intensive problems like those attacked by leadership computing.

AI/ML projects have long been active at the Oak Ridge Leadership Computing Facility (OLCF). Many-node machine learning jobs ran on Titan at least as early as 2015 ([7]; cf. [8]). Summit [9] was designed to handle AI workloads, supported by an Infiniband fat-tree interconnect with adaptive routing, on-node burst buffers, and NVIDIA Volta graphics processing units (GPUs) with Tensor Cores providing over 3 AI-ExaOps mixed precision peak performance. Summit debuted at #1 on the TOP500 list [10] in June 2018 and remains one of the top two computing systems in the world and most powerful system in the United States.

Since its deployment, Summit has provided over 90% of the compute cycles for DOE Office of Science allocation programs, in support of many science domains. Now near the midpoint of its lifespan, Summit's history provides a wealth of information on how scientists use AI/ML methods at scale.

Our purpose is to survey AI/ML usage on Summit from mid- to late-2018 to the present. Our interest is to understand the varieties of AI/ML usage by examining all projects across all allocation programs, with particular attention to AI/ML methods that scale-out to large portions of Summit.

This paper is organized as follows. After reviewing systems and allocation programs to be studied, we outline the study methodology. We then analyze AI/ML usage across all projects, years and allocation programs. After this we look more deeply at AI/ML scale-out projects at large numbers of Summit nodes. This is followed by case studies of projects using AI to coordinate science discovery workflows on Summit. Finally we summarize findings and discuss requirements for future use of AI/ML methods in future scientific projects using HPC.

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

II. Scope and Methodology

A. Systems

The OLCF Summit system consists of over 4600 IBM Power System AC922 nodes [9]. Each of Summit's compute nodes contains six NVIDIA Tesla V100 GPUs and two 22-core IBM POWER9 processors, connected using NVIDIA's high performance NVLINK connections within the node. (One POWER9 core of each processor is reserved for the system, leaving 42 cores per node to run user processes.) All Summit nodes are interconnected with a dual-rail EDR InfiniBand fabric with a non-blocking fat tree topology. Each of Summit's original 4,608 compute nodes contains an aggregate of 96 GB of high bandwidth memory (HBM2) on the GPUs, 512 GB of DDR host memory, and 1.6 TB of non-volatile memory. In Summer 2020, the OLCF added 54 "high memory nodes," each with 192 GB of HBM2 memory, 2 TB of DDR4 memory, and 6.4 TB of non-volatile memory. The hardware configuration and queue policies of these high memory nodes were crafted for even better support of AI/ML workloads than in the original system.

Scientific workflows often include pre- or post-processing steps that do not require the extreme capability of Summit. To support these workflow steps, the OLCF fields a companion commodity Linux cluster with access to the same file systems as Summit. When Summit was first deployed this role was served by Rhea, a cluster with a 512-node CPU partition and 9-node GPU partition. Each CPU partition node contained two 8-core Intel Xeon processors and 128 GB of memory. Each GPU partition node contained two 14-core Intel Xeon processors, 1 TB of host memory, and two NVIDIA K80 GPUs. In late 2020, the OLCF replaced Rhea with the 704-node Andes cluster. Most Andes nodes contain two 16-core AMD EPYC processors with 256 GB memory, but the nine GPU nodes from Rhea have also been incorporated into Andes.

B. Allocation Programs

The OLCF allocates time on leadership resources through three primary allocation programs: the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program, the Advanced Scientific Computing Research Leadership Computing Challenge (ALCC) program, and the Center's Director's Discretionary (DD) Program. The OLCF seeks to enable scientific productivity via capability computing through all of these programs. Accordingly, a set of criteria is considered in making allocations, including the strategic impact of the expected scientific results and the degree to which awardees can effectively use leadership resources. The INCITE program, co-managed by the LCF sites at Oak Ridge and Argonne, is allotted roughly 60% of the available compute hours at OLCF in a given allocation year. INCITE review and selection incorporate a rigorous computational readiness review separate from the scientific merit review that is performed later. The ability and need to take advantage of the full capability afforded by leadership resources are the primary criteria for this review. ALCC—accounting for roughly 20% of the allocable cycles per year—has no formal computational readiness review, but the proposals are assessed by the DOE Office of Advanced Scientific Computing Research (ASCR) for their appropriateness for leadership resources. The DD program allocates the remaining 20% of resources per year. The goals of the DD program include enabling users to prepare for leadership computing competitions, such as INCITE and ALCC, and broadening the community of researchers capable of using leadership computing. Included in this mix are projects that come to the OLCF through the Accelerating Competitiveness through Computational Excellence (ACCEL) Industrial HPC Partnerships outreach, which encourages opportunities for industrial researchers to access the leadership systems to perform research that would not otherwise be possible. In addition, for the years examined here, the DD program allocated up to half of the available time (i.e., 10% of the total) to Exascale Computing Project (ECP) teams to enable GPU porting and scalability testing.

C. Study Methodology

We consider INCITE (2019-2022), ALCC (mid-2019 thru mid-2022) and DD (2019-2021) projects. We also consider projects from the ECP [11], the COVID-19 HPC Consortium [12] (some of which are also DD projects), and Association for Computing Machinery (ACM) Gordon Bell competition finalists (2018-2021) [13], treated here separately.

Our primary information source is the set of project proposals submitted by each principal investigator (PI) and project team. INCITE and ALCC proposals each typically contain about ten pages of narrative, while DD proposals are generally about a page long. In some cases we use other artifacts such as referenced publications or discussions with project computational liaisons in the OLCF Science Engagement section.

Though we only consider Summit-based projects, some projects also use Andes or Rhea for supportive AI/ML analytics work, in a way not always clearly differentiated in proposals. Because of this, we include usage of these systems also when considering whether a project uses AI/ML.

We measure AI/ML usage either by number of projects or by total allocation hours summed across relevant projects. "Allocation hours" refers to the number of Summit node-hours granted to the project at the onset of the project period. Alternatively, one could consider actual hours used or some measure of compute cycles devoted to AI/ML, though this is beyond the scope of this study. Furthermore, this could be misrepresentative, as it could undercount the importance of an ML method that greatly reduces runtime for a science calculation but itself actually requires very little compute time.

A fundamental question is, what constitutes the use of AI/ML in a project? A poorly chosen criterion could distort the results by being too inclusive. As an extreme example, an application could use a vendor linear algebra library which has been, unknown to project members, autotuned using ML methods, and thus be classified as using AI. Or a project could use potentials in a molecular dynamics library that are trained with ML, either unknown to project personnel or so routinized

TABLE I
SCIENCE APPLICATION AI MOTIFS

Motif	Definition	Example
fault detection	detect algorithmic or other failure in execution, send signal for automatic or manual remediation	detect simulation defect caused by execution error
math/cs algorithm	ML is used to enhance some mathematical (non-science-proper) computation	solver's linear system dimension is reduced based on machine-learned parameter
submodel	a (proper) subset of a science computation is replaced by an ML model. molecular dynamics (MD) potentials as special case	physics-based radiation model in a climate code replaced by ML model
steering	automatic steering of the direction of a computation for some internal process	ML method to guide Monte Carlo sampling to include undersampled regions
surrogate model	full science model replaced by ML approximation that captures important aspects, used for speed or science understanding	data from tokamak simulation runs used to train surrogate model
analysis	results from modeling and simulation (modsim) runs are analyzed by a human using ML methods	use graph neural networks to analyze results of MD simulation
ML + modsim loop	both ML and traditional modsim, coupled	MD in loop used to refine deep learning model via active learning
classification	"pure" ML with little or no modsim used to classify some phenomenon; includes some other methods like reinforcement learning	deep neural network inference to detect rare astrophysical event
various	umbrella project with multiple unrelated subprojects using possibly different kinds of AI/ML	CAAR/ESP/NESAP application readiness
undetermined	manner of AI/ML use is undetermined	project is exploring AI/ML use but gives no details

as to have become standard community practice. To avoid such cases, we consider a project to use AI/ML if project personnel have made a conscious decision to implement or use AI/ML and have explicitly indicated this choice.

For AI/ML usage or adoption status, we consider three cases. First, "active" refers to actual usage of AI/ML in the project and given project year. Second, "inactive" refers to usage in a previous project year, planned or possible future use, exploring possibility of use, or usage in a closely-linked companion project. Finally, "none" refers to no serious mention of or interest in AI/ML methods.

Various attempts have been made to categorize the ways AI/ML methods can be used for science, for example "in-the-loop," "on-the-loop" or "around-the-loop" configurations [14]. To add slightly more granularity to this, we define "AI motifs," derived directly from review of the many projects (see Table I). The aim is to enable at-a-glance visibility into what ways scientists incorporate ML methods into their codes. It is recognized that this categorization, though necessary, is inherently somewhat approximate. A project's use of AI/ML could be interpreted to reflect more than one category, or (less commonly) the project could use AI/ML in multiple different ways. In such cases, we select the most prominent category. For example, a project using analysis of results to build a submodel would be classed as "submodel" not "analysis" if the only purpose of the analysis is to build the submodel. It should be emphasized that this taxonomy is tentative and is subject to refinement in the future.

Finally, we study AI/ML usage with respect to science domain. Every OLCF project initially receives one of 48 different 3-letter codes to denote its science subdomain. These are sometimes grouped into a shorter list of nine science domains. Sometimes the subdomain is selected by the project PIs, in other cases OLCF personnel. Science subdomains and domains can refer either to the science problem being solved (e.g., turbulence) or the application area (e.g., engineering). Some projects could conceivably be given multiple designations; for example, a single project could have aspects of turbulence, chemistry, combustion and engineering. For consistency and clarity, in this study we have adjusted the science domain categories and subdomain assignments for OLCF projects in a few cases, to better represent the most prominent scientific theme of each project. Our list is shown in Table II.

TABLE II
SCIENCE DOMAINS AND SUBDOMAINS

Domain	Subdomains
Biology	Bioinfomatics, Biophysics, Life Sciences, Medical Science, Neuroscience, Proteomics, Systems Biology
Chemistry	Chemistry, Physical Chemistry
Computer Science	Computer Science, Machine Learning
Earth Science	Atmospheric Science, Climate, Geosciences, Geographic Information Systems
Engineering	Aerodynamics, Bioenergy, Combustion, Engineering, Fluid Dynamics, Turbulence
Fusion and Plasma	Fusion Energy, Plasma Physics
Materials	Materials Science, Nanoelectronics, Nanomechanics, Nanophotonics, Nanoscience,
Nuclear Energy	Nuclear Fission, Nuclear Fuel Cycle
Physics	Accelerator Physics, Astrophysics, Cosmology, Atomic/Molecular Physics, Condensed Matter Physics, High Energy Physics, Lattice Gauge Theory, Nuclear Physics, Physics, Solar/Space Physics

III. Summit AI/ML Usage Across Programs

We now examine AI/ML usage across all projects, totaling 662 project-years (INCITE 147, ALCC 72, DD 352, COVID non-DD 12, ECP 62, Gordon Bell finalist 17). Gordon Bell finalist projects are analyzed separately in Section IV.

A. *Overall AI/ML usage*

Figure 1 shows overall adoption of AI/ML for Summit projects. For this and following subsections we examine usage across all INCITE, ALCC and DD years as well as ECP projects and also COVID projects that do not overlap with DD, to avoid double-counting. Here we see a substantial number of projects, 1/3 over Summit's lifespan, have actively used AI/ML methods, with another 8% indirect use.

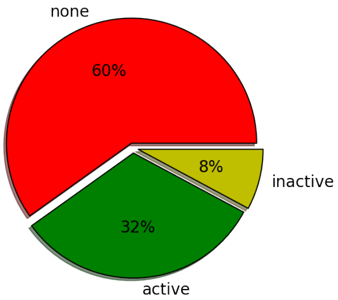
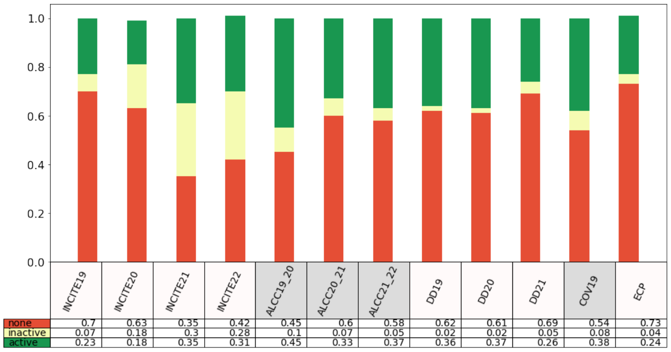


Fig. 1. Overall AI/ML usage, percentage of projects.

B. AI/ML usage by program and year

Figure 2 breaks down this usage by allocation program and year. AI/ML adoption in INCITE, the largest allocation program, has grown steadily from 20% in 2019. ALCC usage has been significant, especially in 2019-20 when a large subset of a smaller number of projects used AI/ML. Note ALCC projects tend to make use of long-established codes and methods, since the program is designed to support specific aims of DOE-supported researchers, unlike INCITE which is more adaptable over time to new communities. DD for every year has a very large number of projects, many using AI/ML. ECP projects understandably use AI/ML less, being more constrained by project goals set early in the program. COVID-19 projects use AI/ML heavily for drug discovery and others.



	INCITE19	INCITE20	INCITE21	INCITE22	ALCC19_20	ALCC20_21	ALCC21_22	DD19	DD20	DD21	COV19	ECP
none	0.7	0.63	0.35	0.42	0.45	0.6	0.58	0.62	0.61	0.69	0.54	0.73
inactive	0.07	0.18	0.3	0.28	0.1	0.07	0.05	0.02	0.02	0.05	0.08	0.04
active	0.23	0.18	0.35	0.31	0.45	0.33	0.37	0.36	0.37	0.26	0.38	0.24

Fig. 2. AI/ML usage by program and year, percentage of projects.

C. AI/ML usage by ML method

Figure 3 shows ML method used, whether deep learning (DL, DNN) or other neural network, or otherwise. We aggregate active and inactive projects. For some projects, such as those only planning to use ML, it was not possible to determine the method used.

Note DL/NN methods are much more prevalent than others. Of the latter, many methods are used, such as SVM, isolation forests, PCA, weighted least squares linear regression, Bayesian regularized regression, multiparameter regression or boosted decision tree regression. DL methods are attractive for their versatility and scalability.

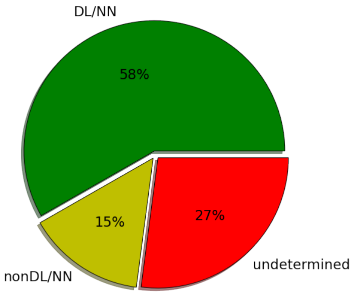


Fig. 3. Usage by AI/ML method, percentage of projects.

D. AI/ML usage by science domain

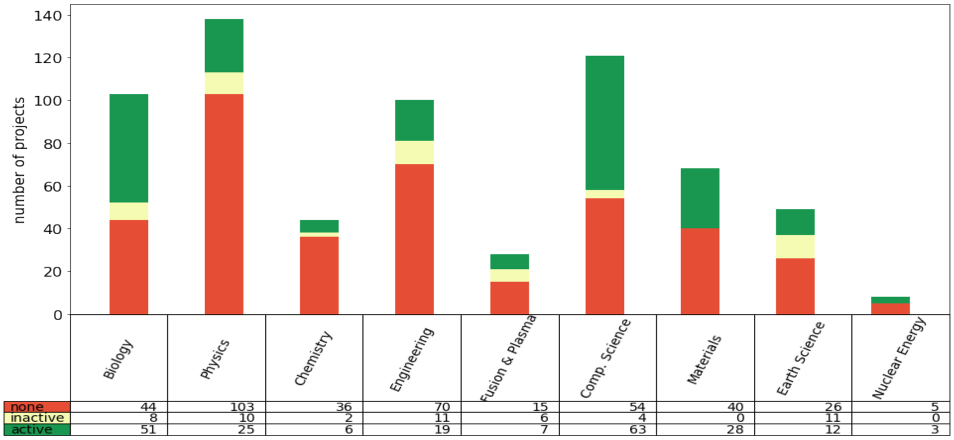
Figure 4 shows results from Figure 1 broken down by science domain. Usage is highly domain-specific. Computer Science has many ML-proper projects, thus high adoption. Biology is a heavy ML user for drug discovery, genomics, COVID-19 research and others. Engineering, Earth Science and Fusion/Plasma, commonly using grids, have significant adoption as well as notable "inactive" use, often reflecting efforts to validate ML models. Materials projects commonly use ML methods to model atomic interactions. Chemistry is represented indirectly under Biology and Materials.

E. AI/ML usage by AI motif

Figure 3 shows AI/ML usage broken down by AI motif. For this and the next subsection we aggregate active and inactive projects and consider only INCITE, ALCC and ECP, for which abundant information is available. The top motif is Submodels, reflecting incorporation of ML models into simulation codes. This with Classification, Analysis, Surrogate Models and MD Potentials account for over 3/4 of usage. AI coordination methods like Steering and ML+Modsim Loop are expected to increase going forward.

F. AI motif vs. science domain

Figure 6 gives breakdown by science domain. Conclusions should be drawn cautiously here on account of small sample size per category. The most prominent usage is Submodels by Engineering. Submodels are also used in other domains like Earth Science often making use of grids. Notably, these domains, often having complex models with CFD, use very little Classification, showing these simulations not presently tractable by fully ML-based methodologies. Biology uses no Submodels (other than MD Potentials), since they generally do not use grids or have spatial resolution issues per se; this points out the highly domain-specific nature of how AI/ML is used. Machine-learned MD Potentials are heavily used in Materials



	Biology	Physics	Chemistry	Engineering	Fusion & Plasma	Comp. Science	Materials	Earth Science	Nuclear Energy
none	44	103	36	70	15	54	40	26	5
inactive	8	10	2	11	6	4	0	11	0
active	51	25	6	19	7	63	28	12	3

Fig. 4. AI/ML usage by science domain, project counts.

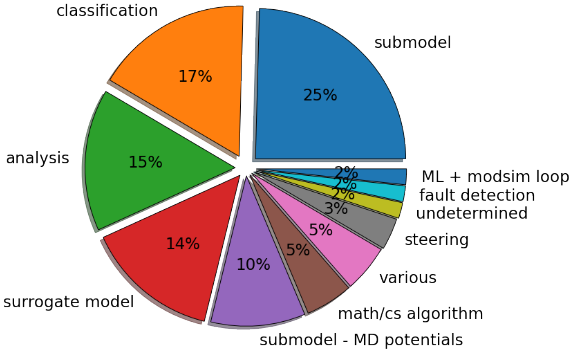


Fig. 5. AI/ML usage by AI motif, percentage of projects.

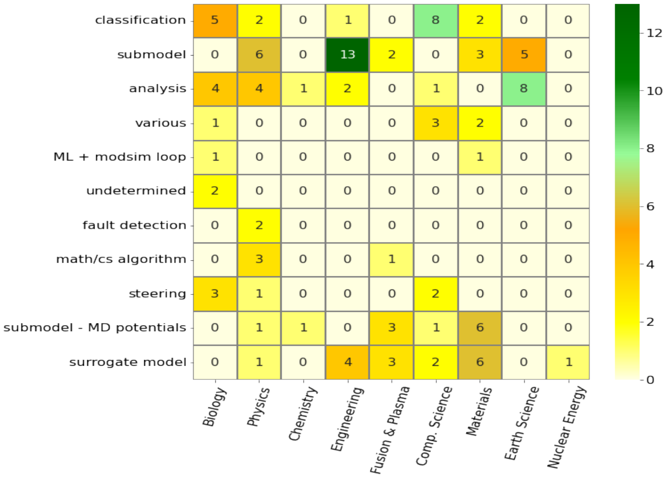
projects; they are used in Fusion/Plasma for plasma/surface interactions. Some Biology projects use MD Potentials but are otherwise classed, e.g., Steering. Computer Science contains many Classification projects; they have no Math/CS Algorithm components since categories like Classification and Various capture the use of such algorithms. Overall, we believe this line of analysis provides rich material for understanding how AI/ML methods can be applied to different science domains.

IV. High Scalability AI/ML on Summit

A. AI/ML-powered Gordon Bell Finalist Projects

We review Gordon Bell award finalists using AI/ML methods on Summit. These well-documented projects all scale to large Summit node counts. Table III summarizes numbers of Summit finalists for both standard and COVID-19 special Gordon Bell competitions. Following is a short overview of the projects and how they use AI/ML methods for science.

1. Ichimura et al., GB/2018 (math/cs algorithm motif) earthquake modeling using a neural network to form the preconditioner for a conjugate gradient solver (scalability to 4096 nodes) [15].



	Biology	Physics	Chemistry	Engineering	Fusion & Plasma	Comp. Science	Materials	Earth Science	Nuclear Energy
classification	5	2	0	1	0	8	2	0	0
submodel	0	6	0	13	2	0	3	5	0
analysis	4	4	1	2	0	1	0	8	0
various	1	0	0	0	0	3	2	0	0
ML + modsim loop	1	0	0	0	0	0	1	0	0
undetermined	2	0	0	0	0	0	0	0	0
fault detection	0	2	0	0	0	0	0	0	0
math/cs algorithm	0	3	0	0	1	0	0	0	0
steering	3	1	0	0	0	2	0	0	0
submodel - MD potentials	0	1	1	0	3	1	6	0	0
surrogate model	0	1	0	4	3	2	6	0	1

Fig. 6. AI motif vs. science domain, project counts.

TABLE III
Gordon Bell award finalist project counts

year category	2018 std	2019 std	2020 std	2020 COVID-19	2021 std	2021 COVID-19
Summit	5	2	4	2	1	3
Summit AI/ML	3	0	1	2	1	3

2. Patton et al., GB/2018 (classification motif) hyperparameter tuning for DNNs to find defect structures in microscopy images (scalability to 4200 nodes, measured 152.5 PF mixed precision) [16].

3. Kurth et al., GB/2018 (classification motif) detection of extreme weather patterns from imagery using adapted Tiramisu, DeepLabv3 DNNs (scalability to 4560 nodes, peak 1.13 EF mixed precision) [17].

4. Jia et al., GB/2020 (MD potentials motif) MD simulations of water and of copper using DeePMD-kit machine-learned potentials (scalability to 4560 nodes) [18].

5. Casalino et al., GB/2020/COVID-19 (steering motif) MD modeling of virus spike dynamics with sampling guided by 3D PointNet-based adversarial autoencoder (scalability to 4096 nodes) [19].

6. Glaser et al., GB/2020/COVID-19 (surrogate model motif) structure-based chemical screening for drug discovery, binding affinity scoring function represented by random forests (scalability to 4602 nodes) [20].

7. Nguyen-Cong et al., GB/2021 (MD potentials motif) MD modeling of carbon atoms in high pressure/temperature environments using machine-learned SNAP MD potentials (scalability to 4650 nodes) [21].

8. Blanchard et al., GB/2021/COVID-19 (classification motif) promising drug candidates found using genetic algorithm search of cross-attention network trained on BERT compound model embedding and transformer protein sequence model embedding (scalability to 4032 nodes, 603 PF mixed precision) [22].

9. Amaro et al., GB/2021/COVID-19 (steering motif) MD simulation guided by DeepDriveMD; also analysis motifs using OrbNet and ANCA-AE (scalability to 4096 nodes) [23].

10. Trifan et al., GB/2021/COVID-19 (steering motif) graph neural operator network, ANCA-AE and CVAE methods used to orchestrate joint MD and finite element simulations to model virus replication-transcription process (Summit scalability to 256 nodes) [24].

It is evident that AI/ML usage at scale is much more variegated than the case of scale-up of a single deep learning model. Benchmark suites like MLPerf [25] composed of pure ML models scaled to many nodes do not fully represent the multitude of ways scientists actually use AI/ML methods in practice. Most uses of ML are as components of a complex application code or workflow, for which the ML method provides meaningful benefits in performance or accuracy. This parallels the broader adoption of AI in the product space, whether as whole device or a component that enhances an existing device.

While the trend is to incorporate AI/ML increasingly into science workflows, this transition will take time. One research group is pursuing a 10-year plan to make ML-only weather forecasting as accurate as conventional modeling and simulation [26], while research in the area is proceeding apace [27], [28]. As will be discussed below, verification, validation, reproducibility and uncertainty quantification must be addressed for ML-powered scientific simulations [29].

B. AI/ML Methods at Extreme Scale

As just shown, AI/ML-powered scale-out projects use AI in different ways, for example a highly scalable science application with in-the-loop AI/ML, or conventional modeling and simulation on many nodes directed by AI/ML on few nodes. Here we consider cases where the AI/ML method itself is scaled up. We consider Gordon Bell finalist projects and several others. Some other OLCF projects are also engaged in scale-out of AI/ML however have not yet made results public.

1. Kurth et al. [17] analyzes image data from climate simulations to detect extreme weather events. Modified DeepLab3+ and Tiramisu networks are trained with LARC learning rate control and gradient lag with data parallelism. Training data input performance is optimized by use of node-local SSDs and MPI transfers of input data between nodes. Scaling to 4560 nodes results in peak 1.13 mixed precision Exaflops and parallel efficiency of 90.7%.

2. Yang et al. [30] models subsurface flow to study nuclear waste remediation. A physics-informed generative adversarial network (PI-GAN) solves the uncertainty quantification problem associated with the relevant stochastic partial differential equation with a network architecture constrained by the problem physics. Batch size limitations for GAN training requires use of a novel model parallelism scheme in addition to data parallelism. The code achieved over 1.2 mixed precision Exaflops performance on 4584 Summit nodes at 93% efficiency.

3. Laanait et al. [31] solves an inverse problem to reconstruct electron density from electron microscopy imagery. An adapted FC-DenseNET network is trained with a LARS/Adam optimizer, global batch size 27,600 and data parallelism. Novel optimizations for gradient reduction enable scalability to 4600 nodes and peak 2.15 mixed precision ExaFlops performance.

4. Khan et al. [32] uses ML to infer the astrophysical parameters of black hole mergers. A modified Wavenet architecture is trained with data parallelism using the LAMB optimizer, achieving 80% scaling efficiency from 8 to 1024 nodes of Summit.

5. Blanchard et al. [22] finds drug candidates using a workflow composed of multiple AI components. The most expensive is a BERT model architecture pretrained on SMILES-represented compound data with a custom model vocabulary. Pretraining uses the LAMB optimizer, data parallelism, gradient accumulation and global batch size up to 5.8 million while maintaining convergence rate. Parallel scaling from 1 to 4032 nodes is 68%; without I/O costs the figure is 83.3%. Peak performance is 603 mixed precision PF at 4032 nodes.

Clearly many kinds of deep learning model scale to a large fraction of Summit, representing diverse science areas (climate, materials, astrophysics, drug discovery). Performance and scaling characteristics depend heavily on model architecture, training settings and input data characteristics. Runtime components such as I/O (cf. [33]) and interconnect can be performance-critical, imposing requirements on system hardware components; for detailed discussion see Subsection VI-B

High floating point rates for model training requires large matrix sizes; this may not be practical for the fastest and most accurate models. Increasing use of sparsity may make this situation more complicated.

We expect the ability to train very large AI/ML models at leadership scale will continue to be a requirement into the future. A clear example of this is the transformer-based models of [22]. In the commercial world, transformer-based language models have scaled past the trillion parameter mark and require tightly integrated HPC systems of similar scale to those at national laboratories [34]. The trend of growing the model size to improve accuracy is expected to continue [35]. Our experience at the OLCF is that every field of science and engineering making use of high-performance computing features some number of projects requiring simulation on a leadership system at full scale to answer important (often fundamental) questions. Importantly, the investigators undertaking these simulations fashion their approximations, implementations, and resolutions based on the available hardware. Current evidence indicates that the same will be true for emerging AI for science workloads.

V. AI-coordinated Workflows: Case Studies

AI methods provide opportunities to orchestrate HPC science workflows in ways not previously possible. One of these is autonomous science discovery workflows. This might involve AI as a replacement of human judgment and control at decision points between the steps in a scientific campaign (e.g., observation, experiment, simulation, or AI training or inference step). AI methods can also enforce consistency between different simulation components, or combine results from other simulations or AI/ML models to generate new science. The "steering" motif described earlier is in some sense an example of this, insofar as a judgment protocol is learned by ML training and applied to guide some part of the simulation.

Below are case studies showing how AI methods are being used for coordinating workflow components on Summit.

A. Materials

An important issue in materials sciences and condensed matter physics that is increasingly being addressed using machine learning workflows is the problem of spanning the gap between models that are capable to address increasing length- and time-scales while maintaining the maximum of the physical accuracy of the more fundamental, higher accurate, yet more expensive, models. The most prevalent use of AI techniques in materials sciences has been in accelerating the exploration of the large search space of possible structures and compositions to achieve desired properties. [36] Yet, in most cases, these approaches have been based on the extraction of succinct feature set from materials databases of sizes that do not require the capabilities of high performance computing. Thus, most materials calculations at scale on Summit to date have followed a traditional modelling and simulation approach with minimal use of AI and machine learning incorporated into their workflow.

One example of applying machine learning techniques to high performance material simulations can be found in [37]. In this work ML is used to achieve high fidelity simulations of the statistical mechanics of multi-component alloys to obtain the finite temperature behavior of these concentrated solid-solution alloys. Previous work on investigating the statistical mechanics of materials resorted to constructing simple models using physical intuition from either experimental observations of a few first principles calculations. To achieve fully first principles accuracy for statistical mechanics calculations of either magnetic or chemical ordering in alloys, attempts were made to directly link scalable, real space density functional theory calculations with classical Monte-Carlo drivers. [38], [39] Machine learning allows a combination of these approaches by training classical models on expensive first principles data and reducing the time to solution compared to this fully density functional theory based approach while retaining most of its fidelity in capturing the quantitative physics. Thus the combination of highly scalable large scale density functional calculations of many materials configurations [40], which provides a sufficiently rich data set for machine learning, together with the use of physics inspired multitasking learning [41] and the use of a Bayesian information criterion [42] to avoid overfitting while still extracting the maximal information from the available data set, allowed Liu et al. [37] to formulate an integrated workflow that makes use of the HPC resources available with Summit to refine the ML derived model with new information obtained during the Monte-Carlo simulation to refine the model for high entropy alloys to obtain qualitative predictions of phase transitions in high entropy alloys.

B. Biology

In [24] an AI-enabled workflow models the replication transcription complex of COVID-19. The centerpiece is a combination of AI/ML components that iteratively couple a mesoscale simulation employing fluctuating finite element analysis (FFEA) and an atomistic-scale simulation based on all atom molecular dynamics (AAMD). The interposed AI/ML methods serve to impose consistency between the two vastly different kinds of simulation. The FFEA model takes 3D Cryo-EM data as input, while conformal changes from the FFEA model are captured by anharmonic conformational analysis enabled autoencoders (ANCA-AE). Conformational changes from the AAMD simulation are in turn captured by a convolution variational autoencoder (CVAE). The simulations are coupled by a graph neural operator (GNO) network.

A novelty of the project is use of AI to coordinate the science campaign across multiple facilities. The AAMD simulation is run using the NAMD code on the full Perlmutter system at NERSC and also the ThetaGPU system at the Argonne Leadership Computing Facility (ALCF). The CVAE model is trained offline on Summit on up to 256 nodes; alternatively it is trained on a Cerebras CS-2 system at the ALCF. The FFEA, ANCA-AE and GNO components are run on ThetaGPU. Balsam used for workflow orchestration.

C. Drug Design

[43] presents a drug lead discovery workflow as an iterative loop infused with AI/ML methods. Two key parts are an MD simulation and a surrogate ML model used for the compound ranking function. MD simulations are performed by OpenMM and NAMD on Summit and are themselves directed by DeepDriveMD using a CVAE to guide sampling. The surrogate model is based on a ResNet-50 network trained on 2D images generated from ligand SMILES strings. The surrogate model computes docking scores to downselect the set of compounds to evaluate by the more precise but more

expensive MD simulations. The workflow is managed by the RAdical-Pilot Task OveRlay (RAPTOR) system.

VI. Discussion

A. AI/ML Method Needs

AI/ML for science discovery has certain distinctive technical requirements pertaining to the AI/ML methods themselves and how they are applied to science problems.

1) Accuracy: Scientists seek confidence that the value inferred from a machine learning model is sufficiently accurate compared to the result from a principles-based modeling and simulation code. As one Summit INCITE PI put it succinctly, "The field of big data and machine learning has become extremely influential but without big theory it remains dogged by a lack of firm theoretical underpinning ensuring its results are reliable." [29]. Indeed, some methods such as neural networks under assumptions possess uniform approximation properties [44], though the bounds may not be adequate for the problem in hand. Properties like consistency and PAC learnability may be inadequate due to applying only in the limit or only probabilistically. Machine learning methods commonly lack the level of theory to give the kind of approximation guarantees available from other tools such as finite element theory. This being said, it must be admitted that some complex multiphysics codes or coupled application simulations also lack a priori approximation guarantees yet are reliably run in practice and produce verifiable results.

Some argue it is acceptable for practice to precede theory in use of deep learning [45]. Indeed, effective but unproven methods in the past have fallen out of favor for a period of time, some might argue needlessly, until convergence theory was developed (an example is the conjugate gradient method). Also, some machine learning methods whose approximation behaviors are not fully understood are nonetheless being applied today, even in mission-critical situations [46].

Our experience at the OLCF is that different projects have different validation requirements. Some simulations run at the OLCF have very demanding validation requirements, such as climate simulations supporting IPCC reports. One differentiating factor seems to be whether an ML-computed result can be confirmed by other means, such as conventional simulation or experiment; an example is the generation of drug candidates for evaluation and testing. We are aware of multiple projects using OLCF resources, currently actively integrating machine learning methods into workflows on multiple fronts (e.g., [26]), but not yet ready for production—limited in some cases by the need to assure confidence in the methods. Some have in fact already developed theory [47] to support accuracy guarantees, for example [48], Summit 2020 Gordon Bell awardee.

Researchers are actively studying approximation properties of machine learning methods for science applications (e.g., [49]). Progress could enable AI/ML adoption by more OLCF projects. However, this research must be problem-centric rather than method-centric. Results in the literature may be inapplicable for various reasons, for example from not answering the right questions or only solving "toy" problems. Though these are not entirely without value, the need here is for results that are actually applicable for production-grade simulations at leadership scale.

2) Generalizability: One aspect of accuracy is out-of-distribution generalization, a model's ability to generalize to input regimes unseen at training time, the failure of which can cause inaccurate simulations. As noted by one Summit Gordon Bell awardee, "In spite of the remarkable success of these ML methods, there is no guarantee for the quality of ML models when they are used to predict the properties of a configuration that is far from the training data set." [47]. This can be caused by required training data being unavailable or too expensive to generate, or the inherent need for more data than is practical [50]. The problem can be acute since exploratory simulations can by nature generate data with characteristics previously unseen. Though misclassified inputs to a network can often be contrived synthetically [51], the problem is not merely academic but in fact can cause failures in physical simulations [52]. Techniques to ensure generalizability or detect out-of-distribution data would be worthwhile.

3) Satisfaction of Constraints: Another facet of accuracy is the need for ML methods to satisfy certain constraints. This takes the form of conserving physical quantities, imposing other physical constraints [53] or preserving symmetries or other invariants. This is sometimes an essential requirement for correctness [52]. As one Summit INCITE climate scientist stated, "If networks are applied iteratively, it will be important to satisfy fundamental conservation properties and to stabilise simulations." [54] Constraints can be imposed exactly (up to roundoff) by choice of network architecture, enforced approximately by loss term, or imposed by a final correction. An example at the OLCF is [48] in which symmetries in molecular dynamics potentials are enforced exactly.

4) Explainability: Some OLCF-hosted projects require explainable AI (cf. [55], [56]). As stated by one Summit INCITE PI, "The inner workings and decision processes of these AIs are opaque. Results can be seen but an understanding why a decision was made is lacking. Therefore, while AI has been a powerful tool for prediction and classification, it has not yet been a tool for knowledge distillation." [57]. Unlike some other AI application domains, scientific inquiry as a collaborative effort requires not only "answers" but also "insight." An ML method with human-level proficiency but no human-level ability to explain its reasoning to a domain scientist can be a "dead-end," providing little guidance toward the next steps of discovery. While some phenomena may be only modelable as a black box, the ability of models to "show their work" yields significant benefit for human-in-the-loop science exploration. It should also be noted that some forms of physical theory are formulated and used to provide fundamental understanding while others are designed to provide predictive power for classes of phenomena. Even an explainable AI rooted in complete theory would not be helpful for the aims of fundamental theory. On the other hand, AI models of this ilk could be quite effective in replacing or, at least, augmenting phenomenology.

B. Hardware and Software Requirements

A distinctive characteristic of learning applications is that most developers interact with high-level frameworks (typically in the Python language), such as TensorFlow and PyTorch, providing nearly all essential building blocks for modeling. Beneath the frameworks, vendors provide full stack support and hence shield underlying hardware and software complexities from the developers. Since most AI/ML workloads boil down to 3 basic types of operations, i.e., convolution, recurrent operations and matrix multiplication, and can take advantage of mixed precision arithmetic, these applications are typically computational bound at the device level. At scale, however, they can be limited by I/O or communication as the trend of ever-growing data and model size continues.

I/O considerations: The I/O pattern of AI/ML workloads follows iterative random access. The aggregated read bandwidth needed to sustain full Summit data-parallel training is roughly estimated from single device training throughput on in-memory synthetic data, multiplying by input data size and number of devices. For the standard ResNet50 on ImageNet benchmark, a total of 20 TB/s is required for ideal scaling. This cannot be achieved on current shared file systems such as GPFS, the read bandwidth of which is only 2.5 TB/s. On the other hand, node-local NVMe has aggregate read bandwidth over 27 TB/s, satisfying needs of typical AI/ML applications. However, the training data of a large-scale scientific application can easily outsize single NVMe volume, hence data partitioning is needed. This can be expensive if per-epoch data shuffling is enforced. Since data on NVMe is not persistent between jobs, data staging is also required, with costs adding up as well (e.g., hundreds of TBs at the start of each training job for hyperparameter search). A high-performance shared file system or NVMe-based caching layer is highly desirable.

Communication considerations: The most common communication pattern for AI/ML workloads is allreduce. Even when highly optimized with GPUDirect for inter-GPU communication, with exploding model size it becomes a common bottleneck. For example, the per device allreduce message size for the ResNet50 and BERT-large models is about 100MB and 1.4 GB, respectively. Given Summit network bandwidth, 25 GB/s, and the algorithm (ring-based allreduce) bandwidth being half of network bandwidth, i.e., 12.5 GB/s, communication time is roughly 8 ms and 110 ms. The latter is close to the time of per-batch forward and backward propagation and hence hard to hide with computation-communication overlap. Thus models larger than BERT-large become communication-bound for the widely used data-parallel training on Summit. High-performance interconnect and/or generic model parallelization is essential for good scaling efficiency on future platforms.

VII. Conclusions

Since its launch in 2018, AI/ML adoption by projects on Summit has risen significantly and stands now at about 31% of INCITE projects actively using AI/ML and another 28% planning, exploring, previously using or indirectly using AI/ML. It is impossible from our vantage point to tell whether the other 1/3 of projects do not use AI/ML due to not having need or from lack expertise or resources.

Deep learning and other neural network methods are most commonly used, though other projects have reasons to use various other ML methods.

AI/ML adoption is highly differentiated by science domain, with Biology, Computer Science and Materials being top categories. Others like Engineering, Physics and Earth Science show significant current usage as well as planning for future use.

AI usage motifs are highly variegated across projects, with use of submodels being most popular—this unsurprising since modern simulations typically rely on many submodels, e.g., turbulence. The pattern of AI motif usage by science domain shows very distinctive patterns underscoring the very domain-specific nature of how projects use AI/ML methods.

Many AI/ML-powered codes scale up effectively. Multiple deep learning training codes scale to full Summit, using increasingly sophisticated models such as transformers. Biologists are substantial users of at-scale ML, the latter well-suited to modeling combinatorial interactions e.g., for protein folding. Reductions and I/O are stress points for scalability.

AI methods are enabling new science workflows, coordinating simulation and ML components, sometimes across systems. We expect use of autonomous workflows to increase.

ML theorists must provide better solutions to problems of accuracy, interpretability and generalization. Computational scientists will need to deeply understand both domain science and ML to evaluate appropriateness and applicability of AI methods for science.

Future use of AI for science will need large systems with good interconnects and file systems with excellent read characteristics and reasonable semantics.

We have presented an early-stage analysis of AI for science at scale. New methods and applications are being developed that were not even imagined several years ago. We expect many new developments in AI for science in coming years.

Acknowledgment

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

[1] "National Artificial Intelligence Initiative Act of 2020," https://www.congress.gov/bill/116th-congress/house-bill/6216/.

[2] N. Baker *et al.*, "Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence," Department of Energy, Tech. Rep., 2 2019.

[3] R. Stevens *et al.*, "AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science," Department of Energy, Tech. Rep., 2 2020.

[4] K. Fagnan *et al.*, "Data and Models: A Framework for Advancing AI in Science," Department of Energy, Tech. Rep., 12 2019.

[5] T. Hey, "Report From Cross-Cutting AI Subcommittee," Department of Energy, Tech. Rep., 9 2020.

[6] John M Jumper et al., “Highly accurate protein structure prediction with alphafold,” *Nature*, vol. 596, pp. 583 – 589, 2021.

[7] S. R. Young *et al.*, “Optimizing deep learning hyper-parameters through an evolutionary algorithm,” in *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments*, ser. MLHPC ’15. ACM, 2015.

[8] W. M. Tang, “Deep learning acceleration of progress toward delivery of fusion energy,” in *NVIDIA GPU Technology Conference*, 2017.

[9] Vazhkudai, Sudharshan S. et al., “The design, deployment, and evaluation of the coral pre-exascale systems,” in *SC ’18 Proceedings*, 2018, pp. 661–672.

[10] “TOP500 Supercomputer Sites,” https://www.top500.org.

[11] “Exascale Computing Project,” https://www.exascaleproject.org/.

[12] “The COVID-19 High Performance Computing Consortium,” https://covid19-hpc-consortium.org/.

[13] “ACM Gordon Bell Prize,” https://awards.acm.org/bell.

[14] Wyatt, Michael, et al., “Is Disaggregation Possible for HPC Cognitive Simulation?” in *MLHPC Workshop, at SC ’21*, 2021.

[15] T. Ichimura *et al.*, “A Fast Scalable Implicit Solver for Nonlinear Time-Evolution Earthquake City Problem on Low-Ordered Unstructured Finite Elements with Artificial Intelligence and Transprecision Computing,” in *SC ’18 Proceedings*, 2018, pp. 627–637.

[16] R. M. Patton *et al.*, “167-PFlops Deep Learning for Electron Microscopy: From Learning Physics to Atomic Manipulation,” in *SC ’18 Proceedings*, 2018, pp. 638–648.

[17] T. Kurth *et al.*, “Exascale Deep Learning for Climate Analytics,” in *SC ’18 Proceedings*, ser. SC ’18. IEEE Press, 2018.

[18] W. Jia *et al.*, “Pushing the Limit of Molecular Dynamics with Ab Initio Accuracy to 100 Million Atoms with Machine Learning,” in *SC ’20 Proceedings*. IEEE Press, 2020.

[19] L. Casalino *et al.*, “AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics,” *The International Journal of High Performance Computing Applications*, vol. 35, no. 5, pp. 432–451, 2021.

[20] J. Glaser *et al.*, “High-throughput virtual laboratory for drug discovery using massive datasets,” *The International Journal of High Performance Computing Applications*, vol. 35, no. 5, pp. 452–468, 2021.

[21] K. Nguyen-Cong *et al.*, “Billion Atom Molecular Dynamics Simulations of Carbon at Extreme Conditions and Experimental Time and Length Scales,” in *SC ’21 Proceedings*, ser. SC ’21. New York, NY, USA: Association for Computing Machinery, 2021.

[22] A. E. Blanchard *et al.*, “Language Models for the Prediction of SARS-CoV-2 Inhibitors,” in *SC ’21 Proceedings*, 2021.

[23] R. Amaro *et al.*, “#COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol,” in *SC ’21 Proceedings*, 2021.

[24] A. Trifan *et al.*, “Intelligent Resolution: Integrating Cryo-EM with AI-Driven Multi-Resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action ,” in *SC ’21 Proceedings*, 2021.

[25] P. Mattson *et al.*, “MLPerf training benchmark,” *arXiv preprint arXiv:1910.01500*, 2019.

[26] P. Dueben, “Progress and challenges for using machine learning in weather and climate prediction,” in *NVIDIA GTC*, March 2021.

[27] J. A. Weyn *et al.*, “Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models,” *Journal of Advances in Modeling Earth Systems*, vol. 13, no. 7, p. e2021MS002502, 2021.

[28] J. A. A. Herrera *et al.*, “Weatherscapes: Nowcasting heat transfer and water continuity,” *ACM Transaction on Graphics*, vol. 40, no. 6, 2021.

[29] P. V. Coveney and R. R. Highfield, “When we can trust computers (and when we can’t),” *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2197, p. 20200067, 2021.

[30] L. Yang *et al.*, “Highly-scalable, physics-informed GANs for learning solutions of stochastic PDEs,” in *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*. IEEE, 2019, pp. 1–11.

[31] N. Laanait, J. Romero, J. Yin, M. T. Young, S. Treichler, V. Starchenko, A. Borisevich, A. Sergeev, and M. Matheson, “Exascale deep learning for scientific inverse problems,” *arXiv preprint arXiv:1909.11150*, 2019.

[32] A. Khan *et al.*, “Physics-inspired deep learning to characterize the signal manifold of quasi-circular, spinning, non-precessing binary black hole mergers,” *Physics Letters B*, vol. 808, p. 135628, 2020.

[33] A. K. Paul *et al.*, “Characterizing Machine Learning I/O Workloads on Leadership Scale HPC Systems,” in *2021 29th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, 2021, pp. 1–8.

[34] “Introducing the AI Research SuperCluster — Meta’s cutting-edge AI supercomputer for AI research,” https://ai.facebook.com/blog/ai-rsc.

[35] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, “Deep Learning’s Diminishing Returns: The Cost of Improvement is Becoming Unsustainable,” *IEEE Spectrum*, vol. 58, no. 10, pp. 50–55, 2021.

[36] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, “Recent advances and applications of machine learning in solid-state materials science,” *npj Computational Materials*, vol. 5, p. 83, 2019.

[37] X. Liu, J. Zhang, J. Yin, S. Bi, M. Eisenbach, and Y. Wang, “Monte carlo simulation of order-disorder transition in refractory high entropy alloys: A data-driven approach,” *Computational Materials Science*, vol. 187, p. 110135, 2021.

[38] M. Eisenbach *et al.*, “A scalable method for ab initio computation of free energies in nanoscale systems,” in *SC ’09 Proceedings*, ser. SC ’09. New York, NY, USA: ACM, 2009, pp. 64:1–64:8.

[39] S. N. Khan and M. Eisenbach, “Density-functional Monte-Carlo simulation of CuZn order-disorder transition,” *Phys. Rev. B*, vol. 93, no. 2, p. 024203, Jan 2016.

[40] M. Eisenbach, J. Larkin, J. Lutjens, S. Rennich, and J. H. Rogers, “Gpu acceleration of the locally selfconsistent multiple scattering code for first principles calculation of the ground state and statistical physics of materials,” *Computer Physics Communications*, vol. 211, pp. 2–7, 2017.

[41] M. L. Pasini, Y. W. Li, J. Yin, J. Zhang, K. Barros, and M. Eisenbach, “Fast and stable deep-learning predictions of material properties for solid solution alloys,” *Journal of Physics: Condensed Matter*, vol. 33, no. 8, p. 084005, 2020.

[42] J. Zhang, X. Liu, S. Bi, J. Yin, G. Zhang, and M. Eisenbach, “Robust data-driven approach for predicting the configurational energy of high entropy alloys,” *Materials & Design*, vol. 185, p. 108247, 2020.

[43] A. A. Saadi *et al.*, “Impeccable: Integrated modeling pipeline for covid cure by assessing better leads,” in *50th International Conference on Parallel Processing*, 2021, pp. 1–12.

[44] K. F. E. Chong, “A closer look at the approximation capabilities of neural networks,” *ArXiv*, vol. abs/2002.06505, 2020.

[45] Y. LeCun, “The epistemology of deep learning,” *Institute for Advanced Studies*, 2019.

[46] D. Martinez *et al.*, “Deep learning evolutionary optimization for regression of rotorcraft vibrational spectra,” in *2018 IEEE/ACM Machine Learning in HPC Environments (MLHPC)*, 2018, pp. 57–66.

[47] L. Zhang, D.-Y. Lin, H. Wang, R. Car, and W. E, “Active learning of uniformly accurate interatomic potentials for materials simulation,” *Phys. Rev. Materials*, vol. 3, p. 023804, Feb 2019.

[48] W. Jia *et al.*, “Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning,” in *SC ’20 Proceedings*. IEEE Press, 2020.

[49] B. Adcock and N. C. Dexter, “The gap between theory and practice in function approximation with deep neural networks,” *ArXiv*, vol. abs/2001.07523, 2021.

[50] G. Yehuda, M. Gabel, and A. Schuster, “It’s not what machines can learn, it’s what we cannot teach,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 831–10 841.

[51] C. Szegedy *et al.*, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.

[52] S. Rasp, M. S. Pritchard, and P. Gentine, “Deep learning to represent subgrid processes in climate models,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 39, pp. 9684–9689, 2018.

[53] T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, “Enforcing analytic constraints in neural networks emulating physical systems,” *Phys. Rev. Lett.*, vol. 126, p. 098302, Mar 2021.

[54] P. D. Dueben and P. Bauer, “Challenges and design choices for global weather and climate models based on machine learning,” *Geoscientific Model Development*, vol. 11, no. 10, pp. 3999–4009, 2018.

[55] J. Streich *et al.*, “Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals?” *Current Opinion in Biotechnology*, vol. 61, pp. 217–225, 2020, plant Biotechnology - Food Biotechnology.

[56] Y. Zhao *et al.*, “RANS turbulence model development using CFD-driven machine learning,” *J. Comput. Phys.*, vol. 411, p. 109413, 2020.

[57] A. L. Harfouche *et al.*, “Accelerating climate resilient plant breeding by applying next-generation artificial intelligence,” *Trends in biotechnology*, vol. 37, no. 11, pp. 1217–1235, 2019.