PUBLIC SUBMISSION

As of: March 21, 2025
Received: March 15, 2025
Status [redacted]
Tracking No. m8b-0m5b-von8
Comments Due: March 15, 2025
Submission Type: API

Docket: NSF_FRDOC_0001
Recently Posted NSF Rules and Notices.

Comment On: NSF_FRDOC_0001-3479
Request for Information: Development of an Artificial Intelligence Action Plan

Document: NSF_FRDOC_0001-DRAFT-7001
Comment on FR Doc # 2025-02305

Submitter Information

Email [redacted]
Organization: Fast Machine Learning Collaboration and AI-Accelerated Algorithms for Data-Driven Discovery Institute

General Comment

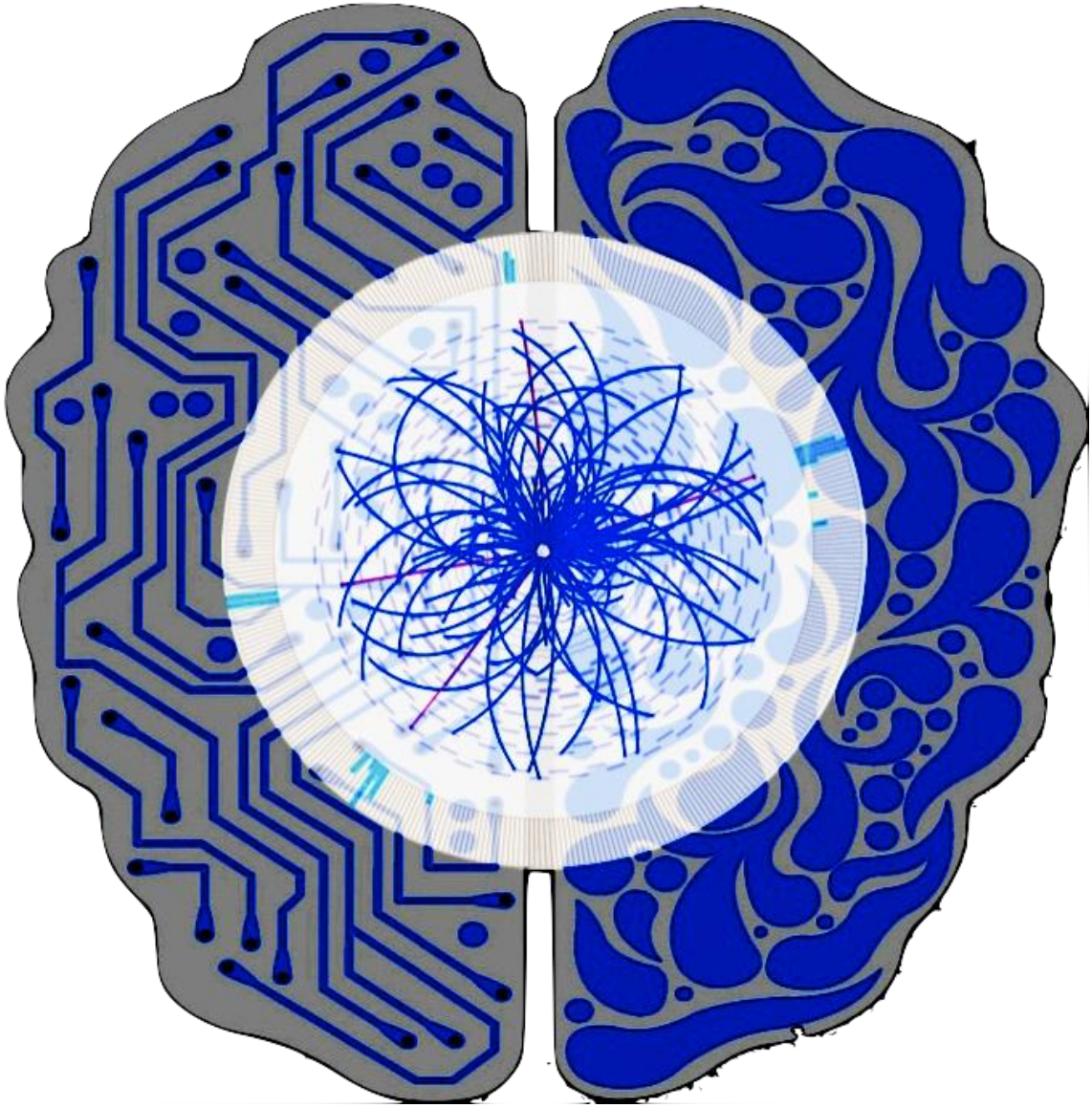
The attached document represents a proposed action plan from the Fast Machine Learning Organization (https://fastmachinelearning.org/) and
AI-Accelerated Algorithms for Data-Driven Discovery(A3D3) Institute (https://a3d3.ai).

Attachments

FastML_AI_Action_Plan

Artificial Intelligence Plan for Fast Machine Learning

Fast Machine Learning Organization (https://fastmachinelearning.org/**)**
AI-Accelerated Algorithms for Data Driven Discovery(A3D3) Institute (https://a3d3.ai/**)**



This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI Action Plan and associated documents without attribution.

Introduction

Artificial Intelligence (AI) has had a transformative impact on society. It is quickly impacting many fields of scientific research, business operations, and everyday life. Despite rapid progress, some areas of AI research advance slower due to difficulties in research and development. In particular, the development of low latency, fast AI requires different goals and AI design to achieve the ultimate performance. Furthermore, the most pressing fast AI problems have originated from attempts to solve specific scientific problems in many different domains, with domain scientists using challenging scientific problems to drive research. Despite that, we are quickly realizing that the lessons learned from the individual domains can be consolidated into a new scientific field that will have significant benefits for future AI technologies. In particular, Fast AI:

- extends the capability of what is possible with AI, leading to the adoption of AI in new systems that avoid current latency restrictions,
- opens new paradigms for AI design and optimization strategies for AI deployment; this leads to new, efficient chip design and heterogeneous computing systems
- motivates the need for extensive use of AI-based algorithm compression strategies, allowing for an understanding of how to make efficient AI
- ultimately leads to the possibility of both faster and lower power usage of all AI algorithms [1,2].

In light of the further advancement of these technologies, a community has emerged over the past six years entitled the Fast Machine Learning community. Within this community, scientists and industry members have come together to develop novel AI solutions that have opened new pathways to the deployment and use of AI. While the drivers of this effort have started from scientific domains, an ecosystem has emerged that brings this technology from one domain to the global community.

The impact of science on data rate and latency requirements can be seen in Figure 1. A broad range of systems exist in many different scientific domains that demand high streaming data rates approaching the petabyte per second data rate. Additionally, the demands of these experiments often require algorithms to respond in fractions of a second, below one millisecond. The existing computational toolkit that serves commercial domains, such as Google Cloud, Netflix, or other extensive computing facilities, cannot address the combined significant data rates and timescales needed to run these scientific experiments. As a result, commercial solutions to these problems do not exist, and novel AI solutions are required to address these challenges. The Fast Machine Learning Community was founded to address these difficult problems and has built a community around the need for solutions for these problems in many scientific domains covering High Energy Physics, Astrophysics, Materials Science, Plasma Physics, Neuroscience, Quantum Computing, and Real-time Systems. Out of this greater effort, several dedicated funded programs have emerged, including the NSF-funded Harnessing the Data Revolution Institute: AI Algorithms for Data-Driven Discovery (A3D3) aimed at developing fast AI algorithms for real-time high-energy physics, astrophysics, and neuroscience experiments.

In this document, we advocate for continued support of scientific research whose focus is on optimized AI strategies for new processing technology. While a large community is working to extend AI algorithm design, the community focusing on processor technology is smaller despite an equally significant need for development. Additionally, we advocate for support for research to deploy algorithms on customized hardware to scale in heterogeneous computing clusters. In particular, support for the optimized deployment of algorithms on heterogeneous clusters is needed to ensure high throughput systems can efficiently utilize modern computing to solve critical problems.

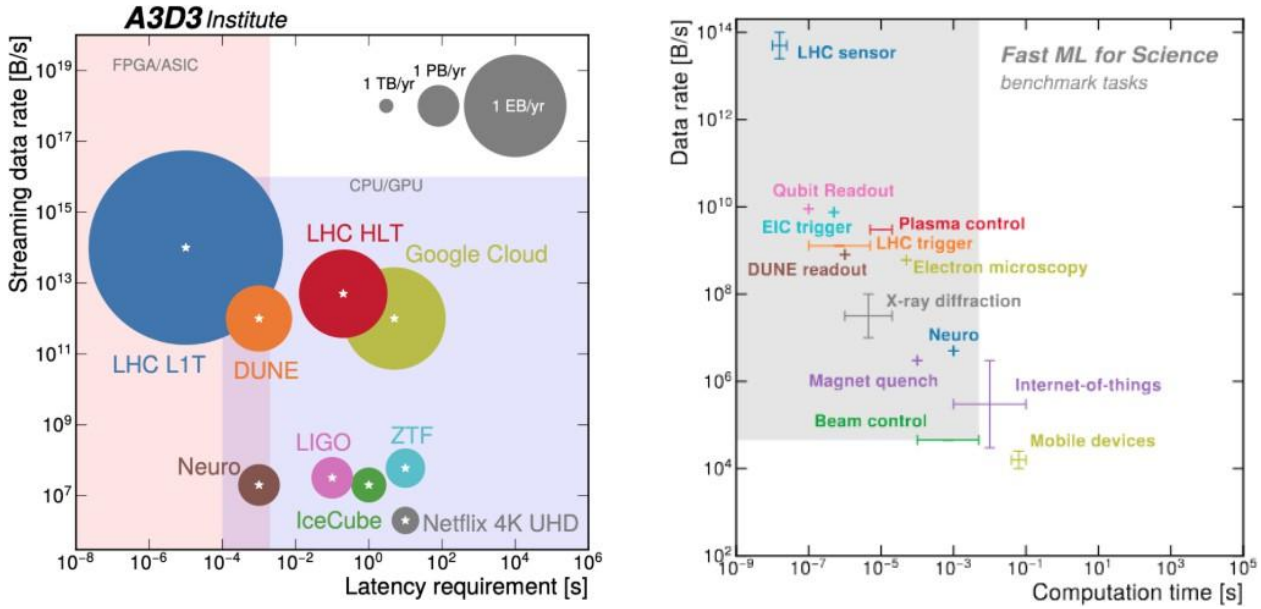


Figure 1: Left: Latency requirements and system-level data streaming rates for various experiments in the physics and neuroscience domain. Right: Latency requirements and task-level data rates for benchmark tasks across many domains of science.

To cultivate an ecosystem where scientific problems drive optimized AI, we advocate for the continued support of cross-disciplinary efforts that aim to bring critical scientific problems from many different domains to the AI community. Building on the existing Fast Machine Learning/A3D3 effort for low-latency, efficient AI (Fig. 2), we aim to grow the ecosystem within their respective scientific communities and build toolkits that allow for solutions to be transferred across domains. Already, we have had a significant amount of success through the development of the software toolkits hls4ml [3] and the SONIC software toolkit [4]. We aim to continue this effort, extending to more domains and finding science drivers where fast AI is needed. This paradigm somewhat differs from other funded AI institutes, which tend to focus on just a few scientific domains.

In the following white paper, we present the impact of Fast Machine Learning in many domains. Our point with this paper is to highlight a significant demand for research and development in these domains, with an emphasis that publicly funded scientific problems are essential to extending further research in this domain. Furthermore, we highlight the need for a cross-disciplinary effort that brings together these domains and produces a common toolkit that can enable fast machine learning for all of the sciences and the rest of the world.

Computational Limitations

Modern processing technology, such as Graphics Processing Units (GPUs), has enabled the rise of AI. However, GPUs have limitations that are implicit in their design. GPUs are best optimized when performing large batch AI processing and are limited in their ability to run at low latencies due to how data is transferred on and off the chip. The above design limitations make it impossible to use GPUs for algorithms that demand latencies below 100 microseconds and difficult for latencies below 1 millisecond. This leads to a demand for different AI-optimized hardware architectures. The Fast Machine Learning Organization has focused on alternative designs by building a toolkit that aims to embed the neural net-

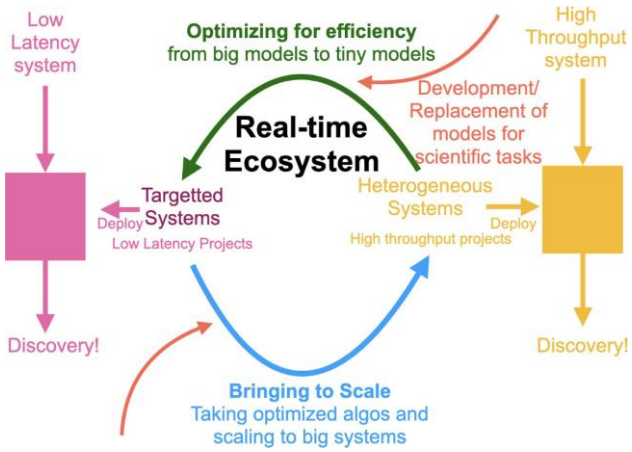


Figure 2: Diagram for the Fast Machine Learning/A3D3 ecosystem aimed at developing AI inference solutions for scientific discovery.

work architecture on the chip, as opposed to using standard chip architecture elements such as systolic arrays. This approach is encased in the HLS4ML [5] design for FPGAs and ASICs, and the cgra4ml [6] design when dedicated AI processor technology is used. Expanding the scope of what is capable in this low latency paradigm requires understanding semiconductor architecture and utilizing the appropriate chip-level tools for the desired problem [7]. Driven by scientific problems, the Fast Machine Learning community is exploring this space and developing a toolkit and knowledge base that exploits the right chip architecture for the right problem.

When dealing with large throughput systems, such as those present for big data experiments, including the Large Hadron Collider and the Vera C. Rubin Astronomical Observatory, there is a need to optimize computing resources to contend with the large data rates while being cost-effective. Current industry solutions are customized and primarily driven by demands for large-scale deployment of large language models. Flexible approaches, which integrate with an existing broad series of workflows, are limited in scope. However, there is a need for an inference toolkit that is flexible to many experiments and provides a rapid adoption of AI tools with optimized data ingestion. Building on the chip-level ideas within the Fast Machine Learning Community, we also pursue strategies to optimize the use of heterogeneous systems (CPU/GPU and other processors) at a large scale. Here, our work targets a solution that builds on existing scientific workflows, allowing for the continual adoption of new AI algorithms as AI solutions emerge.

Connecting the research, fast machine learning work targets optimized use in a generic setting, allowing chip-level optimizations to be propagated to systems-level optimizations. This work leads to a holistic view of how we can adopt Fast machine-learning strategies both at the macro and micro levels. As a result, we aim to build a common toolkit that captures all scales and can be applied to many scientific workflows. A survey of scientific applications, problems, and tools can be found in this A3D3 white paper [8].

Fast Machine Learning in High Energy Physics

With 40 million collisions per second, the Large Hadron Collider(LHC) ingests data at over 1 Petabit/s. This data flow, shown in figure 3, exceeds that of any other device in the world. Furthermore, the high radiation environment provides a constraint that the data needs to be analyzed within a few microsec-

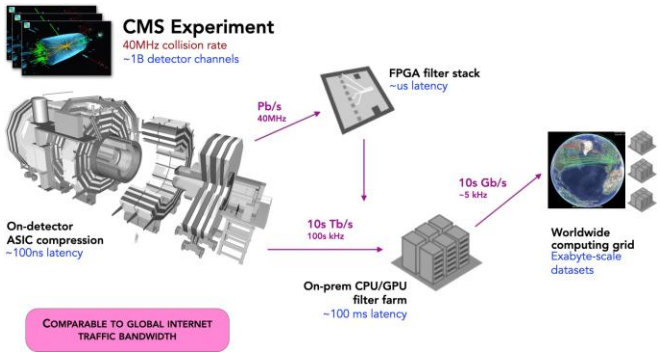


Figure 3: Data flow from the Compact Muon Solenoid detector on the Large Hadron Collider. Data is read at 1 Pb/s to an FPGA custom electronics system that then selects one event in 400 to a second CPU/GPU tier and then selects one event in 100 to be processed in detail and saved to disk.

onds. This analysis selects only the most interesting collision among 400 to be sent to a downstream computing cluster for further analysis. As a result of this enormous data rate and latency restriction, a unique demand for AI algorithms that can operate on a single object (aka collision) within nanosecond timescales has emerged. This work led to the creation of the hls4ml software library aimed at building latency-efficient AI algorithms using ASICs and FPGA processors. The severe restrictions further motivate additional optimizations in algorithm design that involve the reduction of hardware resources through zero-suppression, algorithm compression, and fixed-bit precision operations. As a result, problems with the LHC have been a driver for the development of new strategies to approach low-latency AI inference. The data rates are significantly reduced for downstream data processing at the LHC. Despite the reductions, the data rates are still huge, and the data itself is complex, requiring a tiered set of algorithms to process the data. The entire downstream data rate alone still significantly exceeds the amount of annual data produced at any commercial computing cluster. As a result, there is a need for efficient use of computing and power resources and the same optimized AI algorithmic compression strategies needed for the upstream data. The algorithms used to process high-energy physics data comprise a diverse set of AI and rule-based (non-AI) algorithms, some of which can be run on GPUs/other processors and some not. As a result, an additional challenge emerges: how to efficiently balance computing resources across a mixture of processing architectures and algorithms. Moreover, as time progresses, many of the rule-based algorithms are being replaced by AI algorithms, making it such that the optimal balance of processing elements needed for efficient computing changes over time. To contend with these challenges, members of the community have been developing software frameworks, including the SONIC software framework [4], which build on industry tools but then add additional optimizations to enable dynamic, optimized resource allocation. Other high energy physics experiments, including the neutrino physics experiments Icecube and DUNE and the LIGO gravitational wave experiment, require similar strategies [9–11], and as a result, our building on the SONIC infrastructure for their own low-latency inference pipelines.

By adding the ability to perform AI in real-time experiments in high-energy physics experiments, we open the door to possible major discoveries and new measurements of the universe that were not possible. At the LHC, AI enables the high-precision analysis of every collision instead of just one in 400. Already, we see these new real-time AI approaches are enabling new unprecedented measurements of the Higgs boson properties [12] and novel AI-driven for unexplained phenomena. A deviation in either these Higgs boson measurements or the observation of even a single anomalous event would completely change our whole understanding of how the universe works, causing a revolution in physics [13].

Fast Machine Learning in Astrophysics

Astrophysics may well be regarded as the slowest science with typical latencies of seconds to years or even decades in terms of response time requirements: a new astrophysical phenomenon – a transient – is detected by a satellite, a ground-based telescope, or a detector for non-electromagnetic signals (Gravitational Waves or Neutrinos) but it is not known how long this transient will last. It could be milliseconds or years, but the decision of what needs to be done next, to confirm what it is, or to gather more information to help that confirmation, does need to be made as quickly as possible. The process of making this decision can also be computationally expensive, involving the collation of as much as is known about the region of the sky where the new transient was seen from archives around the world and also sorting through a myriad of computational models that might fit the data so far gathered. For example, gravitational wave interferometers, which have a sampling rate of 16 kHz across a worldwide network of detectors, seek to provide real-time information to the community to enable follow-up by telescopes across the world. Neutrino observatories have similar goals of providing real-time directional information of muon neutrino events. Ground—and space-based astronomical observatories follow up on these detections in as near-real-time as possible to collect time-critical data related to them. In the longer term, these different messengers represent a number of different data modalities, e.g., time series, images, and particles, which together infer the same phenomena, opening the possibility for synergistic analyses to run in coordination in real time.

The other challenge is that this decision process is no longer just happening for a single transient detected every few days or even once per night but for millions of transients detected every night, particularly with the Vera Rubin Observatory coming online in 2025. Fast inferencing is essential to filter out the rare few transients that are genuinely novel or interesting enough to warrant further follow-up using finite resources – there are only so many 10-meter class telescopes on the planet. In the big data game, there are strategic policy decisions that may also influence what happens next: to achieve the best sample of a particular class of rare transient for statistical analysis, it may be worth not wasting resources on following up every single one but only those that are the most useful to the overall program.

Fast Machine Learning in Materials Science

Advancing American leadership in technology and national security requires a fundamental rethinking of how we discover, develop, and deploy critical materials—particularly those underpinning advanced semiconductor technologies. The strength and security of our nation depend on ensuring that sensitive communications, military systems, and critical infrastructure are built upon trusted, reliable, and domestically sourced materials. One crucial area of focus is in backend-of-line (BEOL) semiconductor processes, which integrate chip components vital to secure end-to-end encryption, advanced computing, and ultra-wideband radio frequency (RF) communications. Achieving breakthroughs in these technologies demands not only accelerated discovery but also innovative synthesis approaches. Much of the novel materials require leveraging non-equilibrium processes to create unique functional materials unattainable through conventional methods. Non-equilibrium synthesis enables access to metastable phases, controlled defect structures, and novel material properties essential for next-generation semiconductor components. To exploit these transient phenomena effectively, artificial intelligence (AI) must be embedded in real-time to control experimentation workflows, enabling dynamic, real-time decision-making and control impossible without fast AI control systems. Integrating AI into experimental instruments and manufacturing systems allows immediate identification and stabilization of desired non-equilibrium states, dramatically accelerating the innovation pipeline and increasing quality control, making novel materials and devices economically viable.

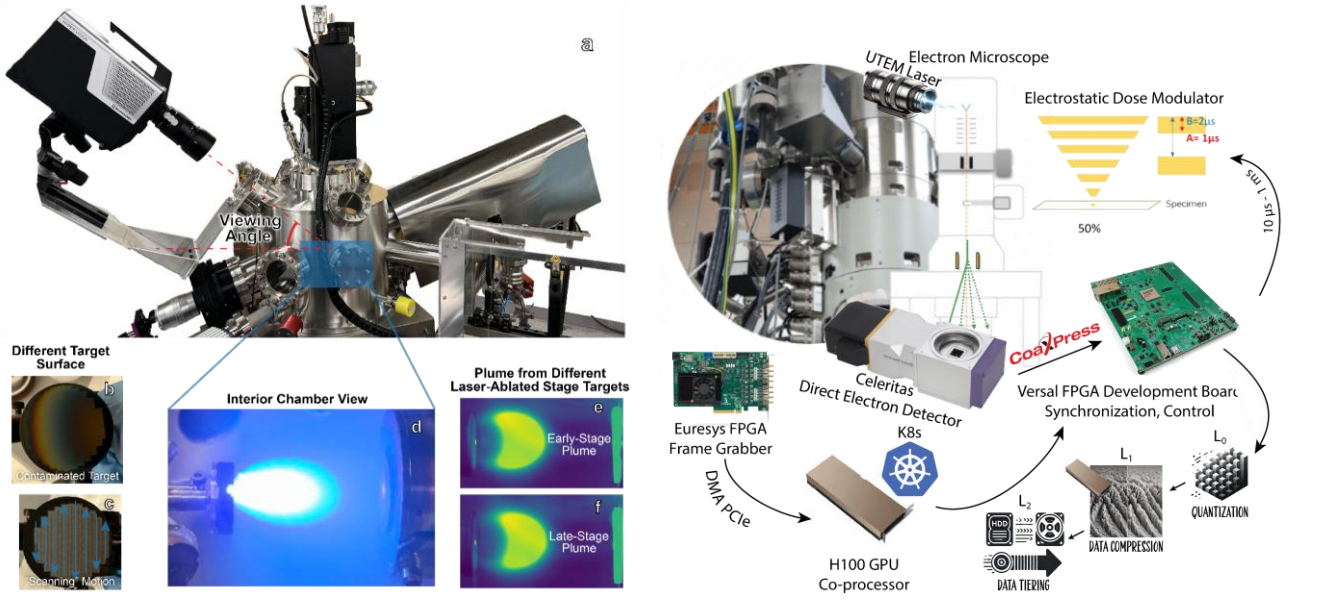


Figure 4: Left: Controlled synthesis of quantum materials using plume-dynamics imaging. Right: Design schematic for AI-in-the-loop ultra-low dose-controlled imaging.

Seeing is understanding, especially at the atomic scale. Yet, despite its historic leadership in science and technology, the United States has fallen behind China and Europe in advanced electron microscopy due to their massive investments. We need to work smarter, not harder. Methods such as four-dimensional scanning transmission electron microscopy (4D-STEM) allow researchers to visualize materials with atomic resolution study of correlated electron effects that underpin quantum computing and information technologies. These microscopes generate immense datasets at extraordinary speeds, capturing transient phenomena like sudden phase transitions, rapid atomic rearrangements, and the formation of defects—all critical events that determine material functionality - and properties of quantum decoherence. Traditional, human-driven data analysis can no longer keep pace; it is too slow and inflexible, often missing fleeting opportunities for breakthroughs. Furthermore, these techniques are inherently destructive. By integrating artificial intelligence (AI) directly into electron microscopy instruments—powered by advanced hardware accelerators such as Field Programmable Gate Arrays (FPGAs)—researchers can instantly analyze and interpret atomic-scale changes while simultaneously maximizing signal-to-noise and minimizing sample damage. Fast-AI control systems for electron microscopy hold the potential of supplanting cryogenic electron microscopy in performance, capabilities, and utility. Furthermore, real-time AI capability allows automated systems to act immediately, precisely steering reactions or atomic assembly processes toward desired quantum states. Reclaiming leadership in atomic-scale electron microscopy through AI-driven instrumentation is therefore essential to America's ability to understand, engineer, and deploy next-generation quantum materials that will define the technological landscape of tomorrow.

Investing in AI-driven electron microscopy and non-equilibrium synthesis is not simply a scientific pursuit—it is a strategic necessity. Reestablishing American leadership in atomic-scale characterization and real-time control positions our nation at the forefront of quantum computing, secure communications, and advanced semiconductor manufacturing. Integrating rapid AI into materials discovery processes enables smarter, faster innovation, reinforcing our technological sovereignty, strengthening national security, and revitalizing domestic supply chains. This convergence of AI, advanced microscopy, and non-equilibrium techniques ensures America's continued global competitiveness and technological leadership for decades to come.

Fast Machine Learning in Plasma Physics

Excerpt from [14] – While magnetic confinement devices can be designed to operate far from stability limits with nominal plasma conditions, transient events or other deviations from desired parameters may lead to instability. Efficient usage of the applied magnetic field to confine higher plasma pressure generally leads to less stable equilibria due to β-driven instabilities, where β is the ratio between the plasma and magnetic pressures. As tokamaks and other magnetic confinement devices achieve fusion-relevant conditions, real-time diagnosis and control of plasma instabilities are essential for maintaining desirable plasma parameters and robust performance. Plasma conditions and instabilities in tokamaks can evolve on an Alfvénic timescale for the most demanding cases, necessitating fast evaluations and control decisions on a timescale of microseconds for quickly growing or rotating kink and tearing instabilities [15]. For this reason, real-time control of plasma properties requires reasonably accurate implementation of control algorithms with latencies commensurate with the plasma dynamics.

Recently, deep learning [16], and other data-driven AI techniques have been applied to a variety of areas in plasma and fusion research. For tokamak control applications, current studies have mainly focused on the prediction and control of disruptive events and equilibrium quantities. These studies can be broadly divided into two categories: 1. by developing a data-driven model to predict the intended plasma quantities [17–20] and using these predictions to inform a non-AI-based control algorithm [21, 22], or 2. by training a deep reinforcement learning agent on a computational [23, 24] or data-driven [25, 26] simulator and using the trained agent to directly generate control policies. Some of these trained models have also been tested in real-time and have achieved latencies on the order of the device's plasma control system cycle time, ranging from milliseconds to as low as 50 μs [18, 21, 23, 26, 27].

Besides disruption and equilibrium, prediction and control of non-equilibrium quantities such as MHD instabilities [28, 29] and edge localized modes [27, 30] with AI approaches have also been investigated. However, applying these models for real-time control can be more challenging as the model needs to perform inference with latency similar to or smaller than the timescale of the plasma phenomena to be controlled, which can be as fast as a few microseconds. Depending on the complexity of the specific algorithm, a real-time implementation may require unique optimization and computing hardware to achieve the target latency.

Fast Machine Learning in Neuroscience

Technology to precisely modulate brain activity is crucial for clinical and basic neuroscience. The ability to deliver targeted neural perturbations relies on technology to process neural data in real time and deliver control signals for neuromodulation (Fig. 5). However, neural data processing typically requires complex computation and powerful hardware. This requires streaming data from the source to the processing hardware, which limits applications in both clinical and research settings. Real-time neural signal processing systems overcome these challenges by operating at faster latencies (sub-millisecond) and low power. In particular, deep learning algorithms for neural signal feature extraction, when sufficiently fast and efficient, have the potential to be developed as a fully mobile, closed-loop system that can be used for clinical and basic neuroscience research.

Many neurological, neurodegenerative, and neuropsychiatric disorders are resistant to current pharmacological treatments. Advances in systems neuroscience indicate a promising treatment option involves directly modulating neural pathways within the brain with invasive or non-invasive neuromodulation technologies such as Deep Brain Stimulation (DBS) and Transcranial Magnetic Stimulation (TMS) [31]. Despite its potential, neuromodulation currently faces many technical and scientific hurdles that restrict their wider adoption. An essential component that determines the efficacy of neuromodulation therapies

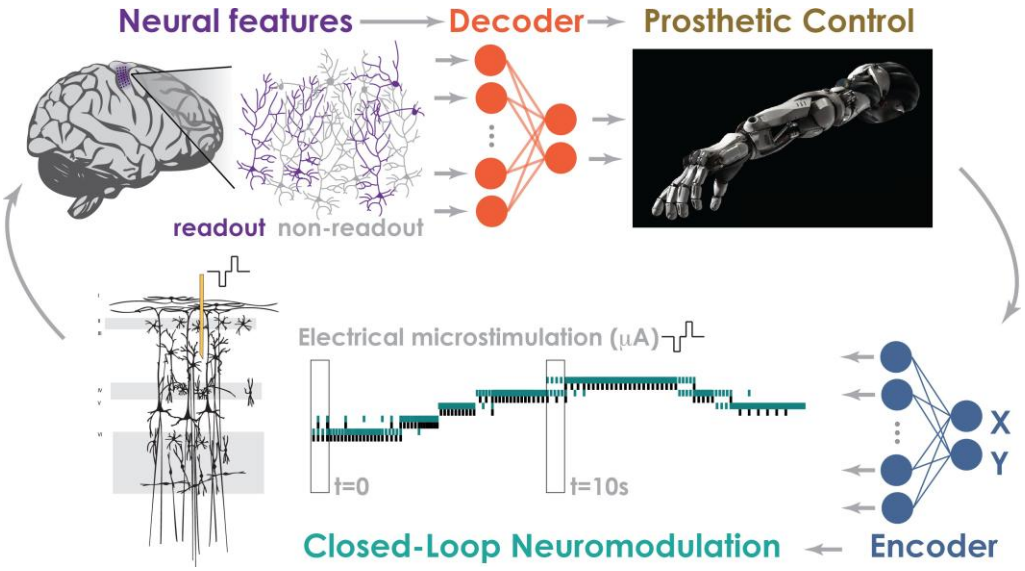


Figure 5: Closed-loop neural prostheses require real-time processing (decoding) of recordings from large neural populations to translate activity patterns into control of a prosthetic (top). Reciprocally, feedback from sensors embedded in the prosthetic must be encoded into spatiotemporal patterns of neuromodulation of the brain to convey artificial sensory feedback that allows for precise motor control (bottom).

is controlling how and when stimulation is delivered. One form of stimulation control with the potential to deliver improved therapeutic outcomes is closed-loop control, where stimulation parameters are selectively adjusted based on symptom states in real time. However, very few devices are currently approved for closed-loop neuromodulation treatment in humans, and existing devices are limited in their ability to detect simple features from neural signals to control stimulation. Thus, there is an unmet need to accelerate the development of next-generation closed-loop control systems to fully unlock the potential for closed-loop neuromodulation for treating currently intractable neurological disorders.

One complex application that requires extensive real-time processing is a closed-loop Brain-Machine Interface (BMI). While natural sensory feedback like vision allows almost all BMIs to be closed-loop systems, the precision and accuracy of movements made with a prosthesis will be limited in the absence of proprioceptive and tactile feedback [32, 33]. Therefore, advances in motor BMIs require that we move past purely open-loop machine-learning approaches to optimize closed-loop performance [34]. Doing so will require careful consideration of all aspects of the system, from neural features to control dynamics and how they interact with the brain's learning computations.

Fast Machine Learning in Quantum Information Science

Excerpt from [35] – Quantum technologies hold the promise to transform a range of applications from computation and communication to sensing. However, realizing these quantum advantages requires scaling from current small-scale prototypes to large-scale quantum processors with increasingly complex qubit arrays. As the number of qubits grows, so does the need for a commensurately scalable and efficient classical co-processing infrastructure. The software, firmware, and hardware responsible for controlling and reading out the quantum states must not only support the expanding qubit counts but also maintain the high fidelity and low latency critical for successful quantum operations.

Parallel to this hardware development, a crucial need exists to integrate classical and quantum hardware

with a robust and scalable software stack. Software challenges include device calibration and tune-up, multiplexed readout, and fast adaptive control. These challenges span both classical and quantum control in the presence of noise, crosstalk, and other errors. For future fault-tolerant quantum machines, fast, high-fidelity measurements across large systems and real-time, low-latency adaptive feedback control are essential [36, 37].

The rise of AI tools in classic computing provides a powerful toolkit for developing adaptive, heuristic algorithms for qubit readout and control. In particular, AI can be used in readout to account for effects that are difficult to compute analytically, such as non-linear behaviors in the time-evolution of signal traces, including noise, multi-qubit correlations, and crosstalk, and the evolution of the system dynamics over time due to external effects [38–42]. Ultimately, an active learning AI approach can provide a path to high-fidelity autonomous and adaptive qubit readout control systems. Toward this goal, it is essential to develop a platform for researchers to study, train, and deploy AI algorithms, bringing together expertise in quantum systems, AI, and readout electronics that meet the constraints of state-of-the-art experiments.

Fast Machine Learning as an Educational Tool

As an educational tool, Fast Machine Learning has helped bring awareness to distinct problems often not highlighted within other contexts. In particular, the science drivers have motivated the need for hardware-aware optimizations in AI algorithm design [1] to yield algorithms that can work in the low-latency/low-power regime. As a result, the workforce that has emerged from the Fast Machine Learning Organization is aware of processing technology and how AI interfaces with this technology. As an example, it is typical for students and postdocs who have worked on Fast Machine Learning problems to train with quantized AI toolkits, such as the Fast Machine Learning toolkit QKeras [43]. Researchers with both an awareness of processors and AI algorithms have often driven many AI advancements. The original transformer paper primarily focuses on the strategies to ensure optimized transformer implementation on a GPU [44]. As the AI toolkits become easier to use and the focus has changed towards the capabilities of AI models, particularly LLMs, this dual awareness of processor and AI has become less prominent within the AI community despite being an essential element.

Beyond developing a workforce, the Fast Machine Learning community provides benchmarks and datasets on which the community can build. The critical scientific problems the Fast Machine Learning community is pursuing are leading to new benchmarks and machine learning challenges [45, 46]. The challenges help to define clear benchmarking problems and grow the community. The most recent machine learning challenge organized by the A3D3 organization yielded more than 610 participating teams; many of these teams were students eager to understand critical scientific AI problems. Out of challenges emerge new industry and scientific benchmarks that help bring awareness to essential AI problems emerging within the field.

Fast Machine Learning Current and Future Initiatives

While we have presented many different science domains, we want to stress that the origins of this organization started from trying to solve a specific problem with high-energy physics. Despite that, we continue to find more scientific fields that benefit from this work. With each domain, we see slightly different challenges that extend our common toolkit, providing further solutions for the rest of the community. We are finding that more and more domains need low latency and high throughput systems and have been able to rapidly solve critical problems through common sharing of the same knowledgebase

and toolkit developed and preserved by the Fast Machine Learning Community.

As processing technology continues to improve, these ideas will become more relevant, especially when considering strategies to improve the computing power budget, make AI more efficient, and make AI run faster. Additionally, new technologies such as nanophotonics [47] could dramatically improve AI latency and power usage. These new technologies often suffer from similar design limitations that mimic the current challenges in ASIC and FPGA design, and we are observing common solutions and science drivers that build on the community's experience are aiding the advancement of these new technologies. Beyond science, Fast Machine Learning is starting to impact other commercial domains, such as low-power, wearable health monitoring, low-power image recognition, and satellite controls. While Fast Machine Learning has enabled new strategies in commercial domains, we stress that scientific problems drive the origins of this new technology and are the main focus within the Fast Machine Organization.

Conclusions

This document presents an overview of the driving science within the Fast Machine Learning organization. The organization aims to use the scientific demands for low latency and high throughput as an engine to advance new designs for processor design, optimizations for AI algorithm design, and flexible software toolkits to integrate AI into current experiments. The existing work within the organization has led to strategies to optimize heterogeneous computing clusters (GPU and other processors) to reduce the total power budget and latency for AI inference. Furthermore, this work has led to faster, more efficient algorithms through optimized chip design for AI architectures embedded on semiconductor chips, including Field Programmable Gate Arrays and Application-Specific Integrated Circuit processors. While the initial conception of this work started to solve high-energy physics problems, the solutions developed within the Fast Machine Learning Organization are opening a broad range of scientific possibilities ranging from quantum information systems to neuroscience. Moreover, the concepts are starting to yield solutions beyond science impacting society as a whole [1–3,48].

Cultivating a cross-disciplinary effort that brings together the many scientific drivers is crucial to preserve the ecosystem and to advance developments in Artificial Intelligence algorithms and chip design. An institute that captures these many scientific domains and produces toolkits and a knowledge base capable of deploying Fast AI algorithms would help preserve and grow the knowledge base. The scientific challenges presented within this action plan each put different demands on what is needed to ensure that algorithms can be run at high throughput and low latency. Their combined efforts are seeds for the creation of cross-domain toolkits capable of deploying Artificial Intelligence algorithms both within their domains and beyond. With existing toolkits, such as HLS4ML and SONIC [4, 49],the Fast Machine Learning Organization is already enabling a community to deploy artificial intelligence in many domains. While working as a single entity, this organization is funded through many different initiatives. The single largest support is from the NSF A3D3 institute, whose mandate ends within 2 years. Continued support for a cross-disciplinary institute would further advance this emerging artificial intelligence field.

The work within the Fast Machine Learning Organization goes beyond the conventional Artificial Intelligence toolkit and enables algorithms to be deployed within systems where computing throughput, latency, and integration demands were previously impossible. The knock-on impacts of this work are substantial and are leading to new devices that will have significant impacts beyond science. Ultimately, the fast machine learning community aims to bring the learned knowledge to the public to ensure that everyone benefits from the technological advances that rapidly appear as artificial intelligence advances.

References

[1] A. M. Deiana *et al.*, "Applications and Techniques for Fast Machine Learning in Science," *Front. Big Data*, vol. 5, p. 787421, 2022. doi: 10.3389/fdata.2022.787421

[2] F. Fahim, B. Hawks, C. Herwig, J. Hirschauer, S. Jindariani, N. Tran, L. P. Carloni, G. Di Guglielmo, P. Harris, J. Krupa *et al.*, "hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices," *arXiv preprint arXiv:2103.05579*, 2021.

[3] J. Duarte, S. Han, P. Harris, S. Jindariani, E. Kreinar, B. Kreis, J. Ngadiuba, M. Pierini, R. Rivera, N. Tran *et al.*, "Fast inference of deep neural networks in fpgas for particle physics," *Journal of instrumentation*, vol. 13, no. 07, p. P07027, 2018.

[4] J. Krupa *et al.*, "GPU coprocessors as a service for deep learning inference in high energy physics," *Mach. Learn. Sci. Tech.*, vol. 2, no. 3, p. 035005, 2021. doi: 10.1088/2632-2153/abec21

[5] J. Duarte *et al.*, "Fast inference of deep neural networks in FPGAs for particle physics," *JINST*, vol. 13, no. 07, p. P07027, 2018. doi: 10.1088/1748-0221/13/07/P07027

[6] G. Abarajithan, Z. Ma, Z. Li, S. Koparkar, R. Munasinghe, F. Restuccia, and R. Kastner, "Cgra4ml: A framework to implement modern neural networks for scientific edge computing," 2024. [Online]. Available: https://arxiv.org/abs/2408.15561

[7] O. Weng, A. Redding, N. Tran, J. M. Duarte, and R. Kastner, "Architectural implications of neural network inference for high data-rate, low-latency scientific applications," 2024. [Online]. Available: https://arxiv.org/abs/2403.08980

[8] M. Agarwal *et al.*, "Applications of Deep Learning to physics workflows," 6 2023.

[9] M. Wang, T. Yang, M. Acosta Flechas, P. Harris, B. Hawks, B. Holzman, K. Knoepfel, J. Krupa, K. Pedro, and N. Tran, "GPU-Accelerated Machine Learning Inference as a Service for Computing in Neutrino Experiments," *Front. Big Data*, vol. 3, p. 604083, 2021. doi: 10.3389/fdata.2020.604083

[10] T. Cai, K. Herner, T. Yang, M. Wang, M. A. Flechas, P. Harris, B. Holzman, K. Pedro, and N. Tran, "Accelerating Machine Learning Inference with GPUs in ProtoDUNE Data Processing," *Comput. Softw. Big Sci.*, vol. 7, no. 1, p. 11, 2023. doi: 10.1007/s41781-023-00101-0

[11] A. Gunny, D. Rankin, J. Krupa, M. Saleem, T. Nguyen, M. Coughlin, P. Harris, E. Katsavounidis, S. Timm, and B. Holzman, "Hardware-accelerated Inference for Real-Time Gravitational-Wave Astronomy," *Nature Astron.*, vol. 6, no. 5, pp. 529–536, 2022. doi: 10.1038/s41550-022-01651-w

[12] CMS Collaboration, "Measurement of boosted Higgs bosons produced via vector boson fusion or gluon fusion in the H $\to b\bar{b}$ decay mode using LHC proton-proton collision data at $\sqrt{s}$ = 13 TeV," *JHEP*, vol. 12, p. 035, 2024. doi: 10.1007/JHEP12(2024)035

[13] ——"Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s}$ = 13 TeV," 12 2024.

[14] Y. Wei, R. F. Forelli, C. Hansen, J. P. Levesque, N. Tran, J. C. Agar, G. Di Guglielmo, M. E. Mauel, and G. A. Navratil, "Low latency optical-based mode tracking with machine learning deployed on FPGAs on a tokamak," *Rev. Sci. Instrum.*, vol. 95, no. 7, p. 073509, 2024. doi: 10.1063/5.0190354

[15] M. S. Chu and M. Okabayashi, "Stabilization of the external kink and the resistive wall mode," *Plasma Physics and Controlled Fusion*, vol. 52, no. 12, p. 123001, oct 2010. doi: 10.1088/0741-3335/52/12/123001. [Online]. Available: https://dx.doi.org/10.1088/0741-3335/52/12/123001

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nat.*, vol. 521, no. 7553, pp. 436–444, May 2015. doi: 10.1038/nature14539

[17] A. Jalalvand, J. Abbate, R. Conlin, G. Verdoolaege, and E. Kolemen, "Real-time and adaptive reservoir computing with application to profile prediction in fusion plasma," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2630–2641, 2022. doi: 10.1109/TNNLS.2021.3085504

[18] M. Boyer and J. Chadwick, "Prediction of electron density and pressure profile shapes on nstx-u using neural networks," *Nuclear Fusion*, vol. 61, no. 4, p. 046024, mar 2021. doi: 10.1088/1741-4326/abe08b. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/abe08b

[19] J. Zhu, C. Rea, R. Granetz, E. Marmar, R. Sweeney, K. Montes, and R. Tinguely, "Integrated deep learning framework for unstable event identification and disruption prediction of tokamak plasmas," *Nuclear Fusion*, vol. 63, no. 4, p. 046009, mar 2023. doi: 10.1088/1741-4326/acb803. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/acb803

[20] J. Kates-Harbeck, A. Svyatkovskiy, and W. Tang, "Predicting disruptive instabilities in controlled fusion plasmas through deep learning," *Nat.*, vol. 568, no. 7753, pp. 526–531, Apr. 2019. doi: 10.1038/s41586-019-1116-4

[21] J. Abbate, R. Conlin, R. Shousha, K. Erickson, and E. Kolemen, "A general infrastructure for data-driven control design and implementation in tokamaks," *Journal of Plasma Physics*, vol. 89, no. 1, p. 895890102, 2023. doi: 10.1017/S0022377822001040

[22] Y. Wei, J. P. Levesque, C. J. Hansen, M. E. Mauel, and G. A. Navratil, "A dimensionality reduction algorithm for mapping tokamak operational regimes using a variational autoencoder (VAE) neural network," *Nuclear Fusion*, vol. 61, no. 12, p. 126063, Dec. 2021. doi: 10.1088/1741-4326/ac3296

[23] J. e. a. Degrave, "Magnetic control of tokamak plasmas through deep reinforcement learning," *Nat*, vol. 602, no. 7897, pp. 414–419, Feb. 2022. doi: 10.1038/s41586-021-04301-9

[24] S. Dubbioso, G. De Tommasi, A. Mele, G. Tartaglione, M. Ariola, and A. Pironti, "A deep reinforcement learning approach for vertical stabilization of tokamak plasmas," *Fusion Engineering and Design*, vol. 194, p. 113725, 2023. doi: https://doi.org/10.1016/j.fusengdes.2023.113725. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920379623003083

[25] J. Seo, Y. S. Na, B. Kim, C. Y. Lee, M. S. Park, S. J. Park, and Y. H. Lee, "Development of an operation trajectory design algorithm for control of multiple 0D parameters using deep reinforcement learning in KSTAR," *Nuclear Fusion*, vol. 62, no. 8, p. 086049, Aug. 2022. doi: 10.1088/1741-4326/ac79be

[26] I. Char, J. Abbate, L. Bardoczi, M. Boyer, Y. Chung, R. Conlin, K. Erickson, V. Mehta, N. Richner, E. Kolemen, and J. Schneider, "Offline model-based reinforcement learning for tokamak control," in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, ser. Proceedings of Machine Learning Research, N. Matni, M. Morari, and G. J. Pappas, Eds., vol. 211. PMLR, 15–16 Jun 2023, pp. 1357–1372. [Online]. Available: https://proceedings.mlr.press/v211/char23a.html

[27] G. Shin, H. Han, M. Kim, S.-H. Hahn, W. Ko, G. Park, Y. Lee, M. Lee, M. Kim, J.-W. Juhn, D. Seo, J. Jang, H. Kim, J. Lee, and H. Kim, "Preemptive rmp-driven elm crash suppression automated by a real-time machine-learning classifier in kstar," *Nuclear Fusion*, vol. 62, no. 2, p. 026035, jan 2022. doi: 10.1088/1741-4326/ac412d. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/ac412d

[28] A. Piccione, J. Berkery, S. Sabbagh, and Y. Andreopoulos, "Predicting resistive wall mode stability in nstx through balanced random forests and counterfactual explanations," *Nuclear Fusion*, vol. 62, no. 3, p. 036002, jan 2022. doi: 10.1088/1741-4326/ac44af. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/ac44af

[29] Y. Wei, J. P. Levesque, C. Hansen, M. E. Mauel, and G. A. Navratil, "Mhd mode tracking using high-speed cameras and deep learning," *Plasma Physics and Controlled Fusion*, vol. 65, no. 7, p. 074002, may 2023. doi: 10.1088/1361-6587/acd581. [Online]. Available: https://dx.doi.org/10.1088/1361-6587/acd581

[30] J. Song, S. Joung, Y.-C. Ghim, S. hee Hahn, J. Jang, and J. Lee, "Development of machine learning model for automatic elm-burst detection without hyperparameter adjustment in kstar tokamak," *Nuclear Engineering and Technology*, vol. 55, no. 1, pp. 100–108, 2023. doi: https://doi.org/10.1016/j.net.2022.08.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1738573322004077

[31] K. K. Sellers, J. L. Cohen, A. N. Khambhati, J. M. Fan, A. M. Lee, E. F. Chang, and A. D. Krystal, "Closed-loop neurostimulation for the treatment of psychiatric disorders," *Neuropsychopharmacology*, vol. 49, no. 1, pp. 163–178, 2024. doi: 10.1038/s41386-023-01631-2

[32] S. N. Flesher, J. E. Downey, J. M. Weiss, C. L. Hughes, A. J. Herrera, E. C. Tyler-Kabara, M. L. Boninger, J. L. Collinger, and R. A. Gaunt, "A brain-computer interface that evokes tactile sensations improves robotic arm control," *Science*, vol. 372, no. May, pp. 831–836, 2021.

[33] R. L. Sainburg, H. Poizner, and C. Ghez, "Loss of proprioception produces deficits in interjoint coordination," *Journal of Neurophysiology*, vol. 70, no. 5, pp. 2136–2147, 1993. doi: 10.1152/jn.1993.70.5.2136

[34] M. C. Dadarlat, R. A. Canfield, and A. L. Orsborn, "Neural Plasticity in Sensorimotor Brain-Machine Interfaces," *Annual Review of Biomedical Engineering*, vol. 15, no. 2, pp. 51–76, 2023.

[35] G. Di Guglielmo *et al.*, "End-to-end workflow for machine learning-based qubit readout with QICK and hls4ml," 1 2025.

[36] M. E. Beverland, P. Murali, M. Troyer, K. M. Svore, T. Hoefler, V. Kliuchnikov, G. H. Low, M. Soeken, A. Sundaram, and A. Vaschillo, "Assessing requirements to scale to practical quantum advantage," 2022. [Online]. Available: https://arxiv.org/abs/2211.07629

[37] M. Mohseni, A. Scherer, K. G. Johnson, O. Wertheim, M. Otten, N. A. Aadit, K. M. Bresniker, K. Y. Camsari, B. Chapman, S. Chatterjee, G. A. Dagnew, A. Esposito, F. Fahim, M. Fiorentino, A. Khalid, X. Kong, B. Kulchytskyy, R. Li, P. A. Lott, I. L. Markov, R. F. McDermott, G. Pedretti, A. Gajjar, A. Silva, J. Sorebo, P. Spentzouris, Z. Steiner, B. Torosov, D. Venturelli, R. J. Visser, Z. Webb, X. Zhan, Y. Cohen, P. Ronagh, A. Ho, R. G. Beausoleil, and J. M. Martinis, "How to build a quantum supercomputer: Scaling challenges and opportunities," 2024. [Online]. Available: https://arxiv.org/abs/2411.10406

[38] P. K. Gautam, S. Kalipatnapu, S. H, U. Singhal, B. Lienhard, V. Singh, and C. S. Thakur, "Low-latency machine learning fpga accelerator for multi-qubit-state discrimination," 2024. [Online]. Available: https://arxiv.org/abs/2407.03852

[39] P. Duan, Z. F. Chen, Q. Zhou, W. C. Kong, H. F. Zhang, and G. P. Guo, "Mitigating crosstalk-induced qubit readout error with shallow-neural-network discrimination," *Physical Review Applied*, vol. 16, no. 2, p. 1, 2021. [Online]. Available: https://doi.org/10.1103/PhysRevApplied.16.024063

[40] E. Magesan, J. M. Gambetta, A. D. Córcoles, and J. M. Chow, "Machine learning for discriminating quantum measurement trajectories and improving readout," *Phys. Rev. Lett.*, vol. 114, p. 200501, May 2015. doi: 10.1103/PhysRevLett.114.200501. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.114.200501

[41] S. Maurya, C. N. Mude, W. D. Oliver, B. Lienhard, and S. Tannu, "Scaling qubit readout with hardware efficient machine learning architectures," in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, ser. ISCA '23. New York, NY, USA: Association for Computing Machinery, 2023. doi: 10.1145/3579371.3589042. ISBN 9798400700958. [Online]. Available: https://doi.org/10.1145/3579371.3589042

[42] N. R. Vora, Y. Xu, A. Hashim, N. Fruitwala, H. N. Nguyen, H. Liao, J. Balewski, A. Rajagopala, K. Nowrouzi, Q. Ji, K. B. Whaley, I. Siddiqi, P. Nguyen, and G. Huang, "Ml-powered fpga-based real-time quantum state discrimination enabling mid-circuit measurements," 2024. [Online]. Available: https://arxiv.org/abs/2406.18807

[43] C. N. Coelho, A. Kuusela, S. Li, H. Zhuang, T. Aarrestad, V. Loncar, J. Ngadiuba, M. Pierini, A. A. Pol, and S. Summers, "Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors," *Nature Mach. Intell.*, vol. 3, pp. 675–686, 2021. doi: 10.1038/s42256-021-00356-5

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[45] J. Duarte, N. Tran, B. Hawks, C. Herwig, J. Muhizi, S. Prakash, and V. J. Reddi, "FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning," in *5th Conference on Machine Learning and Systems*, 7 2022.

[46] E. G. Campolongo, Y.-T. Chou, E. Govorkova, W. Bhimji, W.-L. Chao, C. Harris, S.-C. Hsu, H. Lapp, M. S. Neubauer, J. Namayanja, A. Subramanian, P. Harris, A. Anand, D. E. Carlyn, S. Ghosh, C. Lawrence, E. Moreno, R. Raikman, J. Wu, Z. Zhang, and others, "Building Machine Learning Challenges for Anomaly Detection in Science," *arXiv e-prints*, p. arXiv:2503.02112, Mar. 2025. doi: 10.48550/arXiv.2503.02112

[47] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, no. 7, p. 441–446, Jun. 2017. doi: 10.1038/nphoton.2017.93. [Online]. Available: http://dx.doi.org/10.1038/nphoton.2017.93

[48] F. M. L. Collaboration, "fastmachinelearning/hls4ml," *Version v0*, vol. 8, 2023.

[49] Fast Machine Learning Collaboration, "fastmachinelearning/hls4ml," 2023. [Online]. Available: https://github.com/fastmachinelearning/hls4ml