Strategic vision

AI in the world and at the DOE

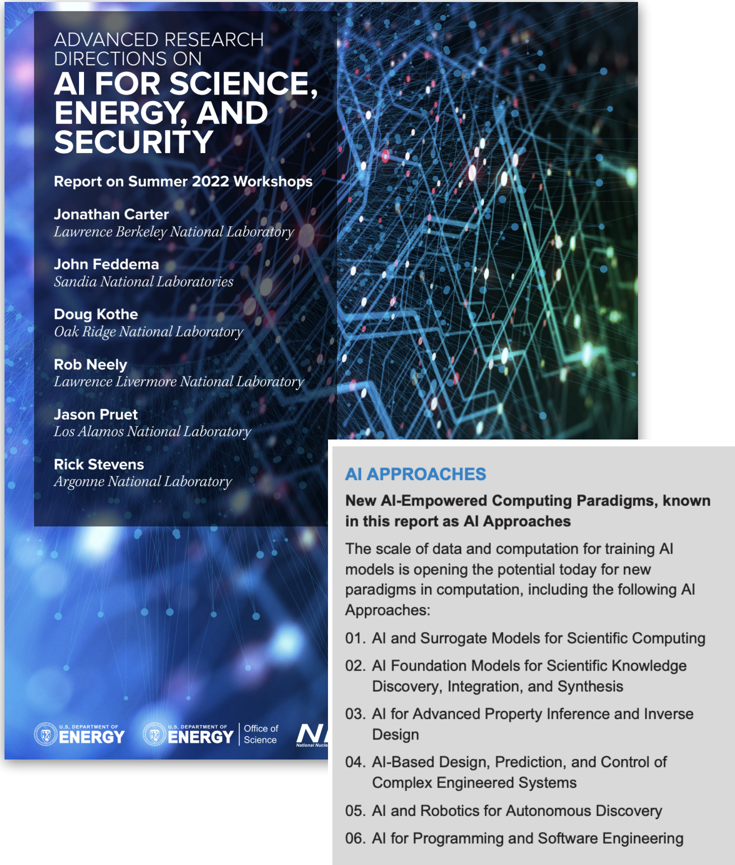
report

- AI is a big space: industry & academia
  - ChatGPT was a game-changer

- The DOE supercomputing facilities are a key resource (ORNL, ANL, LBNL)

- **Where does HEP mission and Fermilab fit into this picture?**

Fermilab strength in **intelligent sensing** and real-time efficient AI towards vision of accelerating scientific discovery at unprecedented data scales



ADVANCED RESEARCH DIRECTIONS ON

AI FOR SCIENCE, ENERGY, AND SECURITY

Report on Summer 2022 Workshops

Jonathan Carter
Lawrence Berkeley National Laboratory

John Feddema
Sandia National Laboratories

Doug Kothe
Oak Ridge National Laboratory

Rob Neely
Lawrence Livermore National Laboratory

Jason Pruet
Los Alamos National Laboratory

Rick Stevens
Argonne National Laboratory

U.S. DEPARTMENT OF ENERGY | U.S. DEPARTMENT OF ENERGY | Office of Science | N

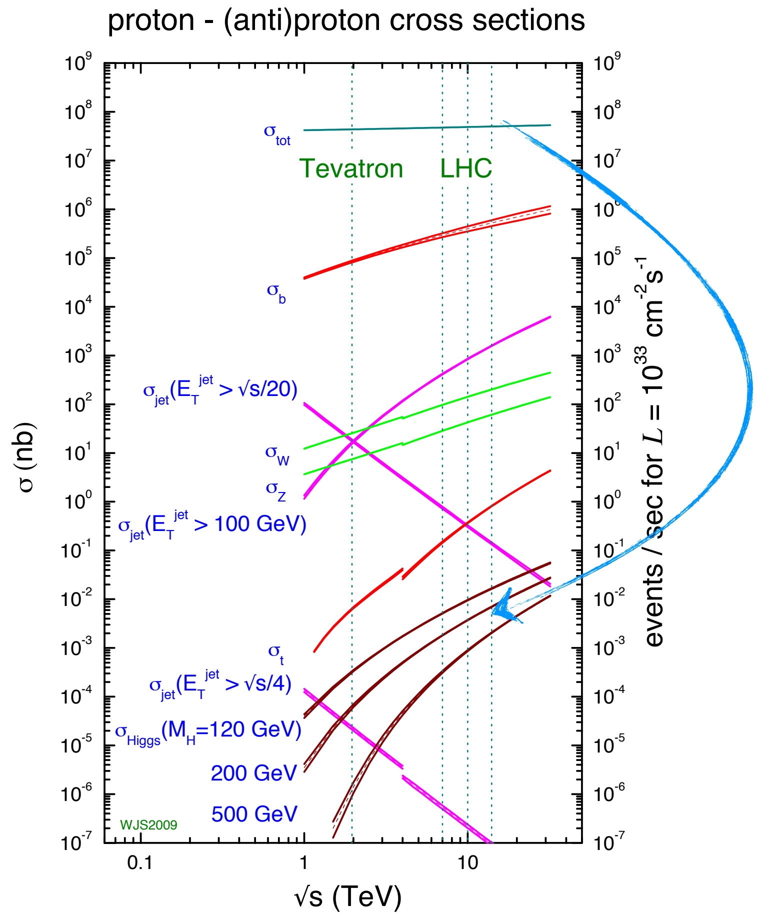
AI APPROACHES

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

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

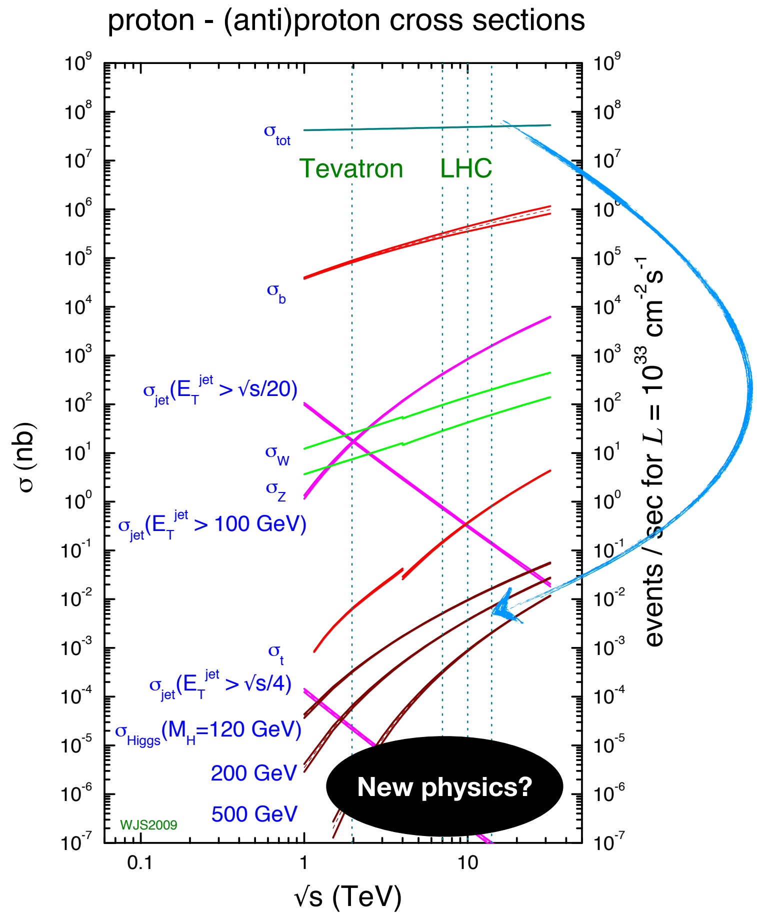
01. AI and Surrogate Models for Scientific Computing
02. AI Foundation Models for Scientific Knowledge Discovery, Integration, and Synthesis
03. AI for Advanced Property Inference and Inverse Design
04. AI-Based Design, Prediction, and Control of Complex Engineered Systems
05. AI and Robotics for Autonomous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

LHC Example

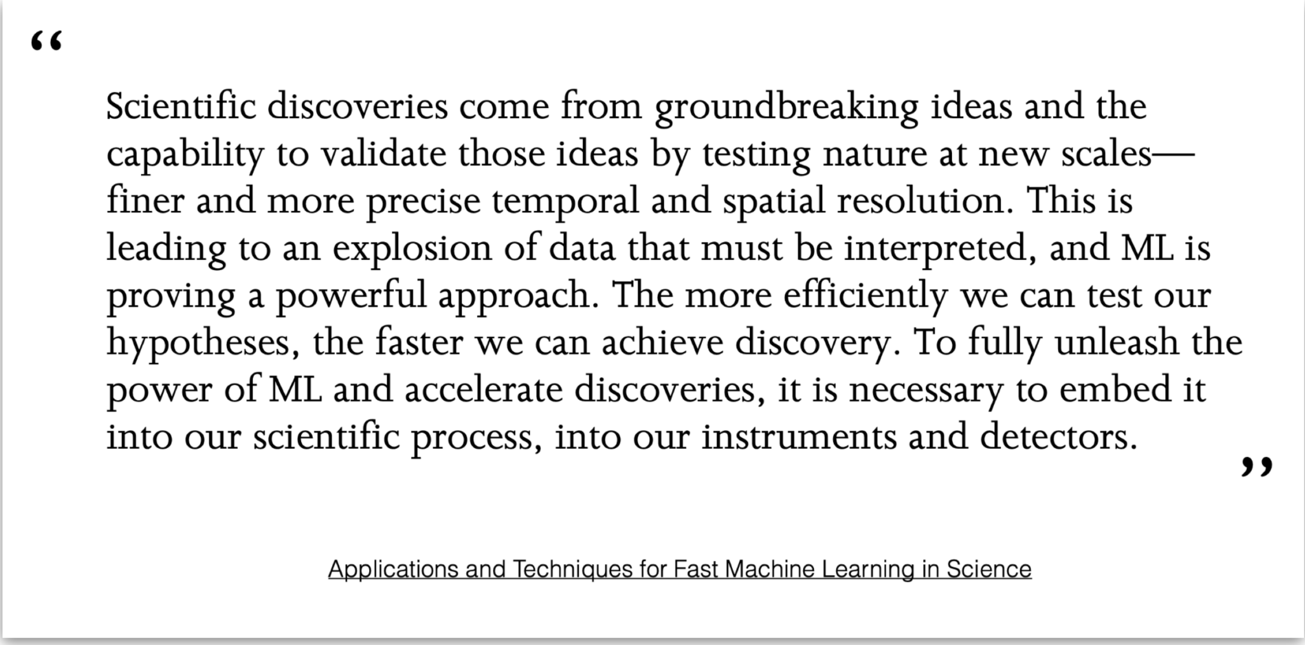


Higgs produced once every billion collisions

LHC Example



Higgs produced once every billion collisions



> “Scientific discoveries come from groundbreaking ideas and the capability to validate those ideas by testing nature at new scales—finer and more precise temporal and spatial resolution. This is leading to an explosion of data that must be interpreted, and ML is proving a powerful approach. The more efficiently we can test our hypotheses, the faster we can achieve discovery. To fully unleash the power of ML and accelerate discoveries, it is necessary to embed it into our scientific process, into our instruments and detectors.”

Applications and Techniques for Fast Machine Learning in Science

Core ML Mission: Efficient, robust, autonomous ML codesign

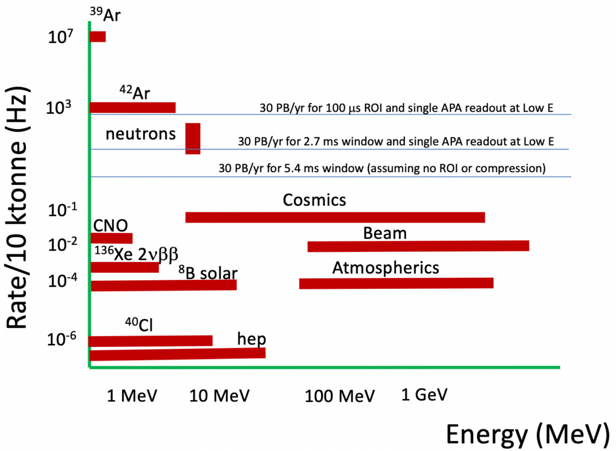
Grand challenges for HEP, examples

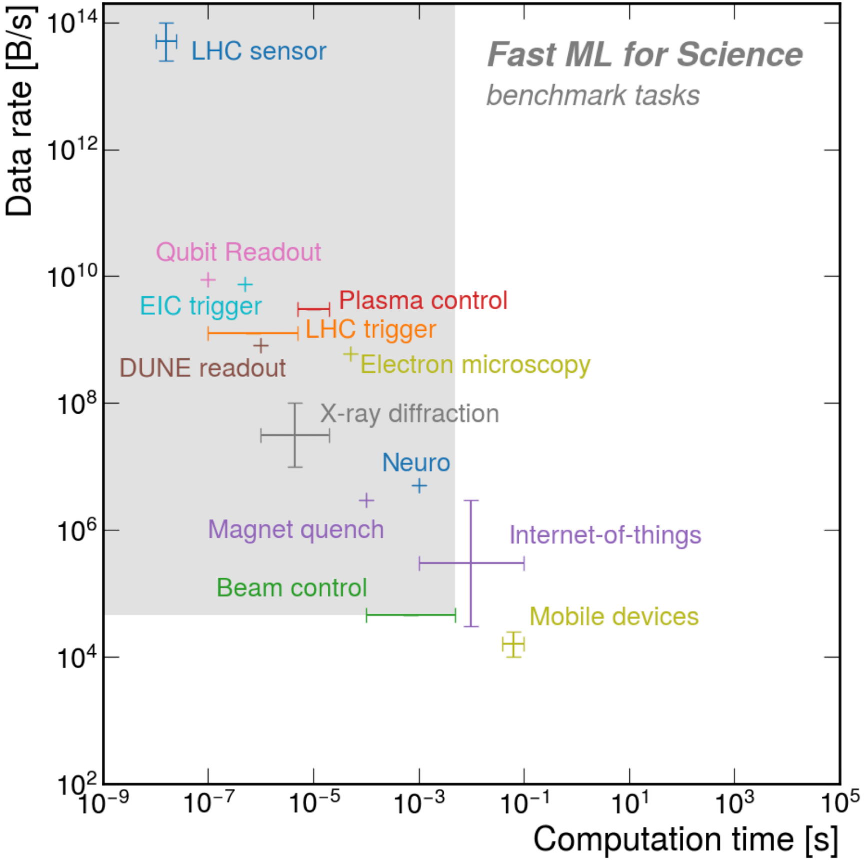
Optimal, continuous readout for DUNE for neutrino physics, multi-messenger astronomy, and other rare measurements

Analyze all 40 MHz of LHC data for the full detector for new physics searches, Higgs measurements, and more

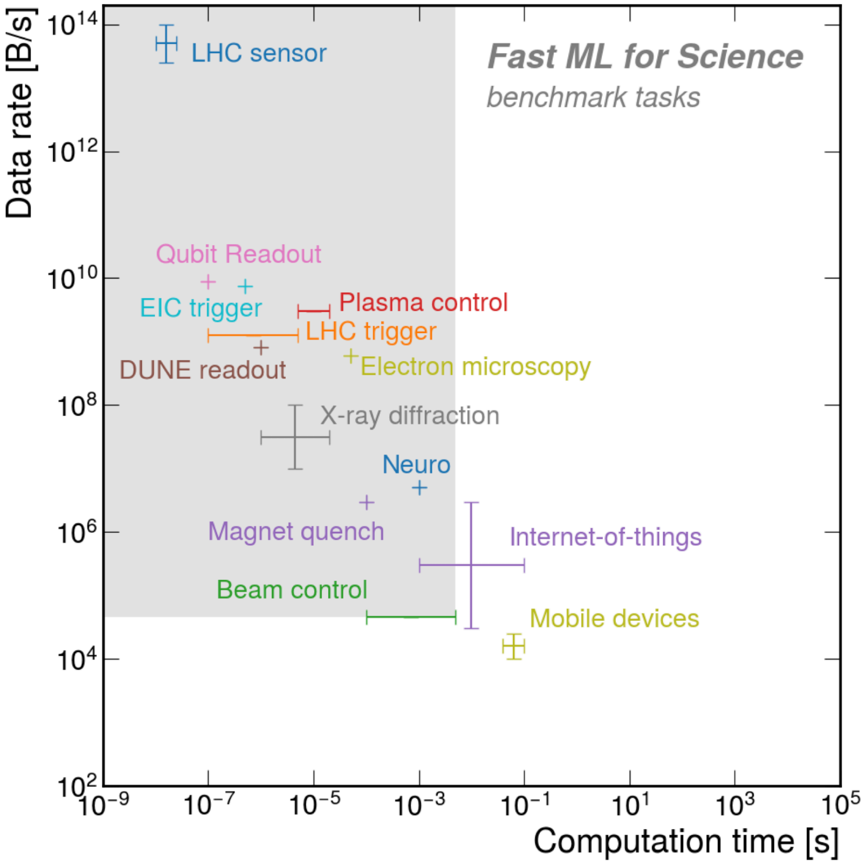
AI-assisted, real-time operation of the Fermilab accelerator complex

Fermilab accelerator complex


Rate/10 ktonne (Hz)
10^7
10^3
10^-1
10^-2
10^-4
10^-6
39Ar
42Ar
neutrons
CNO
136Xe 2νββ
8B solar
40Cl
hep
Cosmics
Beam
Atmospherics
30 PB/yr for 100 µs ROI and single APA readout at Low E
30 PB/yr for 2.7 ms window and single APA readout at Low E
30 PB/yr for 5.4 ms window (assuming no ROI or compression)
1 MeV
10 MeV
100 MeV
1 GeV
Energy (MeV)



Grand challenges spark imaginations!
Benchmarks bring innovation



Grand challenges spark imaginations!
Benchmarks bring innovation

Benefits to HEP: bring new resources to bear on HEP grand challenges (industry partnerships, computer science & engineering researchers)

HEP-born technology brings **transformative technology** to new material research, fusion energy, neuroscience, or industry applications and so on…

Impact: recent *examples*

Leverage core capabilities to deploy **ML <u>at scale</u>** - algorithms + facilities, tools, software, multidisciplinary teams

e.g. large scale user facilities & advanced instrumentation; advanced computer science, visualization, & data; microelectronics

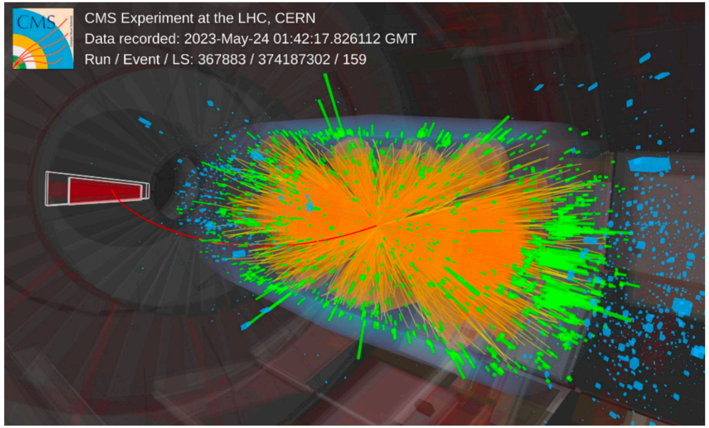
Impact: recent *examples*

Leverage core capabilities to deploy **ML <u>at scale</u>** - algorithms + facilities, tools, software, multidisciplinary teams

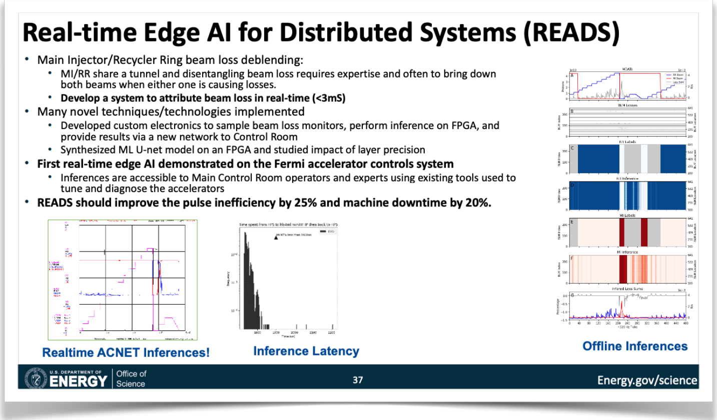
e.g. large scale user facilities & advanced instrumentation; advanced computer science, visualization, & data; microelectronics

- First ever L1 trigger anomaly detection algorithm deployed for LHC CMS Run 3
  - Growth from community benchmarks and collaborations built from community efforts, investment in hls4ml (FastML, AMD, Siemens)
- CMS MLG-23-001 demonstration of accelerated ML workflows with SONIC; working with NVidia, Graphcore, computing operations experts
- First edge AI deployed in Fermilab accelerator complex; working with Intel/NU

https://cds.cern.ch/record/2876546


CMS
CMS Experiment at the LHC, CERN
Data recorded: 2023-May-24 01:42:17.826112 GMT
Run / Event / LS: 367883 / 374187302 / 159

Rameika, HEPAP Aug23



Real-time Edge AI for Distributed Systems (READS)

- Main Injector/Recycler Ring beam loss deblending:
  - MI/RR share a tunnel and disentangling beam loss requires expertise and often to bring down both beams when either one is causing losses.
  - **Develop a system to attribute beam loss in real-time (<3mS)**
- Many novel techniques/technologies implemented
  - Developed custom electronics to sample beam loss monitors, perform inference on FPGA, and provide results via a new network to Control Room
  - Synthesized ML U-net model on an FPGA and studied impact of layer precision
- **First real-time edge AI demonstrated on the Fermi accelerator controls system**
  - Inferences are accessible to Main Control Room operators and experts using existing tools used to tune and diagnose the accelerators
- **READS should improve the pulse inefficiency by 25% and machine downtime by 20%.**

Realtime ACNET Inferences!

Inference Latency

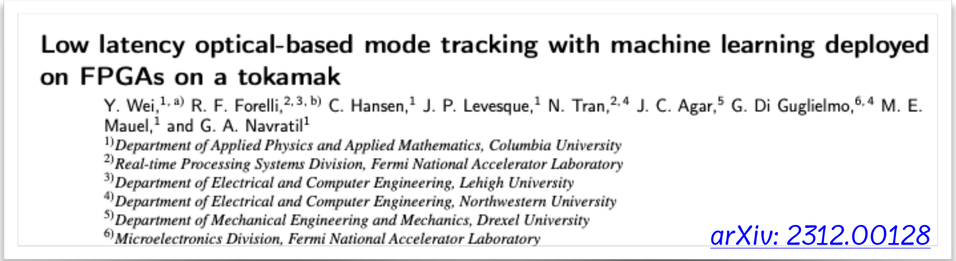
Offline Inferences

U.S. DEPARTMENT OF ENERGY | Office of Science 37 Energy.gov/science

Impact: recent *examples*

Leverage core capabilities to deploy **ML <u>at scale</u>** - algorithms + facilities, tools, software, multidisciplinary teams

e.g. large scale user facilities & advanced instrumentation; advanced computer science, visualization, & data; microelectronics

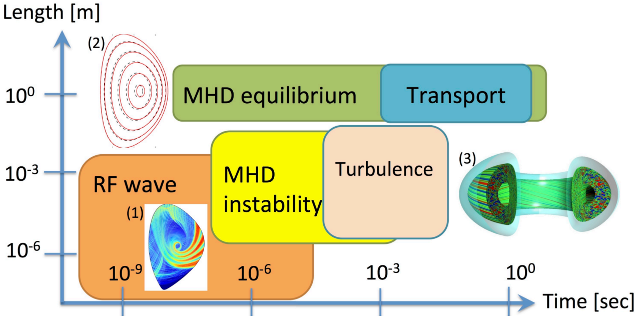


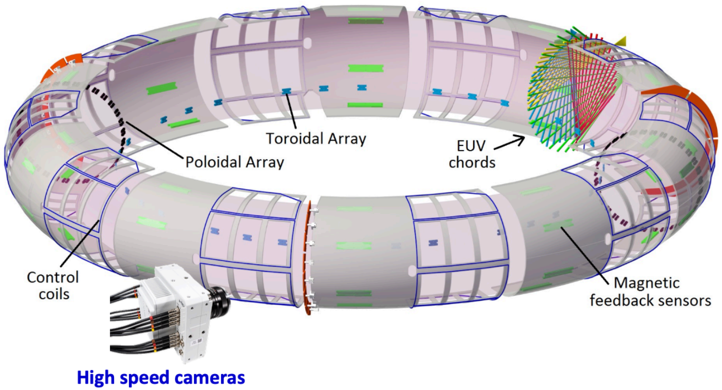
Low latency optical-based mode tracking with machine learning deployed on FPGAs on a tokamak

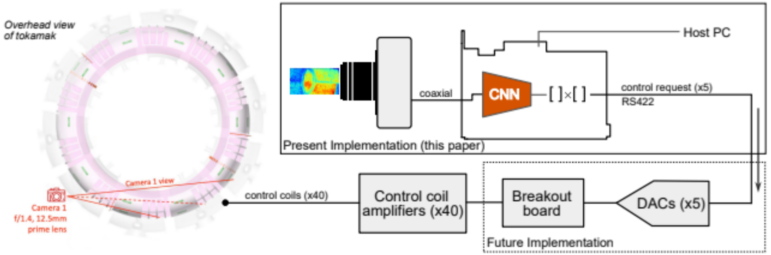
Y. Wei,[1, a)] R. F. Forelli,[2, 3, b)] C. Hansen,[1] J. P. Levesque,[1] N. Tran,[2,4] J. C. Agar,[5] G. Di Guglielmo,[6, 4] M. E. Mauel,[1] and G. A. Navratil[1]

[1] *Department of Applied Physics and Applied Mathematics, Columbia University*
[2] *Real-time Processing Systems Division, Fermi National Accelerator Laboratory*
[3] *Department of Electrical and Computer Engineering, Lehigh University*
[4] *Department of Electrical and Computer Engineering, Northwestern University*
[5] *Department of Mechanical Engineering and Mechanics, Drexel University*
[6] *Microelectronics Division, Fermi National Accelerator Laboratory*

arXiv: 2312.00128



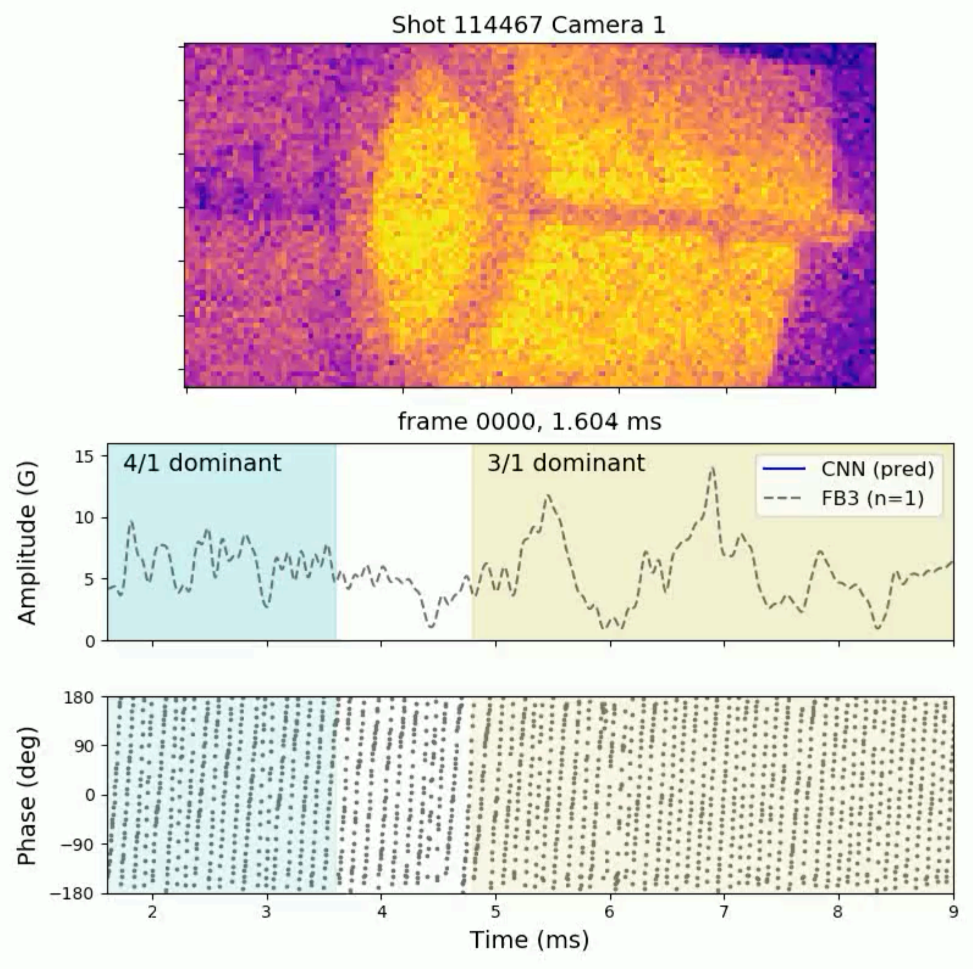




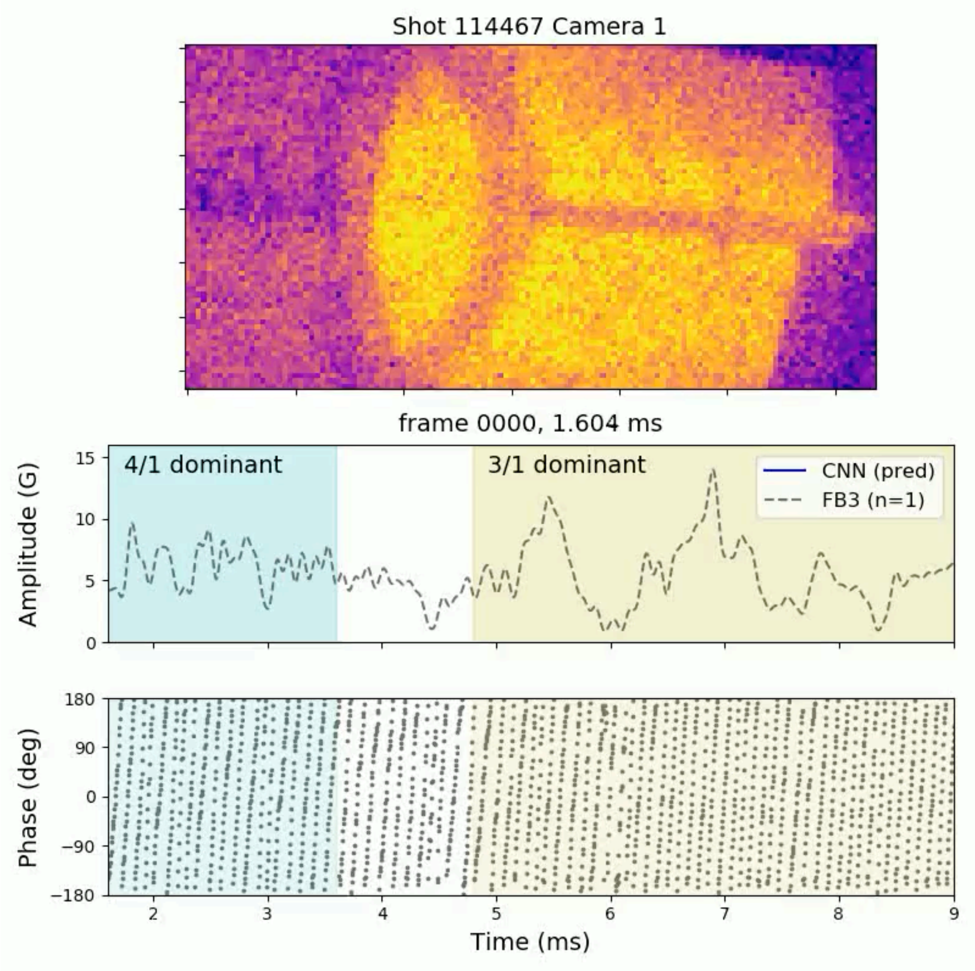
120 kfps throughput, 17.6 μs latency

Enabling new capabilities for fusion experiments!

Credit: Yumou Wei



Credit: Yumou Wei



Summary

Executive summary

Charge: Review the status of the AI/ML activities at the laboratory and of the recommendations made at past meetings: Formulate a strategy to respond to future AI/ML *research* calls, not necessarily just for AI/ML centers.

Framing: AI research is advancing rapidly; one primary area of Fermilab strength is in intelligent sensing and real-time efficient AI

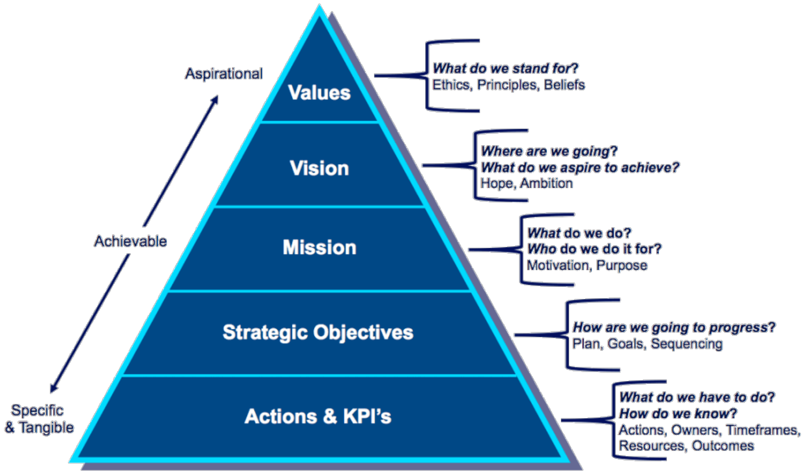
Vision: Accelerate scientific discovery at unprecedented data scales while creating enabling technology for society

Mission: Efficient, robust, autonomous ML codesign

A. Catalyze inclusive, multidisciplinary *Fast ML* community around grand challenges and benchmark tasks
B. Leverage relevant Fermilab core capabilities and strengths to build tools to support the community

Strategy:

A. Identify and grow appropriate sustainable funding streams to support community tools
B. Advance cutting-edge intelligent sensing, real-time AI research
C. Develop industry/academic partnerships to support the core mission



Key performance indicators:

1. Sustainable funding sources for supporting community tools and users on 2 year timescale
2. New and existing partnerships & collaborations resulting in: research output; new projects on AI technology and research; technology transfer; and community growth (users, downloads, etc.)