Status of AI/ML program at Fermilab

Nhan Tran on behalf of the AI Project Office and Fermilab AI community
January 17, 2023

Outline

- Vision & strategic drivers
- AI Project Office and program organization
- Program milestones and highlights
- Leveraging unique & core capabilities

Charge: We ask the PAC to review the status of the AI/ML program at the laboratory and to assess whether the laboratory is in position to make a compelling case to become an AI/ML center

Vision

ai.fnal.gov
Snowmass Talk on Cross-boundary AI

AI for physics, physics for AI

- Develop **AI capabilities to accelerate HEP science** and contribute **greater science/industry AI ecosystem**
- Build **diverse, inclusive community; assemble multi-disciplinary collaborations** around cross-cutting HEP AI challenges

Motivation

DOE HEP builds and operates among the most difficult and biggest projects with the most complex devices in science -- accelerators and detectors. Our priority is using AI for real-time controls, operations, and data processing to **accelerate HEP science**.

Pillars for AI-accelerated discovery

Algorithms for HEP science
Physics-inspired data & models; Robust & generalizable learning; Fast and efficient algorithms

Computing hardware and infrastructure

Operations and control systems

Real-time AI systems at edge

AI for HEP

Drivers to accelerate discovery

- **Deeper insights & better performance**
  Maximize science by getting the most out of machines and experiments; reduce systematics and understand anomalies
- **Accelerate time-to-physics**
  Enable powerful/robust ML at each stage of data processing; mitigate computing and data analysis challenges; automate scientific method and discovery
- **Improve operational efficiency**
  Optimize experimental “control” via triggers, data monitoring; recover lost data and physics

AI for HEP

Drivers to accelerate discovery

- **Deeper insights & better performance**
  Maximize science by getting the most out of machines and experiments; reduce systematics and understand anomalies
- **Accelerate time-to-physics**
  Enable powerful/robust ML at each stage of data processing; mitigate computing and data analysis challenges; automate scientific method and discovery
- **Improve operational efficiency**
  Optimize experimental “control” via triggers, data monitoring; recover lost data and physics

Algorithms for HEP science

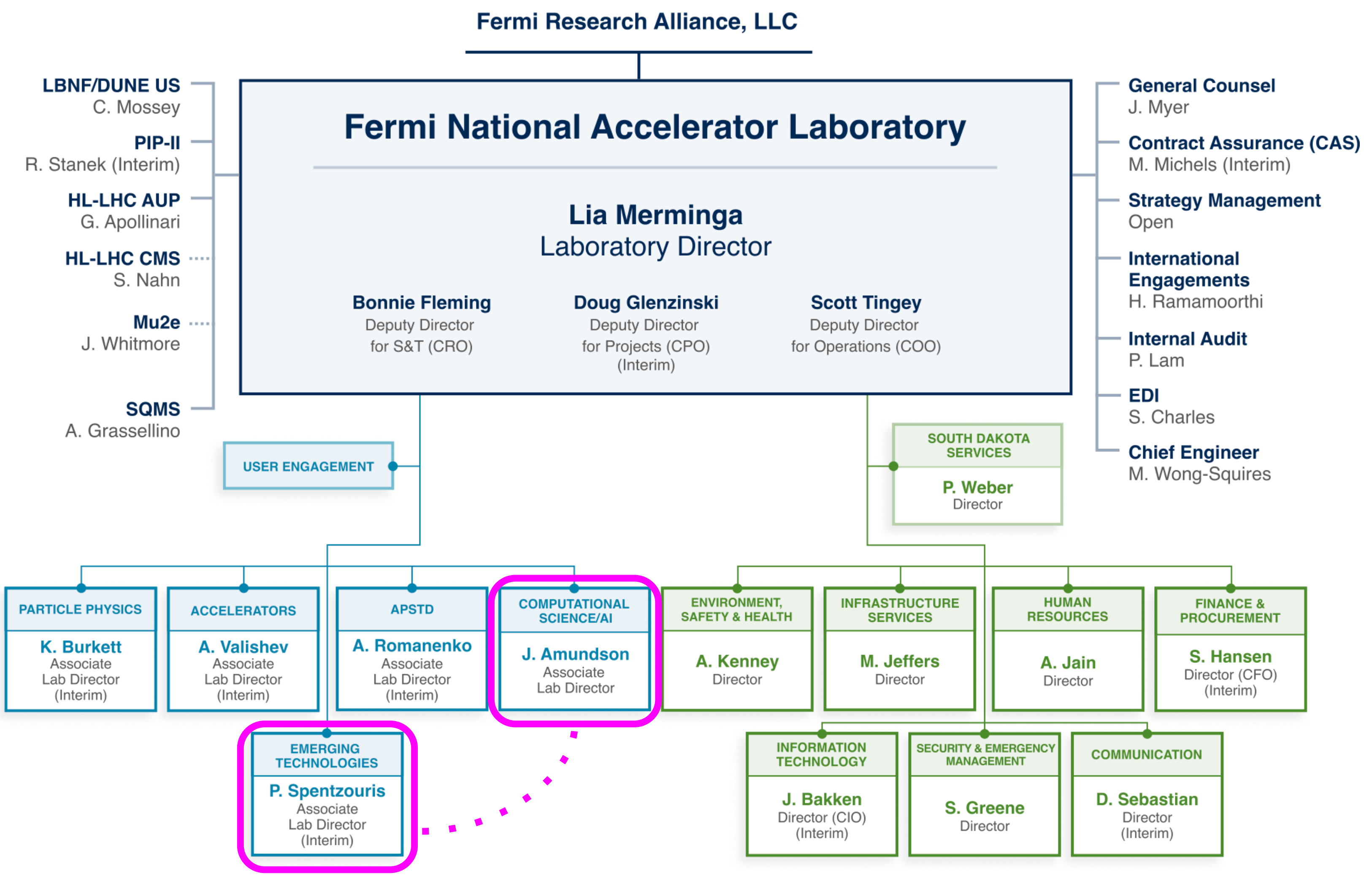
- **Physics-inspired data & models**
  Models tailored to physics and machine data representations that integrate our physics knowledge
- **Robust & generalizable learning**
  Build robust models to adapt/ extrapolate; quantify uncertainties and understand anomalies; towards explainable algorithms
- **“Fast” & efficient algorithms**
  ML in hardware-constrained environments for real-time operations and decision-making

Executive summary

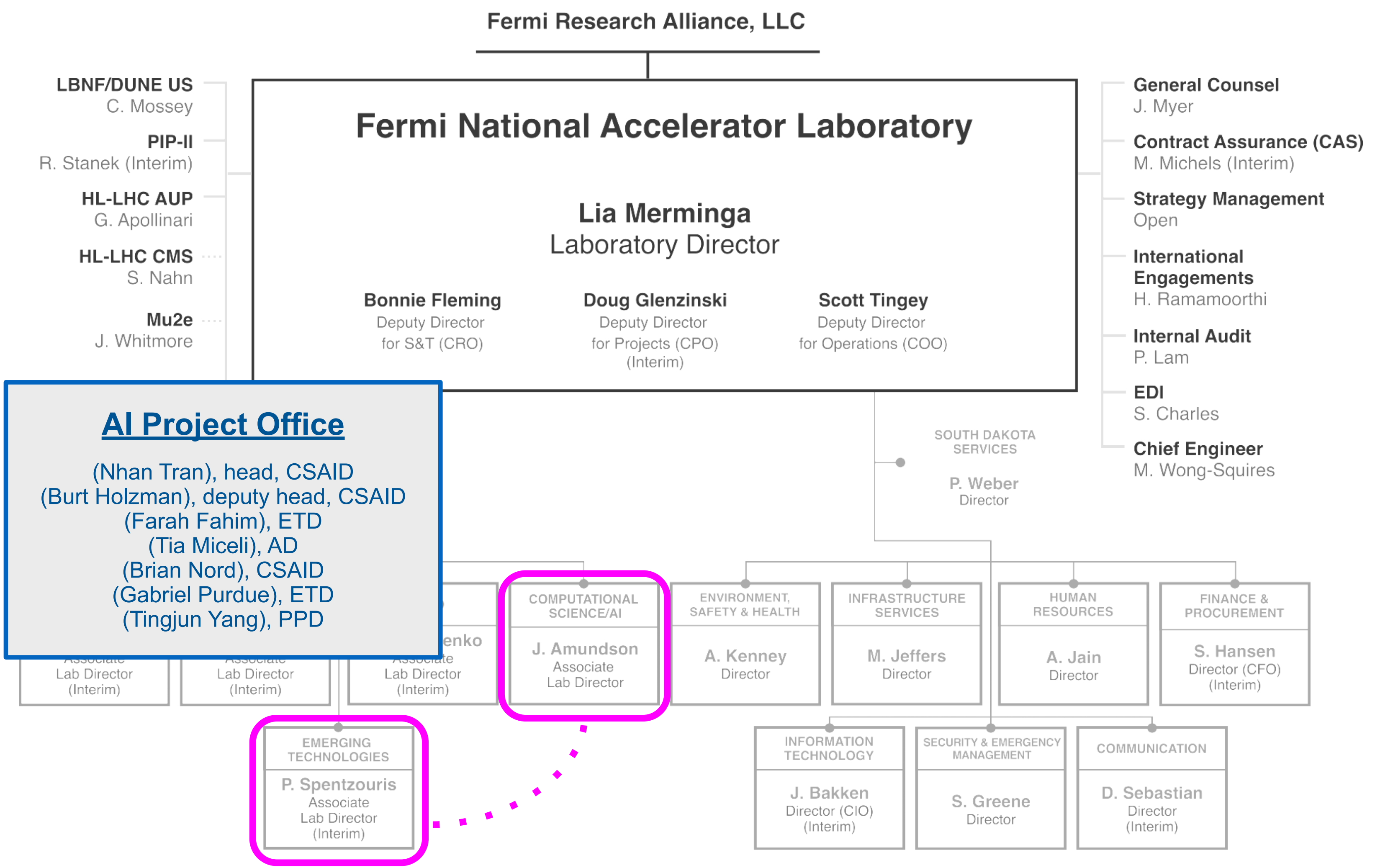
- **Fermilab AI/ML program focused on *accelerating science***
  - Program pillars connect algorithm advancements with sensing, computing, and operations to solve HEP challenges
  - Identified areas where Fermilab contributes to the greater DOE AI needs
- AI Project Office coordinating overall **strategy and building community**
- Portfolio of research strong case for AI center involvement
  - Center lead would focus on real-time AI and edge sensing
    - Additional focus areas could complement other centers (digital twins, automated discovery and design)
  - Modest funds needed to seed efforts during **upcoming critical 1 year period**
    - Opportunities to develop collaborations & projects focused on *core AI research, strategic HEP applications, and industry/academic partnerships*

Outline

- Vision & strategic drivers
- **AI Project Office and program organization**
- Program milestones and highlights
- Leveraging unique & core capabilities

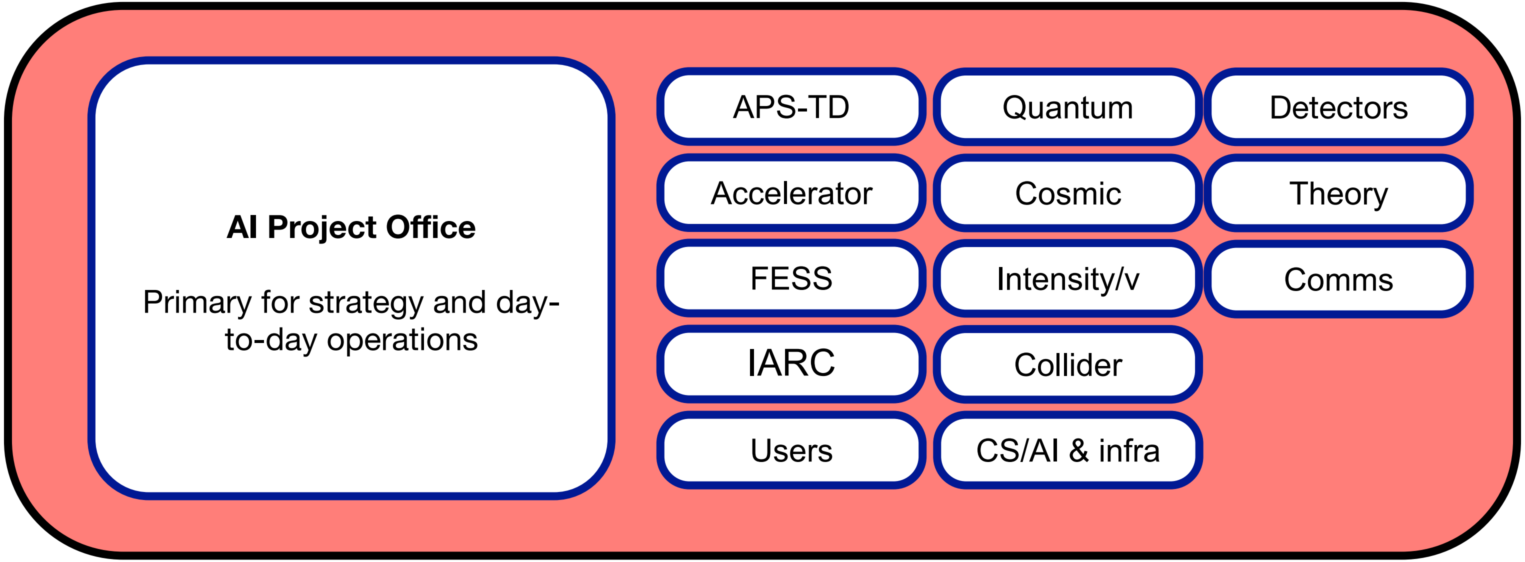


Fermi Research Alliance, LLC
Fermi National Accelerator Laboratory
Lia Merminga
Laboratory Director
Bonnie Fleming
Deputy Director for S&T (CRO)
Doug Glenzinski
Deputy Director for Projects (CPO) (Interim)
Scott Tingey
Deputy Director for Operations (COO)
LBNF/DUNE US
C. Mossey
PIP-II
R. Stanek (Interim)
HL-LHC AUP
G. Apollinari
HL-LHC CMS
S. Nahn
Mu2e
J. Whitmore
SQMS
A. Grassellino
General Counsel
J. Myer
Contract Assurance (CAS)
M. Michels (Interim)
Strategy Management
Open
International Engagements
H. Ramamoorthi
Internal Audit
P. Lam
EDI
S. Charles
Chief Engineer
M. Wong-Squires
USER ENGAGEMENT
SOUTH DAKOTA SERVICES
P. Weber
Director
PARTICLE PHYSICS
K. Burkett
Associate Lab Director (Interim)
ACCELERATORS
A. Valishev
Associate Lab Director (Interim)
APSTD
A. Romanenko
Associate Lab Director (Interim)
COMPUTATIONAL SCIENCE/AI
J. Amundson
Associate Lab Director
EMERGING TECHNOLOGIES
P. Spentzouris
Associate Lab Director (Interim)
ENVIRONMENT, SAFETY & HEALTH
A. Kenney
Director
INFRASTRUCTURE SERVICES
M. Jeffers
Director
HUMAN RESOURCES
A. Jain
Director
FINANCE & PROCUREMENT
S. Hansen
Director (CFO) (Interim)
INFORMATION TECHNOLOGY
J. Bakken
Director (CIO) (Interim)
SECURITY & EMERGENCY MANAGEMENT
S. Greene
Director
COMMUNICATION
D. Sebastian
Director (Interim)



Fermi Research Alliance, LLC
Fermi National Accelerator Laboratory
Lia Merminga
Laboratory Director
Bonnie Fleming
Deputy Director
for S&T (CRO)
Doug Glenzinski
Deputy Director
for Projects (CPO)
(Interim)
Scott Tingey
Deputy Director
for Operations (COO)
LBNF/DUNE US
C. Mossey
PIP-II
R. Stanek (Interim)
HL-LHC AUP
G. Apollinari
HL-LHC CMS
S. Nahn
Mu2e
J. Whitmore
General Counsel
J. Myer
Contract Assurance (CAS)
M. Michels (Interim)
Strategy Management
Open
International Engagements
H. Ramamoorthi
Internal Audit
P. Lam
EDI
S. Charles
Chief Engineer
M. Wong-Squires
SOUTH DAKOTA SERVICES
P. Weber
Director
AI Project Office
(Nhan Tran), head, CSAID
(Burt Holzman), deputy head, CSAID
(Farah Fahim), ETD
(Tia Miceli), AD
(Brian Nord), CSAID
(Gabriel Purdue), ETD
(Tingjun Yang), PPD
Lab Director
(Interim)
Lab Director
(Interim)
enko
Lab Director
(Interim)
COMPUTATIONAL SCIENCE/AI
J. Amundson
Associate
Lab Director
ENVIRONMENT, SAFETY & HEALTH
A. Kenney
Director
INFRASTRUCTURE SERVICES
M. Jeffers
Director
HUMAN RESOURCES
A. Jain
Director
FINANCE & PROCUREMENT
S. Hansen
Director (CFO)
(Interim)
EMERGING TECHNOLOGIES
P. Spentzouris
Associate
Lab Director
(Interim)
INFORMATION TECHNOLOGY
J. Bakken
Director (CIO)
(Interim)
SECURITY & EMERGENCY MANAGEMENT
S. Greene
Director
COMMUNICATION
D. Sebastian
Director
(Interim)

AI Program and Liaisons



Liaisons: link across the laboratory

communicate interests and needs of focus area to AI project and focus area participants

providing input to overall AI project strategy

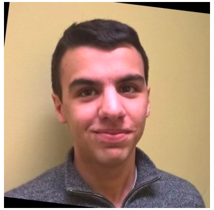
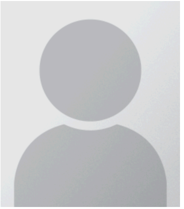
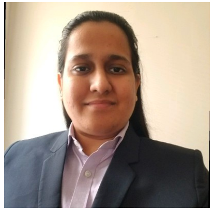
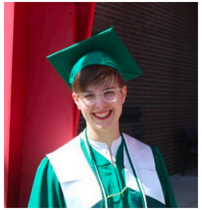
organize materials, inputs for AI-related funding calls and communications.

Mission

- Developing **strategic capabilities** within the (inter)national AI ecosystem
  – AI to advance lab scientific mission, and where Fermilab can advance AI research
- Building **community** around cross-cutting problems, tools, and educational opportunities
  – Connecting teams across the lab and keeping a big-picture view of what is going on
  – Develop infrastructure for AI research — both people (e.g. AI associate program) and hardware (e.g. GPU access)
- Establish a strategy to support a **strong funding profile** through network of stakeholders and partners
- **Sharing** Fermilab and HEP's AI work with the world

Workforce development

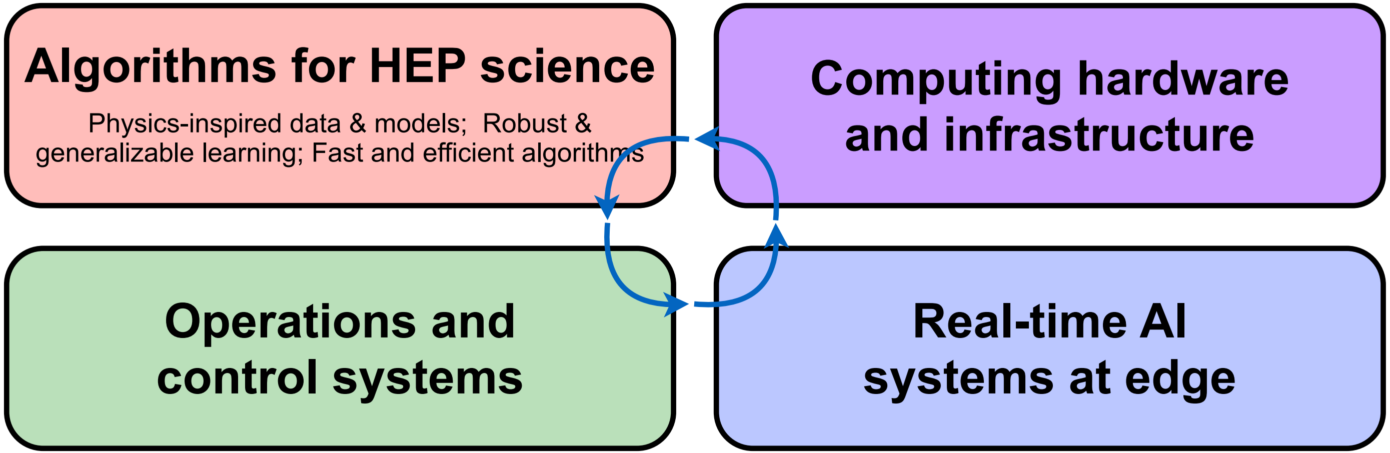
- New job type developed for AI research: **AI associate program**
  - New job family for advancement at Fermilab
- Modeled after industry 1-year internships
- Provides scientific AI research opportunities
  - Primarily Bachelors/MS with background in computer science & AI
- Concept emulated in other areas - e.g. engineering, quantum

Outline

- Vision & strategic drivers
- AI Project Office and program organization
- **Program milestones and highlights**
- Leveraging unique & core capabilities

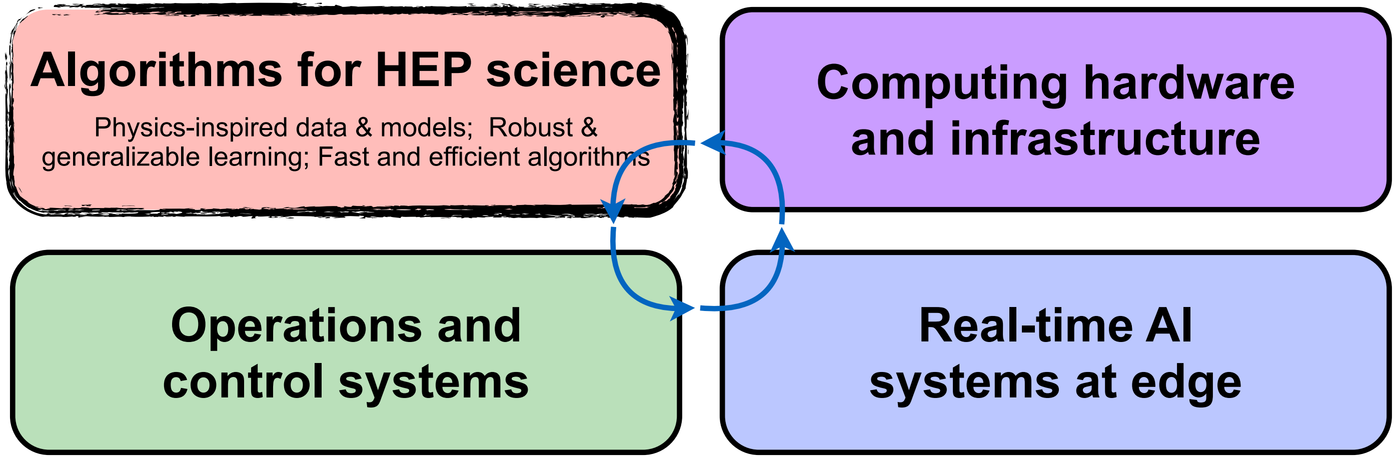
Program context



AI program in ~15 minutes

- Algorithms for HEP science
  - **Physics-inspired models and data**
    - Graph learning
    - Generative models
    - SBI/likelihood-free inference
    - Accelerating theory
  - **Robust and generalizable learning**
    - Domain adaptation
    - Anomaly detection
    - Semi-/self-supervision
  - **Fast and efficient algorithms**
    - Multi-objective optimization
    - Quantization/sparsity
    - Knowledge distillation
- Operations and controls
  - Real-time accelerator controls
  - Telescope design and operations
  - Quantum machine learning
- Computing hardware and infrastructure
  - Resources for AI practitioners
  - Efficient AI-in-production
- Real-time systems at the edge
  - Hardware-algorithm codesign for HEP and beyond
  - Near-detector, low latency AI
  - On-sensor/detector AI

Program context

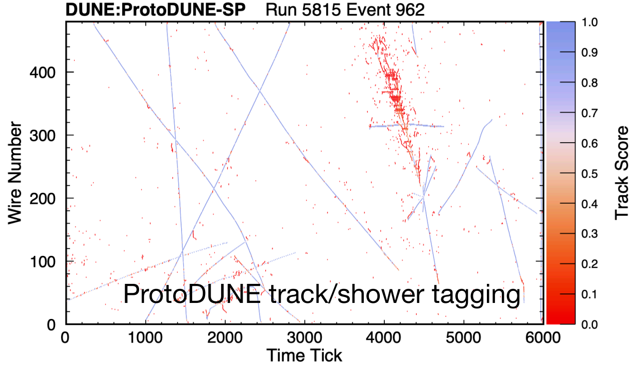


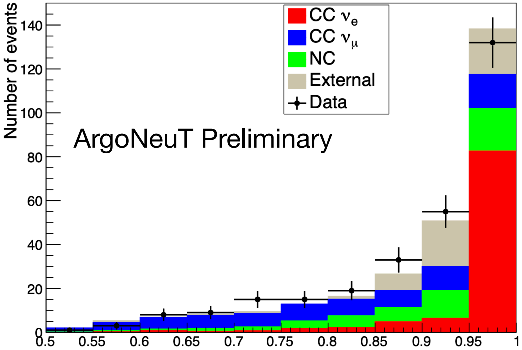
Reconstruction and pattern recognition

Uboldi et al, Nucl. Instrum. Meth. A 1028 (2022) 166371
ArgoNeuT JINST 17 (2022) P01018
DUNE Eur.Phys.J.C 82 (2022) 10, 903

Convolutional NNs to provide crucial information in neutrino interactions

- **Waveform ROI identification**
  - 1D CNN to identify signals in the raw waveforms.
  - Works for both TPC and photon detector waveforms.
- **Hit tagging**
  - 2D CNN to flag each hit as track, shower or Michel activity.
  - Validated using ProtoDUNE data.
- **Neutrino ID**
  - 2D CNN to flag each neutrino interaction as numu, nue or NC interaction.
  - Developed for DUNE and validated using ArgoNeuT data.
- **MicroBooNE open data!**
  - A tool for collaborative AI developments
  - https://microboone.fnal.gov/documents-publications/public-datasets/



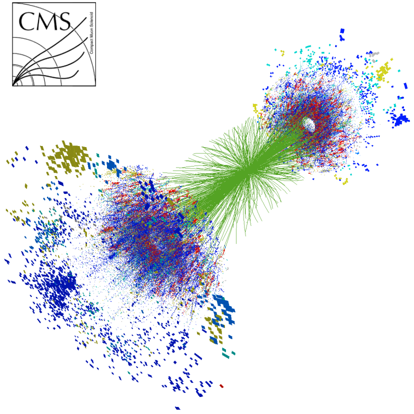


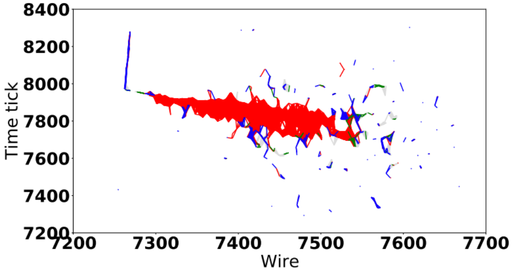
Reconstruction and pattern recognition

CMS-DP-2018-046
Gray, Kljinsma, et al., arXiv:2003.08013
Cerati, Kowalkowski, Gray, Klijnsma, et al., https://arxiv.org/abs/2103.06233

Including Graph Neural Networks to extract optimal performance from **complex, high-dimensional, sparse data**

- **Broad applications across HEP**
- **LHC Jet tagging** natural application for Graph NNs
  - Boosted Higgs (→ bb) gives 2x more signal efficiency
  - Enables new analyses → ggHcc!
- **Graphs for clustering & tracking**
  - CMS HGCal (High Granularity Calorimeter) clustering - leading performance for multi-particle reconstruction
  - ECal clustering application for Run 2/3 targeting - improves γγ significance by ~7%
  - Other applications include MET, pileup mitigation, etc
  - Exploration for LArTPC reconstruction for tracking + clustering

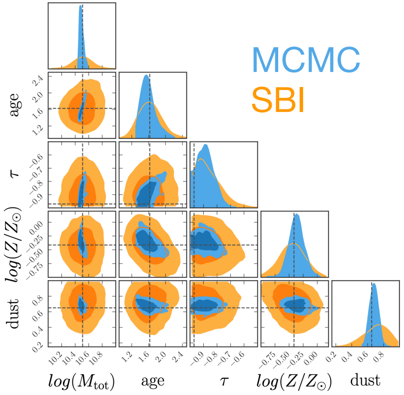




Simulation-based Inference (SBI) for Cosmic Analysis

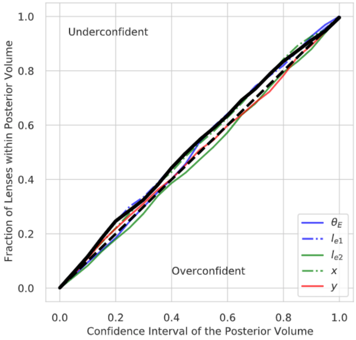
Nord ECA
Galaxy Spectra
Khullar, Nord, Ciprijanovic, Poh, Xu 2022 (MLST & Neurips)
Strong Lenses
Poh et a.l, 2022 in Neurips Workshop

- **Goal:** Maximal information extraction from high-dimensional data to **rapidly find/measure** objects, dark energy, dark matter
  - **Traditional methods use explicit analytic functions** with simplified assumptions; typically **slow** and **inaccurate**
- **Forward modeling and SBI** permits flexible likelihoods
  - Simulated datasets until matching observation
  - Can be 10^5 times faster than traditional methods
- Applications across many surveys (DES, LSST, CMB-S4) and objects (Strong Lenses, Spectra, Quasars, Galaxy Clusters)
  - Connections across all of HEP



Proof-of-concept:
Simple SBI method (not highly tuned) is just as accurate as MCMC, but much faster

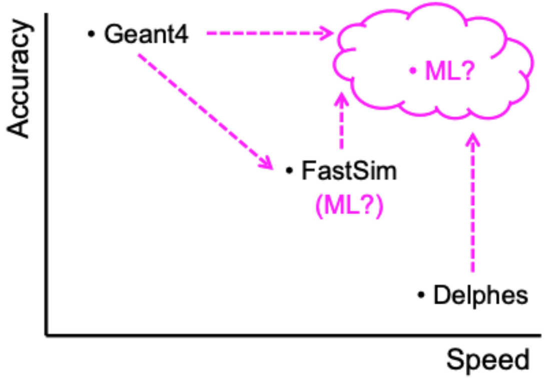
SBI shows correct level of confidence in estimates.



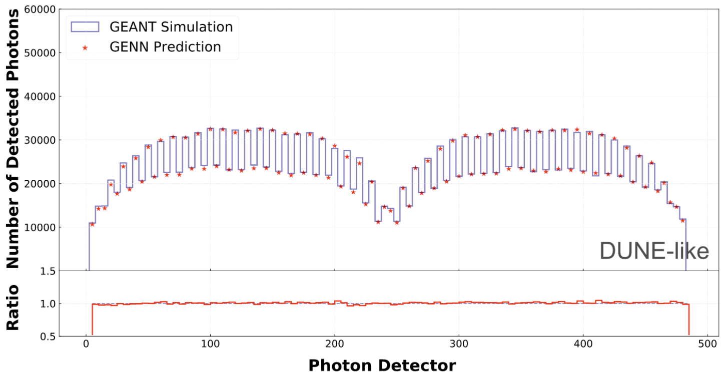
Generative models for simulation

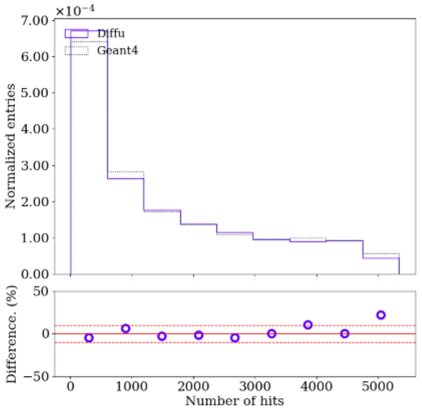
Pedro et al., arXiv:2202.05320, ACAT2021
Pedro et al., arXiv:2203.08806
Mu, Himmel, Ramson, Mach. Learn. Sci. Tech. 3 (2022) 1, 015033

- **High fidelity ML-based parameterized simulation** to mitigate computing bottleneck for DUNE and LHC
  - Find way to fuse GEANT full-sim with ML
  - More naturally run on coprocessors
- GENN for photon transport simulation
- Stable diffusion (CaloDiffusion) for LHC calorimeter



20-50 times faster than Geant4 simulation



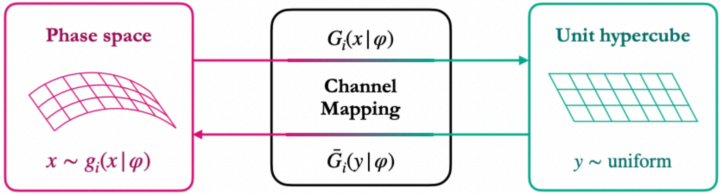


Diffusion model: avoids pitfalls of GANs, high quality output

Competitive results on the CaloChallenge dataset

Accelerating theory

Rocco et al., arXiv: 2206.10021
Issacson et al., arXIv:2212.06172



Develop **flexible hidden-nucleon, neural network ansatz suitable to solve the nuclear many-body Schrodinger equation**

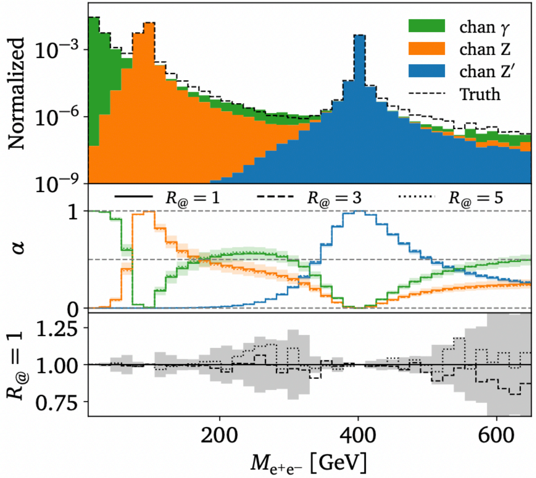
Non-exponential scaling with number of nucleons

Light and medium-mass nuclei's energy and spatial density distributions in excellent agreement with theory calculations

Neural importance sampling with normalizing flows:

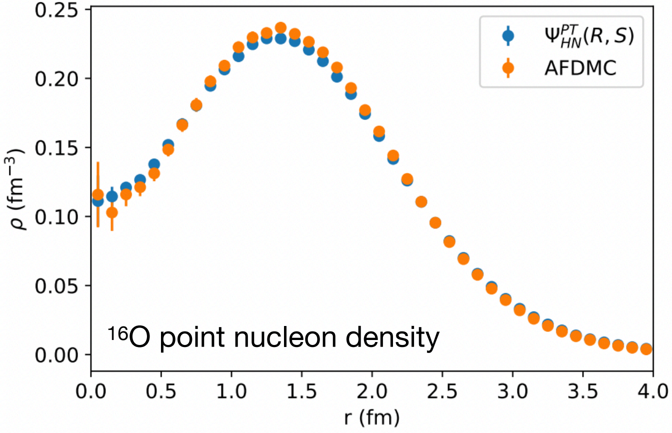
Models a complex probability density as an invertible transformation of simple base density.

Machine-learned multi-channel Drell-Yan



Hidden nucleon ansatz (AI)

Perturbatively-corrected theory calculation



Domain Adaptation for dataset shift

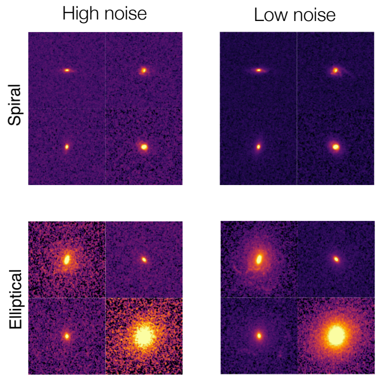
Deep Universal Domain Adaptation
Ciprijanovic, Lewis, Pedro, Madireddy, Nord, Perdue, Wild
(2022 in Neurips Workshop, 2023 in prep for journal)
CMS, Stealth SUSY search arXiv:2102.06976

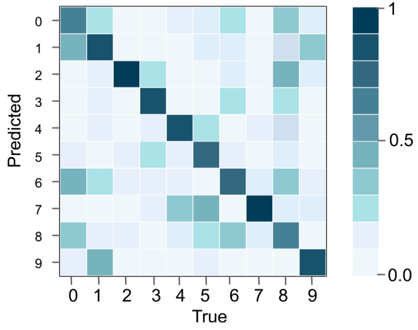
- **Adapting AI models** as data changes — different datasets, simulation vs. observation, etc.
  - **Mitigate bias** from training sample
- **Deep Universal Domain Adaptation (DUDA) for cosmic analysis**

  A. reduces the need hyperparameter tuning and

  B. reduces the requirement for overlap between training and observed data
- Applications across many surveys (DES, LSST, CMB-S4) and objects (Strong Lenses, Spectra, Quasars, Galaxy Clusters) — connections across all of HEP
  - Unsupervised domain adaptation from gradient reversal is used for data-driven in LHC analysis for Stealth SUSY background estimation

Example images of simulated galaxy morphologies with different levels of telescope noise.





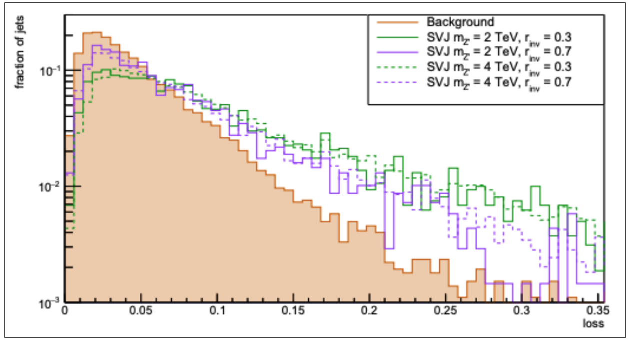
Confusion matrix for classification of galaxy types using DUDA

Robust learning from data

Pedro et al., JHEP 02 (2022) 074
Ngadiuba et al., arXiv: 2107.02157
Ngadiuba et al., Nature Machine Intelligence 4, 154 (2022)
Ngadiuba et al., arXiv: 2110.08508
Feng, Tran et al., submitted to EPJC

Deeper insights with less reliance on simulation

- **Anomaly detection**
  - **At the LHC**
    - Studied on dark QCD showers - autoencoder trained only on bag-only exceeds performance of BDT trained on a "wrong" signal model
    - In L1 trigger - enable sensitivity to hidden or suppressed new physics scenarios (leptoquarks, new scalars, …)
  - **For accelerator controls**, L-CAPE project using 2022 Linac Data
    - LSTM autoencoder to identify faults — higher operational efficiency
    - Most common Linac faults being identified, and some with actionable precursors
- **Semi-supervised models**
    - Semi-supervised graph learning for PU mitigation reduces reliance on simulation (modeling, truth info) - trains on charged particles in data
    - Improves on expert algo by > 20% for jet mass resolution

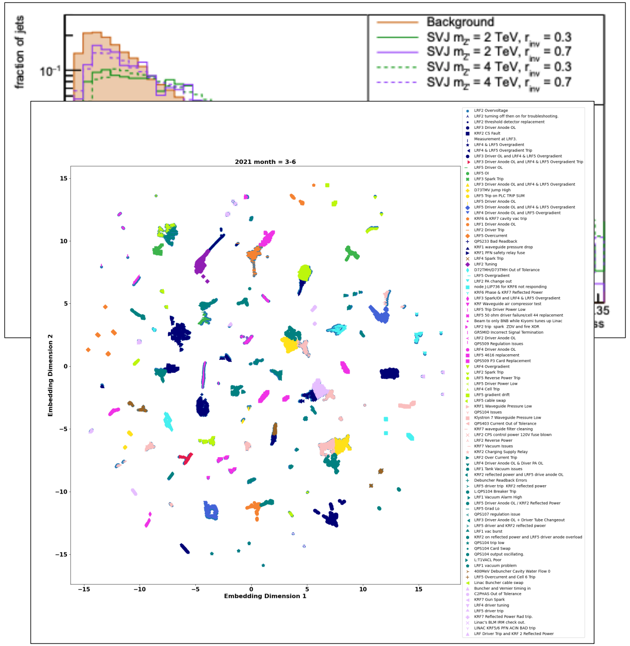


Robust learning from data

Pedro et al., JHEP 02 (2022) 074
Ngadiuba et al., arXiv: 2107.02157
Ngadiuba et al., Nature Machine Intelligence 4, 154 (2022)
Ngadiuba et al., arXiv: 2110.08508
Feng, Tran et al., submitted to EPJC

Deeper insights with less reliance on simulation

- **Anomaly detection**
  - **At the LHC**
    - Studied on dark QCD showers - autoencoder trained only on bag-only exceeds performance of BDT trained on a "wrong" signal model
    - In L1 trigger - enable sensitivity to hidden or suppressed new physics scenarios (leptoquarks, new scalars, …)
  - **For accelerator controls**, L-CAPE project using 2022 Linac Data
    - LSTM autoencoder to identify faults — higher operational efficiency
    - Most common Linac faults being identified, and some with actionable precursors
- **Semi-supervised models**
    - Semi-supervised graph learning for PU mitigation reduces reliance on simulation (modeling, truth info) - trains on charged particles in data
    - Improves on expert algo by > 20% for jet mass resolution

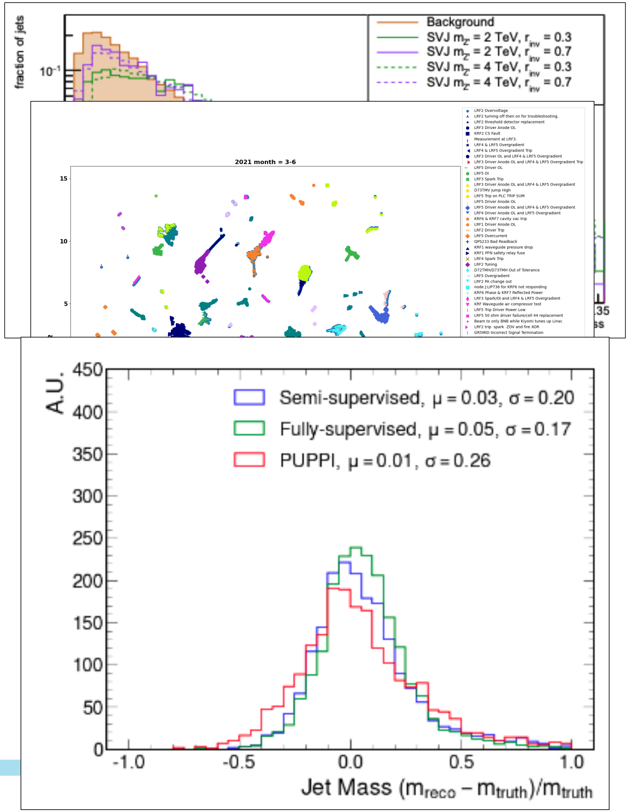


Robust learning from data

Pedro et al., JHEP 02 (2022) 074
Ngadiuba et al., arXiv: 2107.02157
Ngadiuba et al., Nature Machine Intelligence 4, 154 (2022)
Ngadiuba et al., arXiv: 2110.08508
Feng, Tran et al., submitted to EPJC

Deeper insights with less reliance on simulation

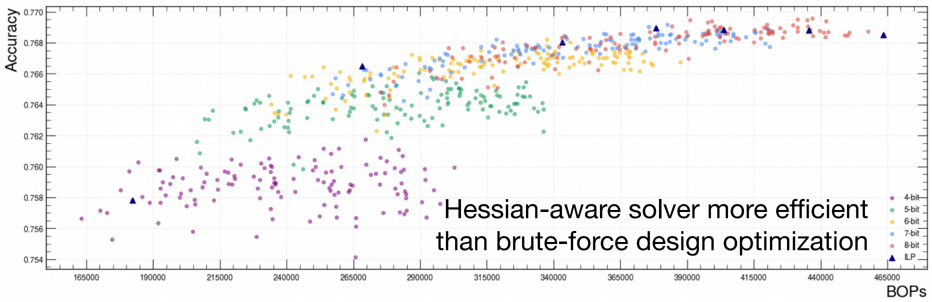
- **Anomaly detection**
  - **At the LHC**
    - Studied on dark QCD showers - autoencoder trained only on bag-only exceeds performance of BDT trained on a "wrong" signal model
    - In L1 trigger - enable sensitivity to hidden or suppressed new physics scenarios (leptoquarks, new scalars, …)
  - **For accelerator controls**, L-CAPE project using 2022 Linac Data
    - LSTM autoencoder to identify faults — higher operational efficiency
    - Most common Linac faults being identified, and some with actionable precursors
- **Semi-supervised models**
  - Semi-supervised graph learning for PU mitigation reduces reliance on simulation (modeling, truth info) - trains on charged particles in data
  - Improves on expert algo by > 20% for jet mass resolution



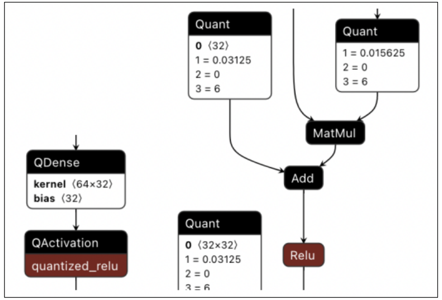
Fast and efficient algorithms

Hawks, Tran, Quantization-aware pruning, arXiv:2102.11289
Mitrevski, Hawks, Muhizi, Tran, QONNX, arXiv:2206.07527
An end-to-end codesign workflow of Hessian-aware quantized neural networks for FPGAs and ASICs
Campos, Hawks, MItrevski, Tran
Quantized Distilled Autoencoder Model for 4D Transmission Edge Microscopy
Forelli, Muhizi, Tran

- **Real-time and efficient AI**: driver for scientific sensing/compute
  - Core research into quantization and sparsity and optimization techniques
  - Important for hardware implementation (more on this later)
    - Developing training frameworks for quantization-aware AI and hardware translation
    - QONNX - build industry standards - interchange formats for quantized AI
- Building techniques for broader scientific community
  - Quantized model distillation for microscopy



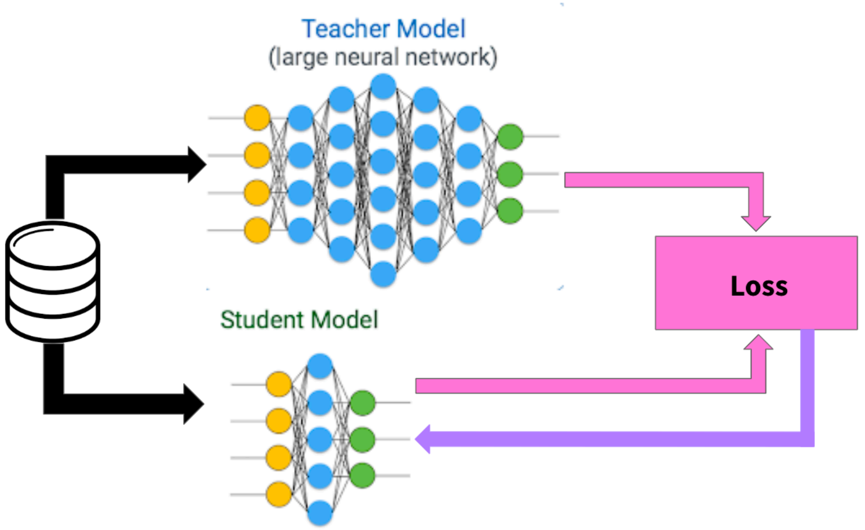
Collaboration with industry/community on common standards for representing quantized neural networks



Fast and efficient algorithms

Hawks, Tran, Quantization-aware pruning, arXiv:2102.11289
Mitrevski, Hawks, Muhizi, Tran, QONNX, arXiv:2206.07527
An end-to-end codesign workflow of Hessian-aware quantized neural networks for FPGAs and ASICs
Campos, Hawks, MItrevski, Tran
Quantized Distilled Autoencoder Model for 4D Transmission Edge Microscopy
Forelli, Muhizi, Tran

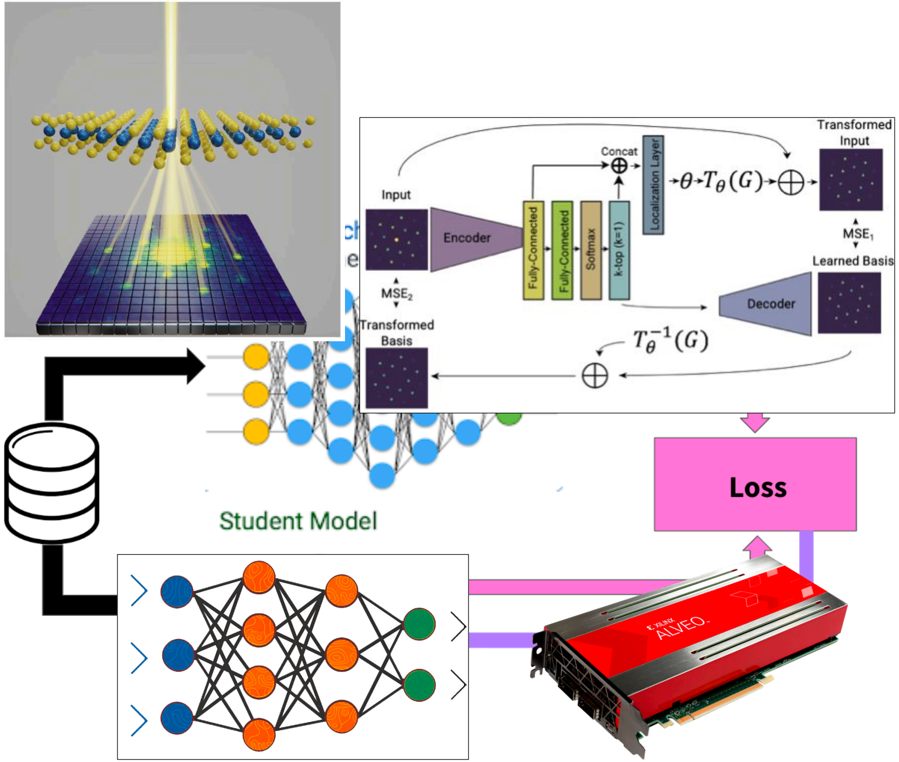
- **Real-time and efficient AI**: driver for scientific sensing/compute
  - Core research into quantization and sparsity and optimization techniques
  - Important for hardware implementation (more on this later)
    - Developing training frameworks for quantization-aware AI and hardware translation
    - QONNX - build industry standards - interchange formats for quantized AI
- Building techniques for broader scientific community
  - Quantized model distillation for microscopy



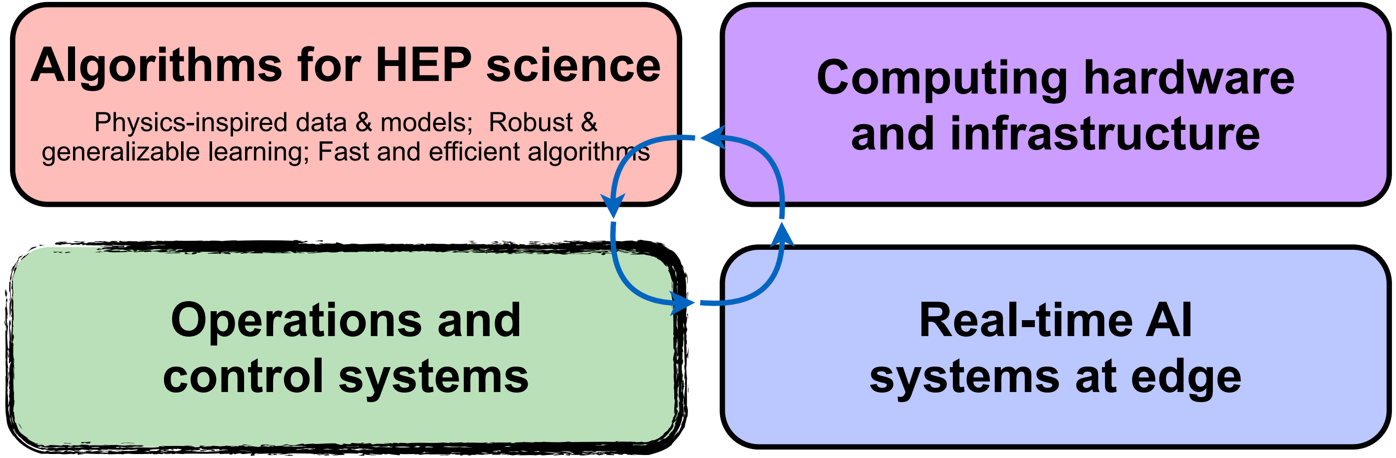
Fast and efficient algorithms

Hawks, Tran, Quantization-aware pruning, arXiv:2102.11289
Mitrevski, Hawks, Muhizi, Tran, QONNX, arXiv:2206.07527
An end-to-end codesign workflow of Hessian-aware quantized neural networks for FPGAs and ASICs
Campos, Hawks, MItrevski, Tran
Quantized Distilled Autoencoder Model for 4D Transmission Edge Microscopy
Forelli, Muhizi, Tran

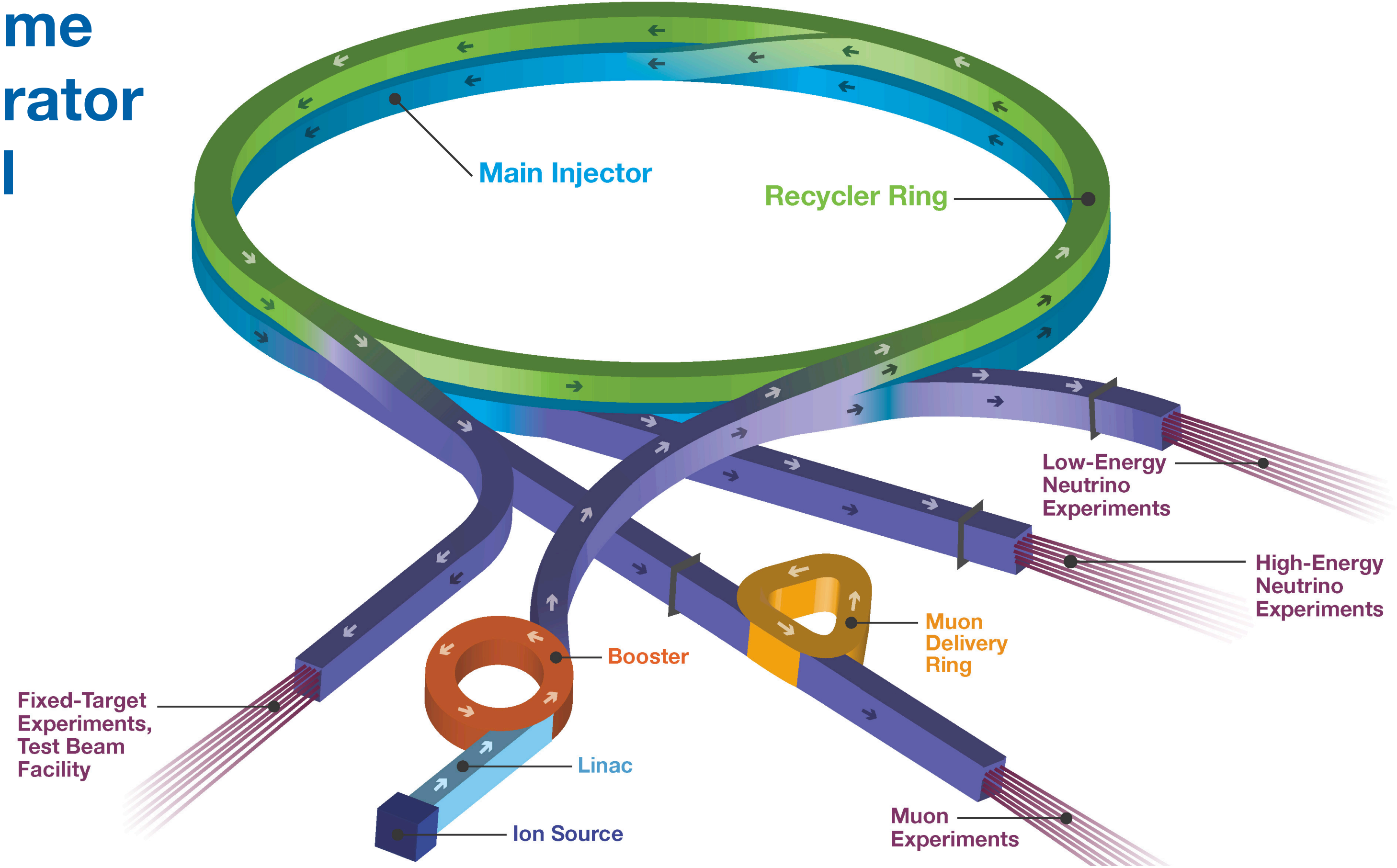
- **Real-time and efficient AI**: driver for scientific sensing/compute
  - Core research into quantization and sparsity and optimization techniques
  - Important for hardware implementation (more on this later)
    - Developing training frameworks for quantization-aware AI and hardware translation
    - QONNX - build industry standards - interchange formats for quantized AI
- Building techniques for broader scientific community
    - Quantized model distillation for microscopy



Program context



Real-time accelerator control

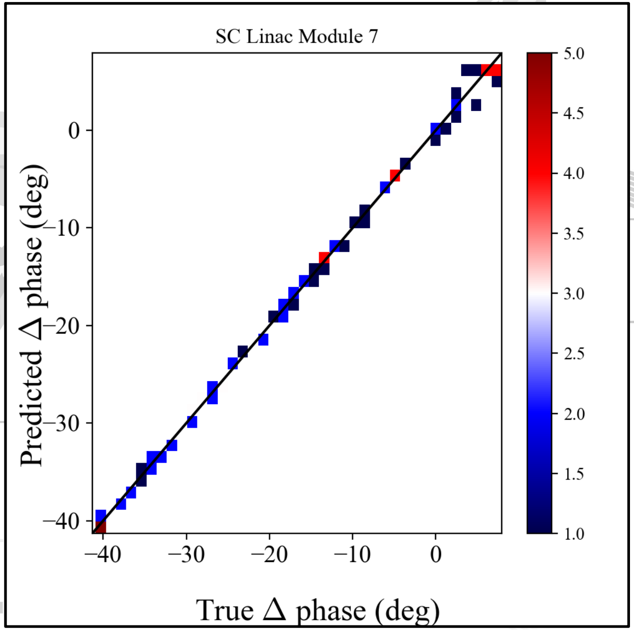


Real-time accelerator control

Linac RF optimization

Predict RF parameters to keep beam energy constant and minimize emmitance

Proof-of-concept with single cavity phase regulation; multi-cavity promising

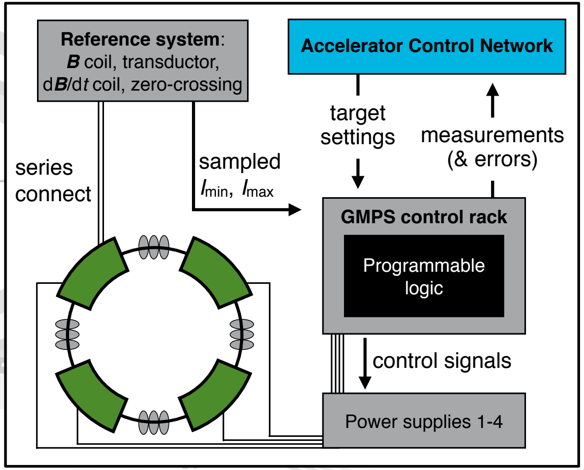


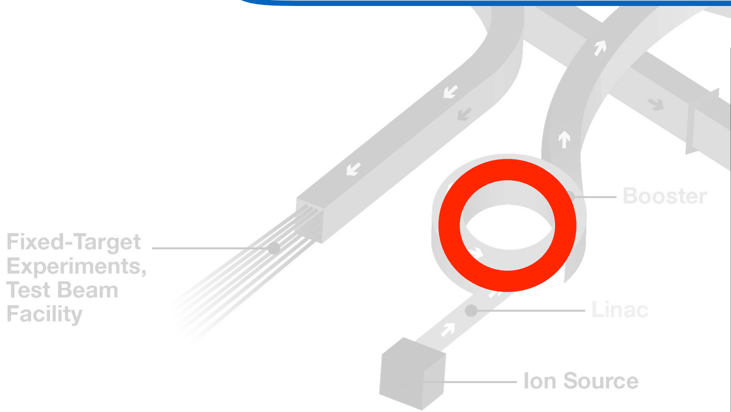
Real-time accelerator control

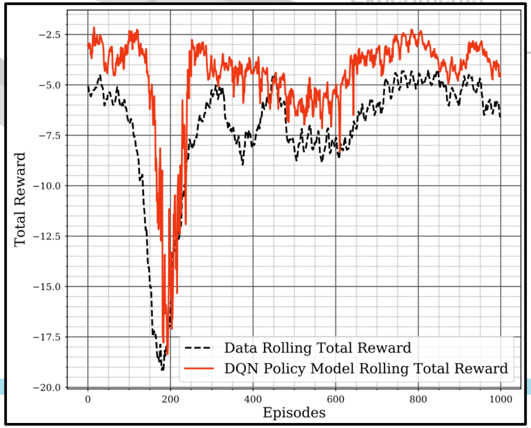
Booster GMPS

Real-time **reinforcement learning agent** in FPGA to regulate Gradient Magnet Power Supply; replace a traditional PID loop — shows improvement in reward (reduced magnet current error)

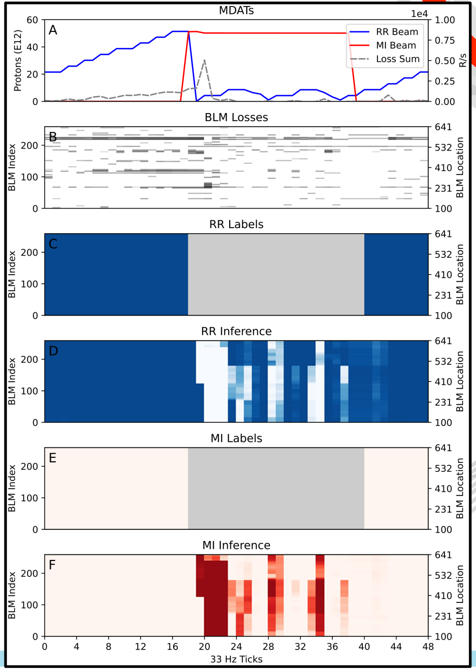
Development of digital twin for simulation framework

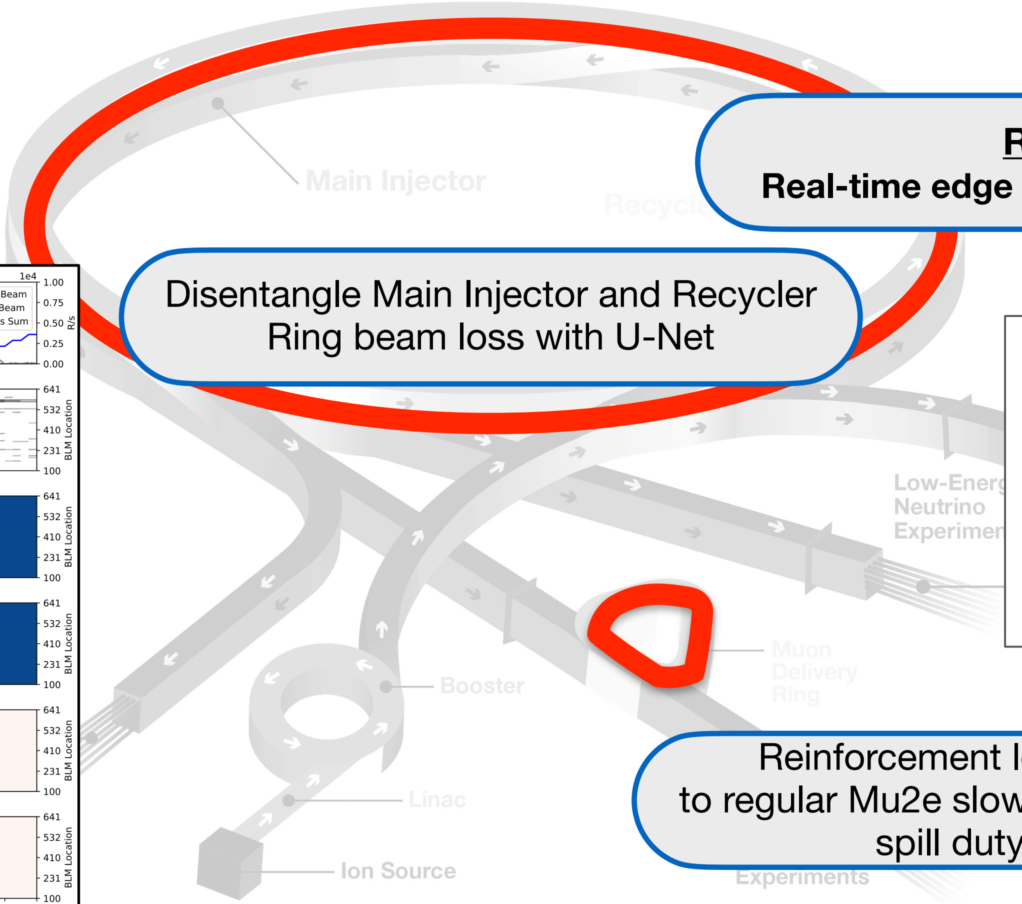






Real-time accelerator control





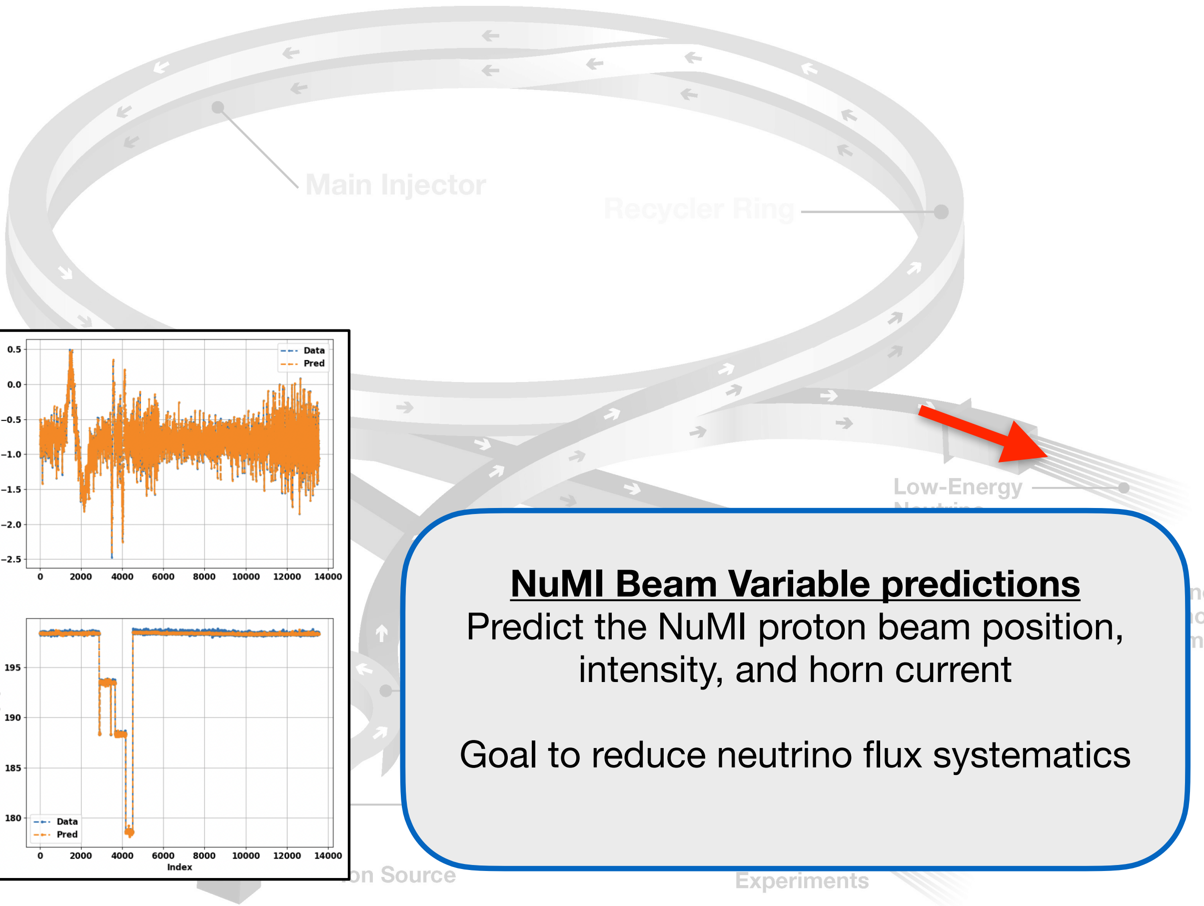
READS

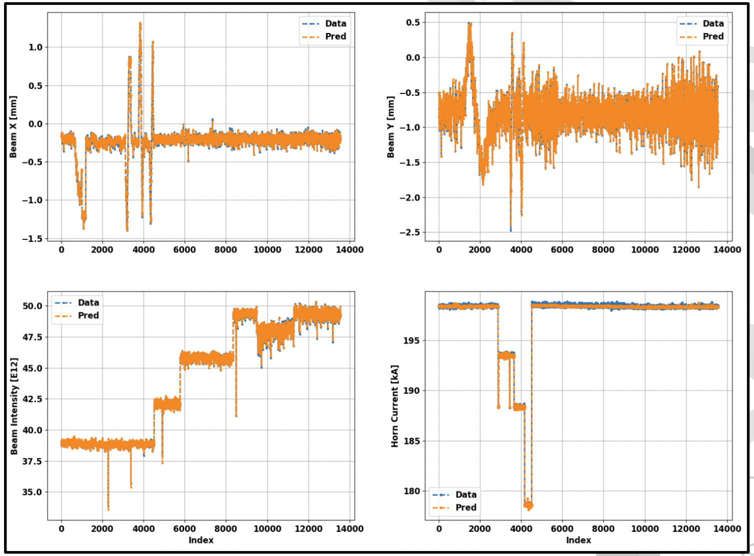
Real-time edge AI distributed system

Disentangle Main Injector and Recycler Ring beam loss with U-Net

Reinforcement learning agent to regular Mu2e slow spill and increase spill duty factor

Real-time accelerator control





NuMI Beam Variable predictions

Predict the NuMI proton beam position, intensity, and horn current

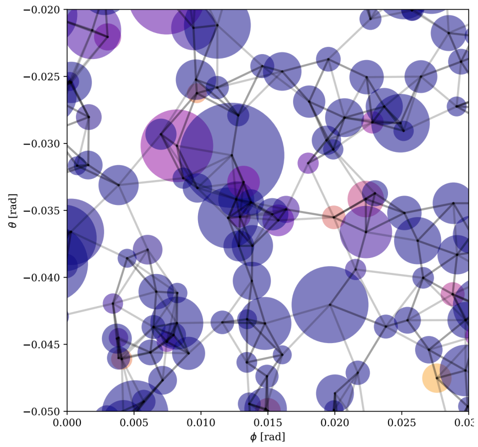
Goal to reduce neutrino flux systematics

Automation for cosmology experiments

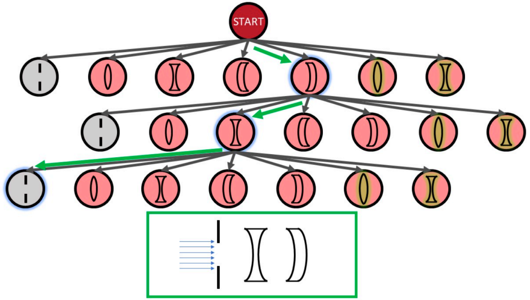
Spectroscopic Survey Optimization
Cranmer, Melchior, Nord, 2021 (Neurips workshop)
Optical System Design
Cohen (HS student) and Nord, 2023 (in prep.)

Self-driving telescopes:
Adaptive optimization for survey scheduling

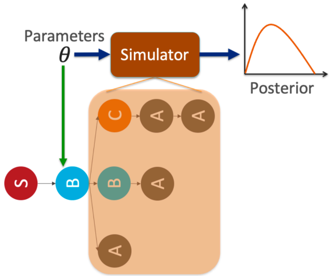
- **Unsupervised Graph Neural Networks**: optimize an observation strategy to constrain cosmological parameters
- **Supervised Reinforcement Learning**: build a decision-making algorithm to prepare or adapt observations



A network of galaxies optimally selected for cosmic matter estimation



Schematic example of generating an optical system - Green arrows show optimized tree traversal



Overview: tree produces optical system; posteriors are of element shape parameters

Automated instrument design: replace expensive optics simulation

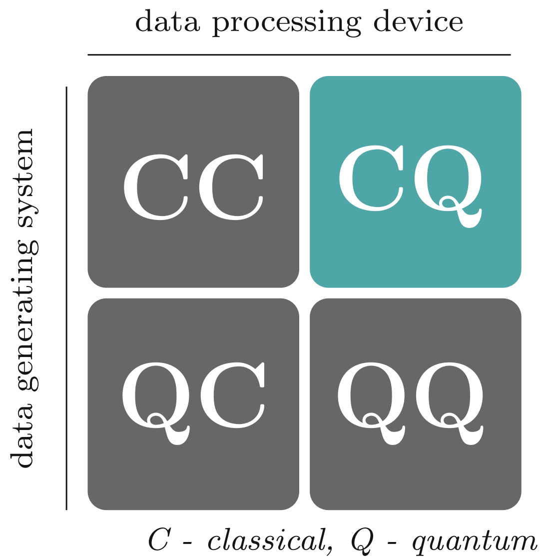
Use decisions trees + simulation-based inference to *arrange optics* and *choose optical element*

Quantum machine learning

Science mission: explore application of "quantum machine learning" to *scientific data*

"Quantum ML" means different things depending on what the data source and algorithm's physical substrate are

- Anecdotally, most experimental and theoretical work in QML focuses on using a quantum processor to analyze classical data ("CQ").
  - Also almost certainly the least promising area for study, especially for HEP (large datasets, science motivations not well-aligned).
- Analyzing quantum data on a classical machine ("QC") usually becomes a control problem, or a program optimization problem.
- Analyzing quantum data with a quantum processor ("QQ") makes the most sense in the context of analyzing the output of quantum sensors or the output of another quantum computer - we can't store entangled states for long periods of time!

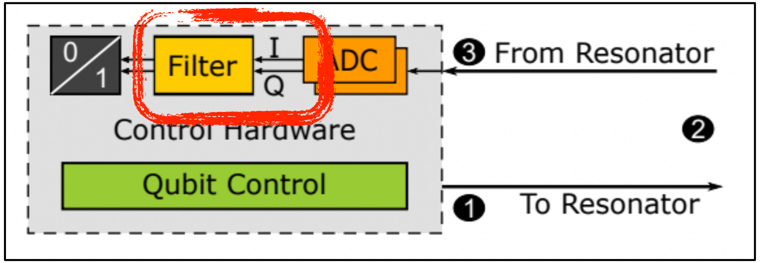


M. Schuld and F. Petruccione
Springer Press, 2018

QC and QQ — interesting areas to explore at Fermilab!

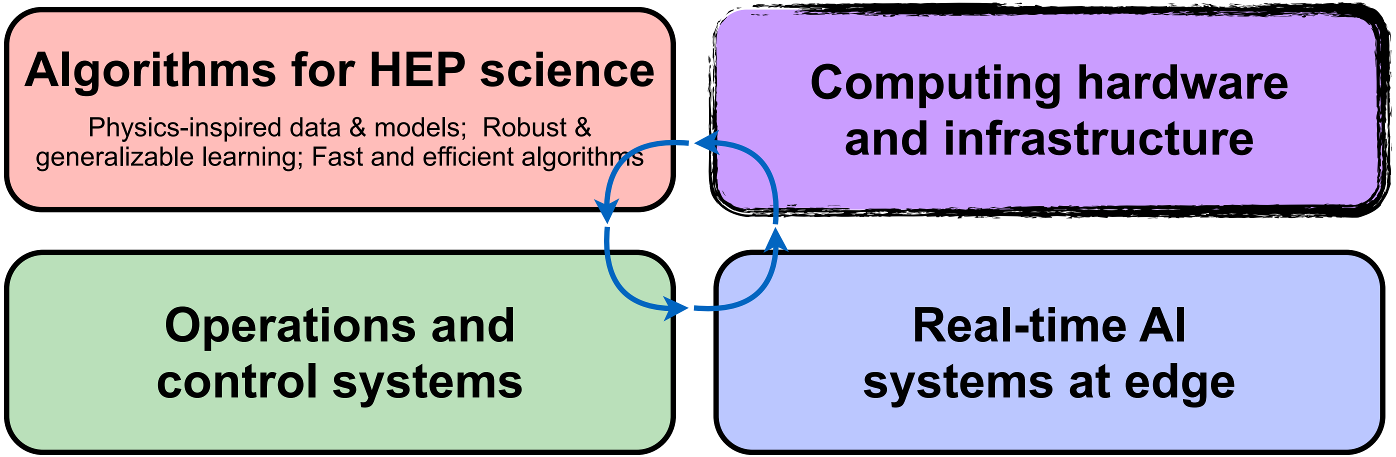
Practical QML (QC and QQ) at Fermilab

- AI/ML for controlling and optimizing quantum computers
  - Exciting effort couples to microelectronics and edge AI applications to improve quantum readout
  - Classical AI for de-noising quantum computations in theory calculations and event generators — QuantISED program studying quantum computing for neutrino scattering calculations
  - Classical AI for predicting quantum circuit fidelity on noisy hardware - important for HEP field theory problems involving extremely deep quantum circuits



- Quantum AI for quantum data
  - Exciting efforts involve theoretical work on enhancing the sensitivity of quantum sensors connected by a quantum network (SQMS and FQI).
    - Very early days although proof of principle theoretical and experimental work has been done on optical test benches.
  - Quantum ML techniques for enhancing signal extraction from quantum simulation (FQI, joint with U. Trento, CERN).
    - No clear advantages discovered yet - may be a hammer searching for nails, but potentially interesting.

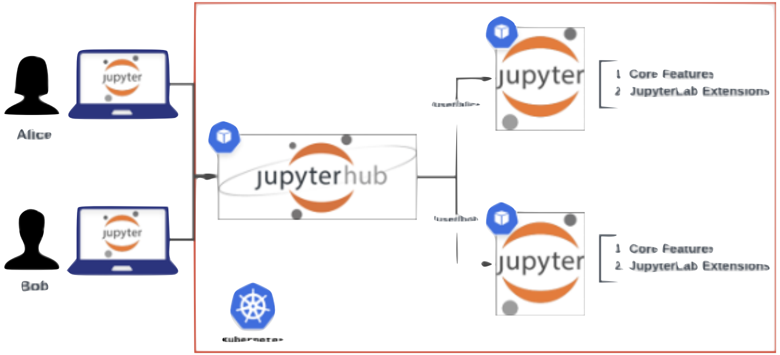
Program context

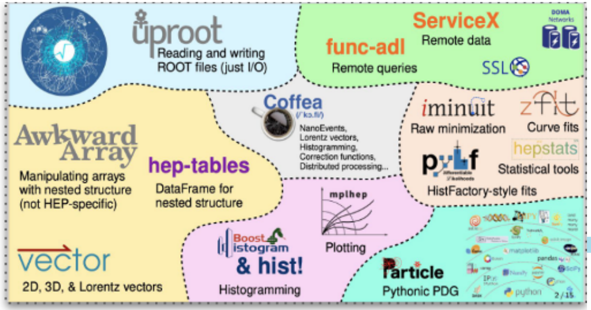


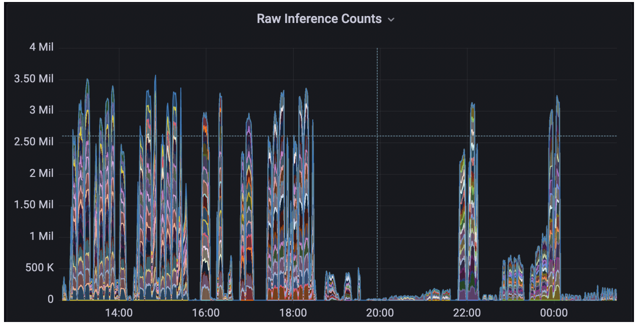
Elastic Analysis Facility & Fermilab Computing Facilities

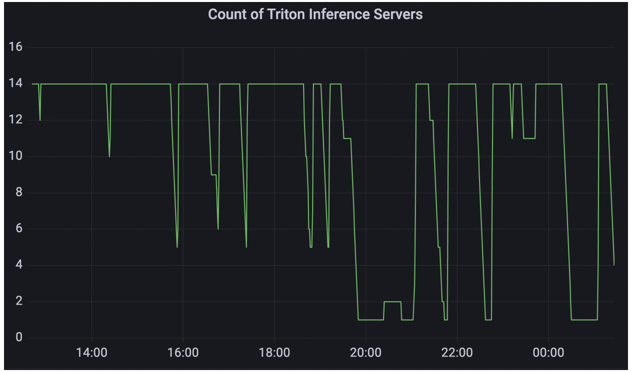
- **Elastic Analysis Facility** @ Fermilab provides **resources** and **data-science standard industry tools** for AI training and inference
- Additional GPU resources available on CMS LPC, Wilson Cluster
- Capable of **bursting** to O(100k) batch computing CPU cores

Flechas et al., arXiv:2203.10161
Benjamin et al., arXiv:2203.08010





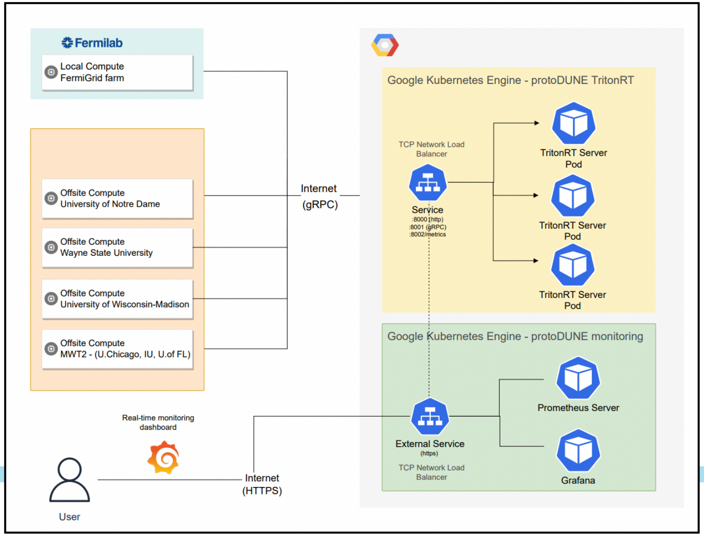




Accelerating ML processing

Jindariani, Ngadiuba, Pedro, Tran, [Comput Softw Big Sci (2019) 3:13]
Kljinsma, Pedro, Tran, [Mach. Learn.: Sci. Technol. 2 (2021) 035005]
Kljinsma, Pedro, Tran, [IEEE/ACM H2RC 2020]
Wang,Yang,Flechas,Hawks,Holzman,Knoepfel,Pedro,Tran, [arXiv:2009.04509]
Cai,Herner,Yang,Wang,Flechas,Holzman,Pedro,Tran, [arXiv:2301.04633]

- To alleviate future HEP **computing will be bottlenecks** - **enable more powerful algorithms** on optimal hardware
- **Coprocessors** (GPUs, FPGAs, ASICs, …) naturally accelerate ML workloads **by orders of magnitude**
- No way to guarantee access to HW at all production sites
- Leverage industry hardware and tools - provide **coprocessors as-a-service**



SONIC:

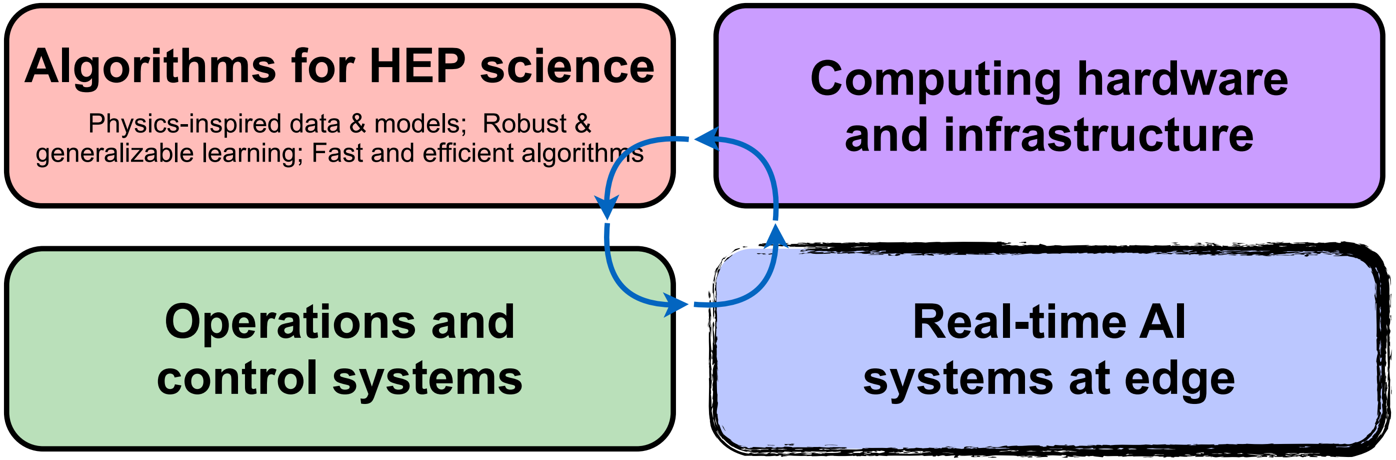
Services for Optimized Network Inference on Coprocessors

- Explore with on-prem, clouds, HPC and also for analysis facilities for all types of emerging hardware
- Testing now on CMS production workflows for Run 3
- ProtoDUNE production run (~7M) events demonstrates > 2x acceleration with GPU

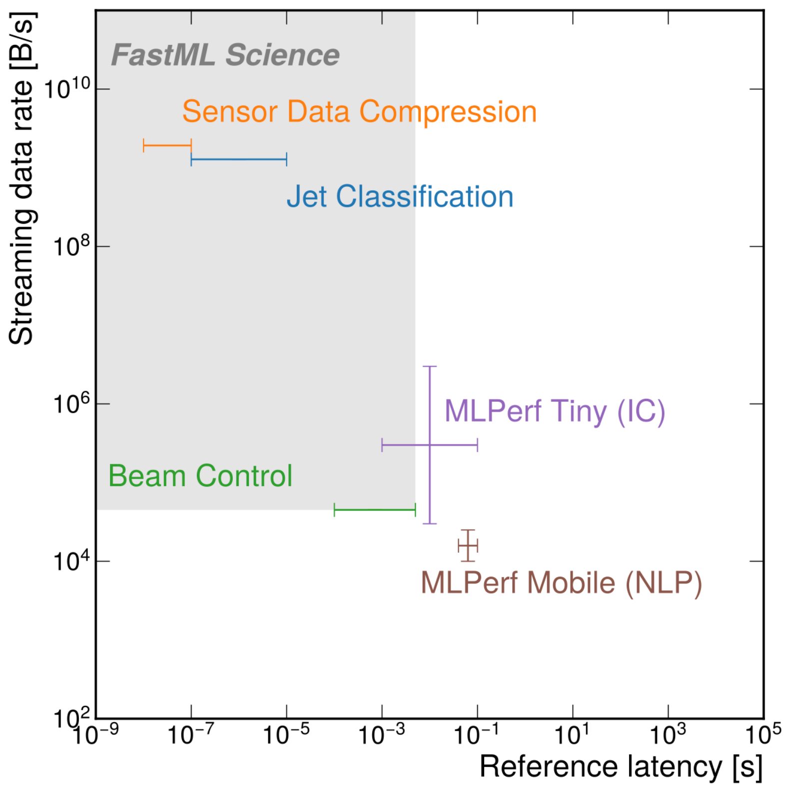
Monitoring of 100 GPU run

Fermilab AI-in-production workshop coming soon!

Program context



“Fast” ML at the extreme edge



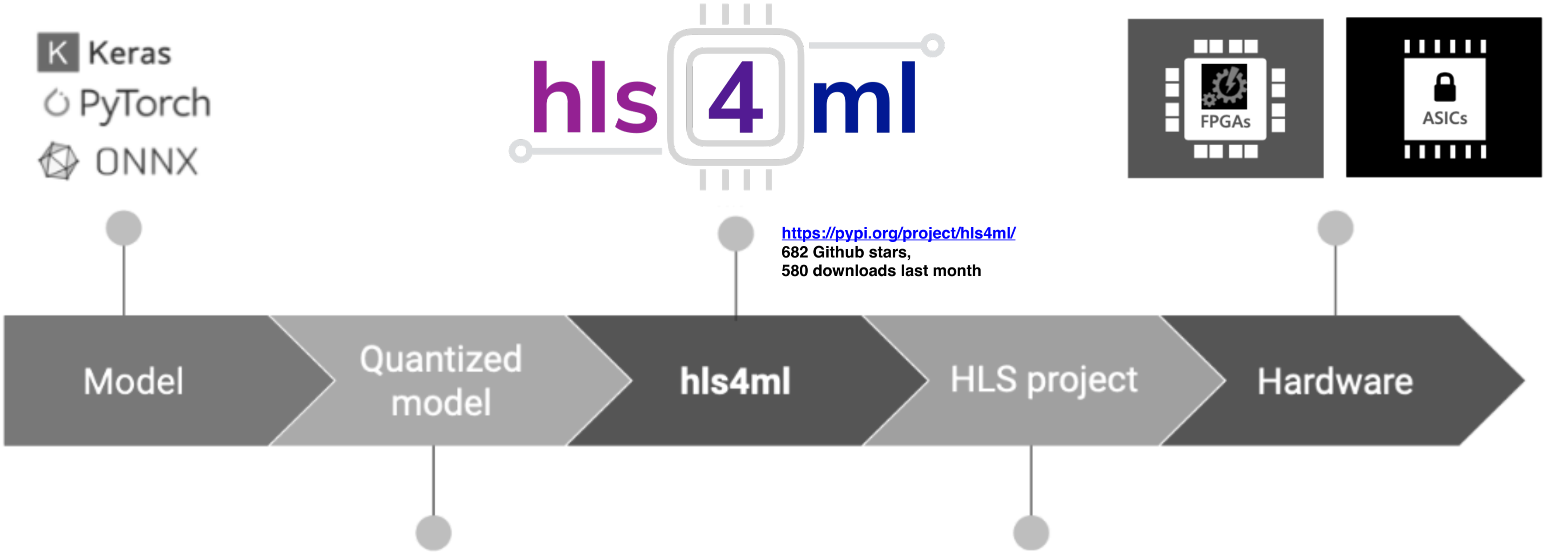
Cutting-edge scientific experiments explore nature at the **finest temporal and spatial scales**
Leads to data rates far surpassing industry — requires developing **innovative techniques**

- ML in specialized embedded architectures require in ***real-time*** to reduce and filter data
- Optimal data selection enables **more efficient operation and control, saves lost data, and accelerates time-to-discovery**

Efficient ML hardware software codesign

https://fastmachinelearning.org/hls4ml

Enabling efficient algorithms and workflows for non-experts into hardware



https://pypi.org/project/hls4ml/
682 Github stars,
580 downloads last month

QKeras (Google)
Brevitas (AMD)
HAWQ (UC Berkeley)
QONNX (Microsoft/AMD)

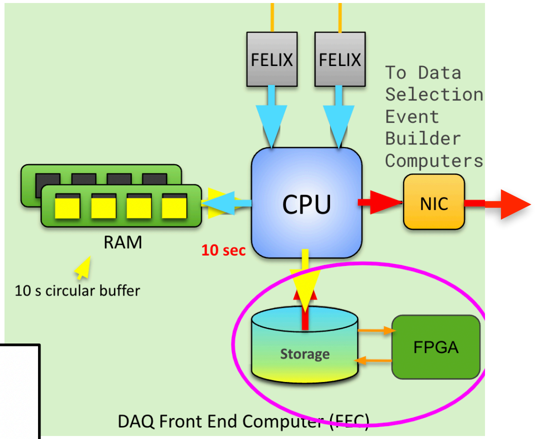
VIVADO
Mentor A Siemens Business
Quartus Prime

Ngadiuba, Tran et al., JINST 13 P07027 (2018)
https://fastmachinelearning.org/hls4ml
Ding, Hawks, Junk, Mitrevski, Wang, Yang, IEEE NSS

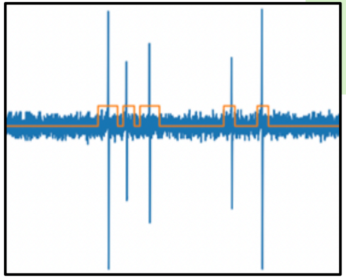
hls4ml for near-detector, low-latency

hls4ml in FPGA applied broadly across the sciences and beyond (more on this later)
HEP experiments, accelerator control, magnet training, nuclear physics, microscopy/material sciences, quantum controls/readout, fusion, …

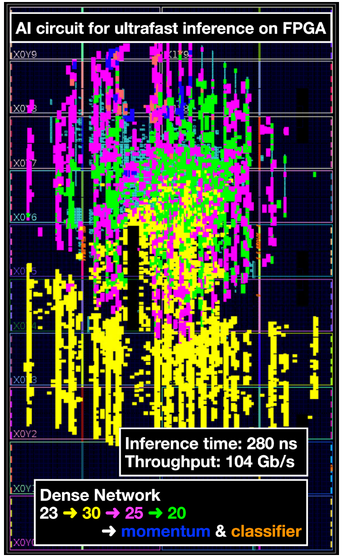
Per wire ROI finder for extracting low energy neutrino signals



Region of Interest



DUNE Supernova Filter + MMA



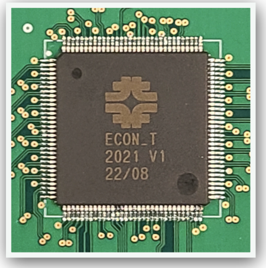
- Run 3 displaced muon ID **enables completely new capability;** muon momentum regression **cuts rate by > 2x** for HL-LHC
- **Active program**
  - Applications for Run 3 & HL-LHC from low-level data compression to cluster calibration to high level physics topology selections to anomaly detection

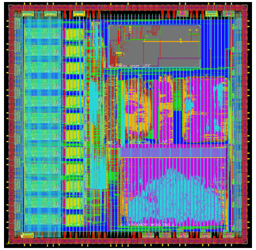
LHC Trigger - CMS and ATLAS

hls4ml for on-sensor/detector AI

Fahim ECA
Herwig, Hirschauer, Kwok, Ngadiuba, Tran, et al.,
[IEEE Trans. Nucl. Sci. 68, 2179 (2021)]
Dickinson et al. [CPAD talk]

On-detector/sensor AI can be a game-changer for extreme environments
Extreme data bandwidths, radiation environments, low power, cryogenic, etc.





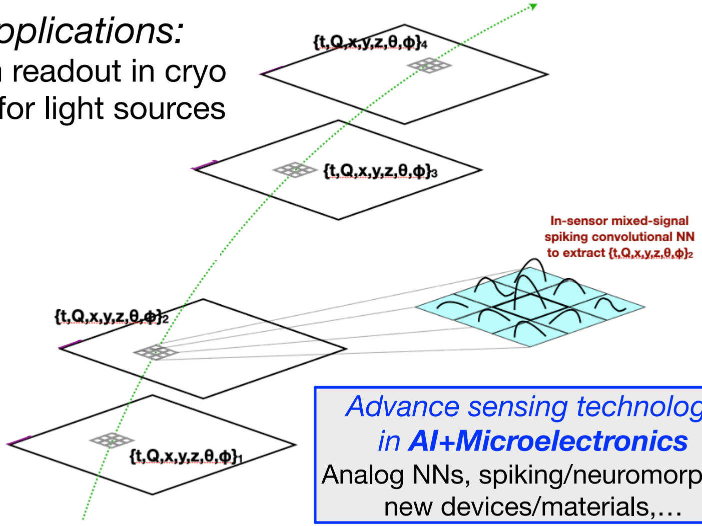
Data compression encoder ASIC for CMS HGCal

- First **design** and **implementation** of modern DL for HEP on ASIC
- **Enables powerful non-linear data compression schemes** on detector; better trigger primitives downstream
- Chips fabricated and tested, performed well under functional/ radiation validation

Pushing state-of-the-art of technology

Goal: **40 MHz pixel detectors**

Other applications:
Quantum readout in cryo
Sensors for light sources
etc.



*Advance sensing technology in **AI+Microelectronics***
Analog NNs, spiking/neuromorphic, new devices/materials,…

See more in Farah's talk next!

Outline

- Vision & strategic drivers
- AI Project Office and program organization
- Program milestones and highlights
- **Leveraging unique & core capabilities**

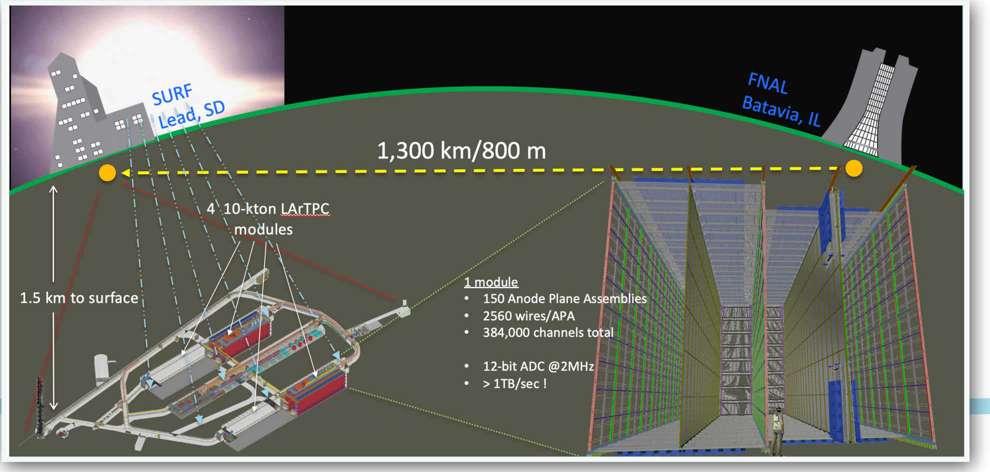
Fermilab AI strengths

- Expertise in state-of-the-art detectors, accelerators, and device readout synergistic with **real-time, edge AI and intelligent sensing**
  - *Complementary* to supercomputing and HPC facilities
  - Strong community around "Fast ML" collective in the past 4-5 years

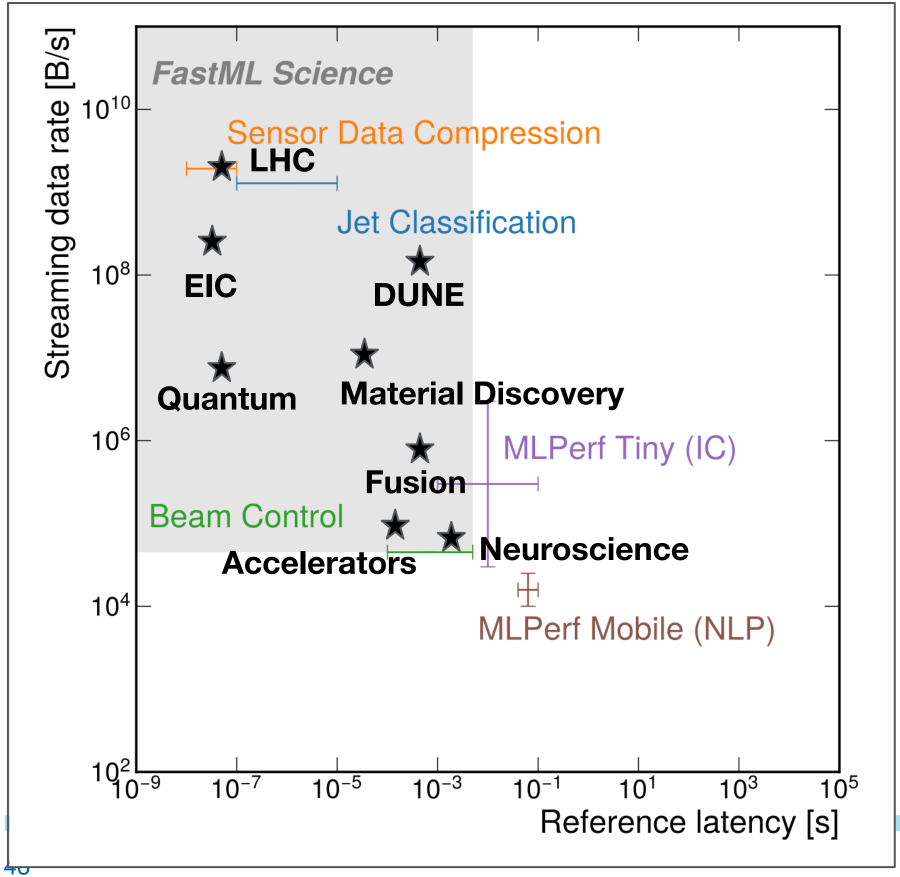
- Additional focus areas w/strong connections to (FNAL-led or other) AI centers
  - **Automated operations and digital twins** — accelerators and other large experiment controls
  - Needs for science
    - **Automated scientific method and discovery**
    - **Uncertainty quantification** (error bars), **Bias/domain shift** (domain adaptation)

Grand challenges towards AI centers

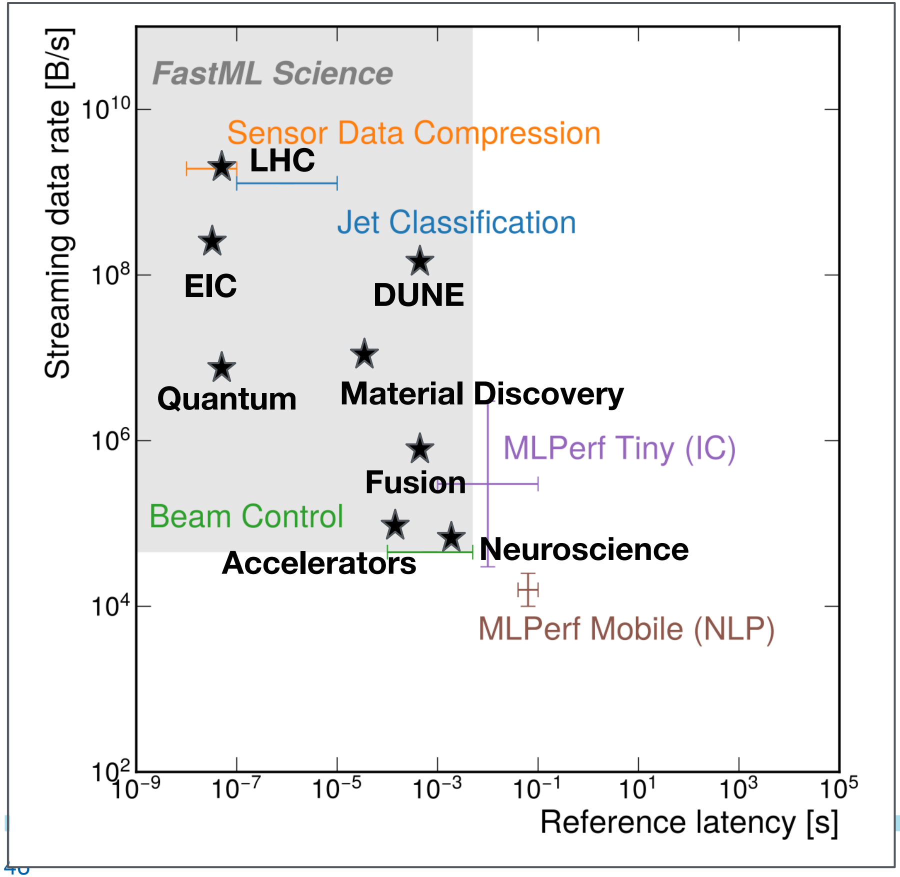
- **Real-time edge AI center driven by grand scientific challenges**
  - A multi-faceted DUNE program sensitive to extremely rare signals
    - Supernova burst, proton decay, neutrinoless double beta decay
  - LHC and future energy frontier experiments that can analyze every collision (e.g. complete 40MHz readout)
  - Automated accelerator complex driven by AI agents and digital twins



Grand Challenges in HEP and beyond



Grand Challenges in HEP and beyond



- Development of unique techniques to democratize edge AI, build benchmarks and community tools (hls4ml, open data, SONIC, DeepBench, Fast ML benchmarks,…)
- Good partnerships w/multi-disciplinary collaborators in electrical/computer engineering, core AI, computing (HPC labs), and industry
- Connected to other research focus areas on robust AI, domain shift, UQ

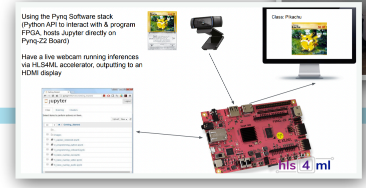
Examples of on-going DOE awards

- READS for **Fermilab accelerator controls** [DOE HEP]
- Extreme data reduction at the edge: **CMS and 4D Transmission Edge Microscopy** [DOE ASCR]
- Autonomous triggers for **sPHENIX/EIC** [DOE NP]
- **Efficient AI from physics phenomena** [DOE HEP]
- **Smart Pixels** [DOE HEP]

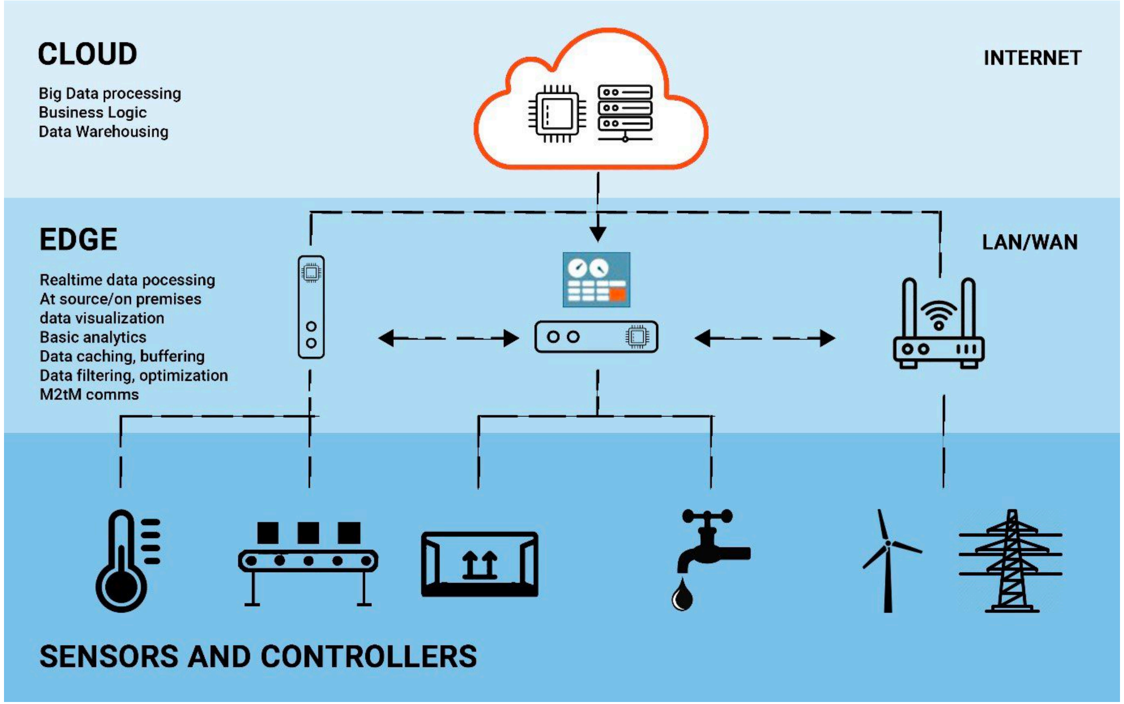
+ collaborations for other areas (quantum readout, fusion, …)

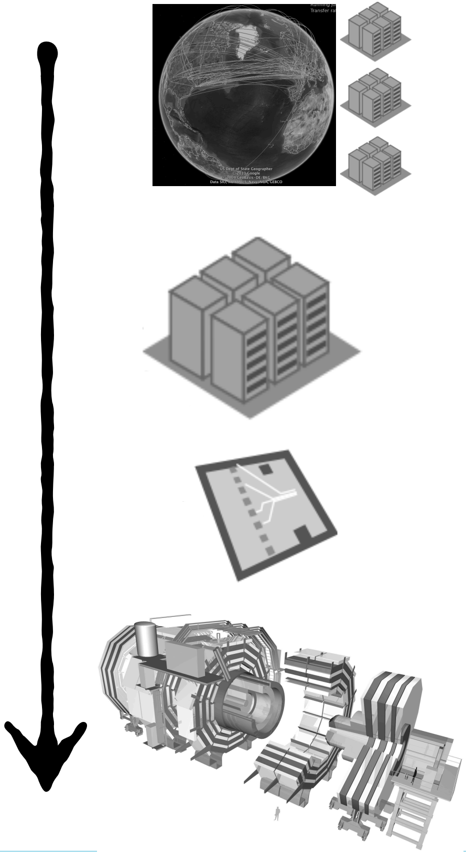
Workforce development and outreach

- Focused on bridging the gap between AI and HEP researchers
  - **AI associate program** has brought researchers from different backgrounds (CS, ECE) to Fermilab than usual
  - **Multi-disciplinary collaborations cross-pollinate** research teams and backgrounds
- **Outreach, education, tech transfer**
  - Major thrust of program is developing tools for science and industry
  - hls4ml tutorials, demos, and materials are a part of graduate school curriculum for ECE class, physics and engineering schools and conferences, and broader tech conferences



Connection to society and industry





Fermilab

Connection to society and industry

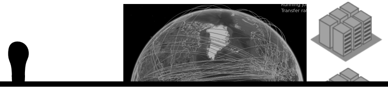
MLCommons launches machine learning benchmark for devices like smartwatches and voice assistants

by Ben Wodecki 6/16/2021

With experts from Qualcomm, Fermilab, and Google aiding in its development

MLCommons, the open engineering consortium behind the MLPerf benchmark test, has launched a new measurement suite aimed at 'tiny' devices like smartwatches and voice assistants.

MLPerf Tiny Inference is designed to compare performance of embedded devices and models with a footprint of 100kB or less, by measuring



INTER

LAN/W

SENSORS AND CONTROLLERS

ML Commons

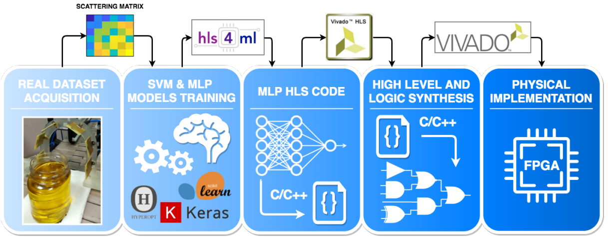
TINY ML

AMD

XILINX

SIEMENS

Mentor®
A Siemens Business



Next steps and collaborations

- **Step 1: Continue delivering cutting edge AI science and technology**
- **Collaborations and partnerships**
    - Large community of HEP universities and labs
    - Strong computer science and engineering collaborations and growing!
        - University groups and national labs
    - Industry connections
        - Established collaborations with hardware and software industry leaders: Nvidia, Microsoft, AMD/Xilinx, Siemens/Mentor,…
        - Developing connections with application areas, e.g. important and large Siemens customer led to common projects, Hawkeye360 (Satellites), etc.
    - Continue to build up more local connections, particularly with UChicago and ANL

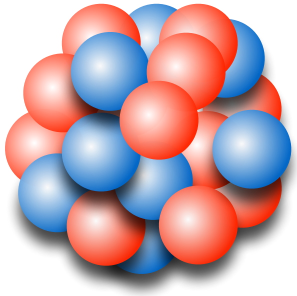
Executive summary (reprise)

- **Fermilab AI/ML program focused on *accelerating science***
  - Program pillars connect algorithm advancements with sensing, computing, and operations to solve HEP challenges
  - Identified areas where Fermilab contributes to the greater DOE AI needs
- AI Project Office coordinating overall **strategy and building community**
- Portfolio of research strong case for AI center involvement
  - Center lead would focus on real-time AI and edge sensing
    - Additional focus areas could complement other centers (digital twins, automated discovery and design)
  - Modest funds needed to seed efforts during **upcoming critical 1 year period**
    - Opportunities to develop collaborations & projects focused on *core AI research, strategic HEP applications, and industry/academic partnerships*

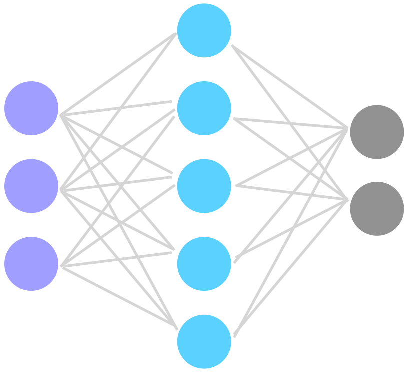
Extra

Understanding how protons and neutrons self-organize to form atomic nuclei requires solving the nuclear many-body Schrödinger equation

$$H\Psi = E\Psi$$



Machine learning methods allows to devise accurate **nuclear wave functions** suitable for quantum Monte Carlo calculations that **do not scale exponentially** with the number of nucleons



An artificial neural network wave function that involves additional ``hidden'' degrees of freedom has been introduced to improve the accuracy of the solution systemically

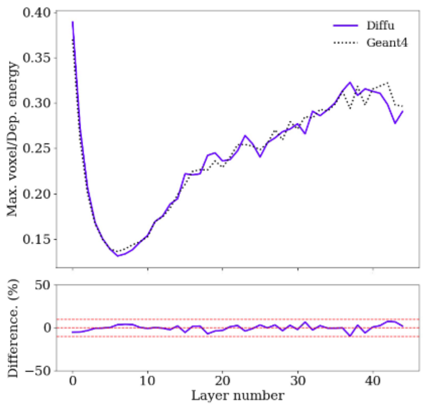
$$\Psi_{HN}(R,S) \equiv \det \begin{bmatrix} \phi_v(R,S) & \phi_v(R_h,S_h) \\ \chi_h(R,S) & \chi_h(R_h,S_h) \end{bmatrix}$$

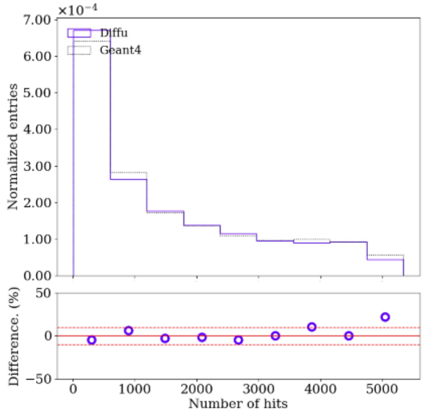
Light and medium-mass nuclei's energy and spatial density distributions are in excellent agreement with those obtained utilizing exact-diagonalization and diffusion Monte Carlo approaches.

CaloDiffusion: ML for Simulation

O. Amram, A. Lewis, K. Pedro

- Full detector simulation using Geant4 is too slow to keep up with HL-LHC data volumes and computing constraints
- Use generative ML techniques to increase speed & retain accuracy



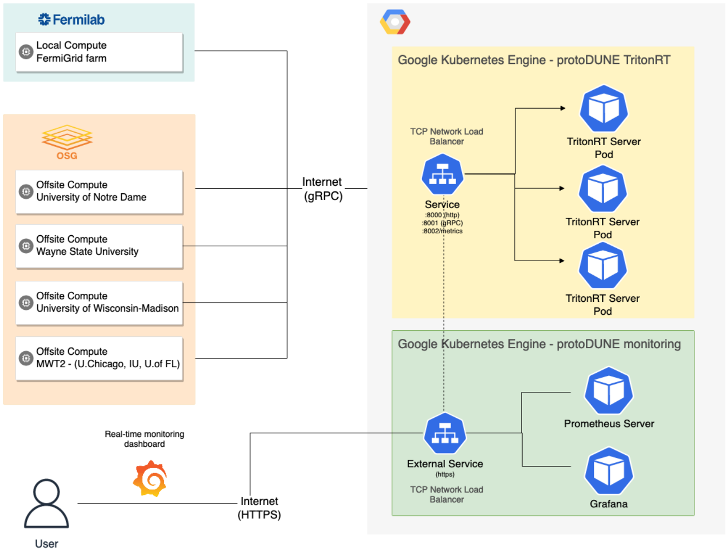


- Diffusion model: avoids pitfalls of GANs, high quality output
  - w/ improvements (preprocessing, RZ conditioning, cylindrical convolutions, cosine noise schedule), competitive results on CaloChallenge dataset
- Next steps: latent space optimization, reduce # diffusion steps, hybrid approaches w/ classical FastSim

Inference acceleration and real-time applications

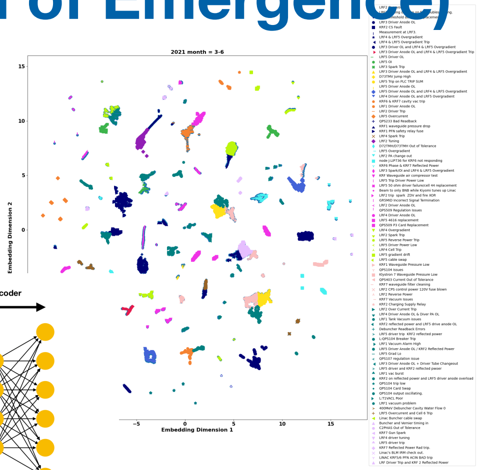
Wang et al. Front. Big Data 3 (2021) 604083
Cai et al. arXiv:2301.04633

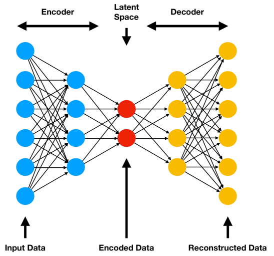
- **NuSONIC**
  - Accelerate ML inference using GPUs/FPGAs
  - Processing ProtoDUNE data using GPUs on the google cloud.
  - Saturation from network bandwidth well understood and important for grid jobs
- **Real-time applications**
  - Physics Inspired Neural Nets (PINNs).
  - AI for event triggering (new DOE AI funding)
  - Applications for ICEBERG, SBND and DUNE



L-CAPE (Linac Conditional Anomaly Prediction of Emergence)

- The 'L-CAPE' effort is to use ML to automate accelerator operations which will result in higher efficiencies and cost savings.
- The challenge is being able to use large, diverse data sets with changing nominal operating points, to generate an accurate ML accelerator model. The ML needs to be accurate, reliable and nearly 'real time'.



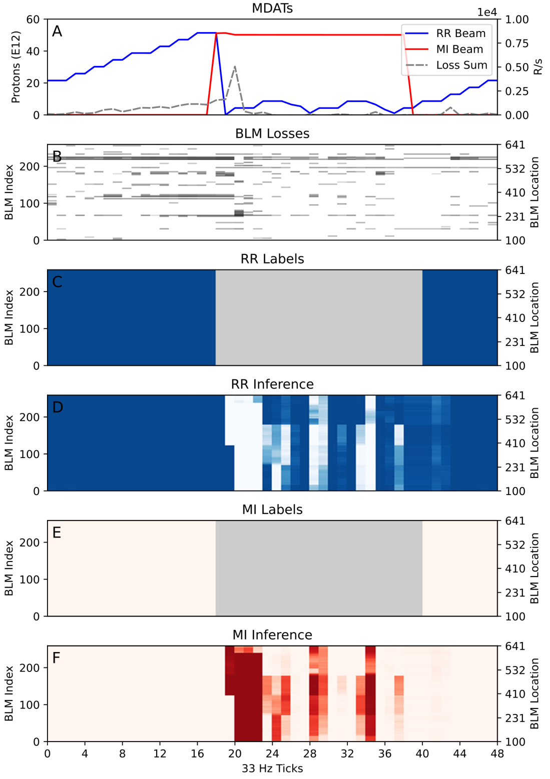


- The work has focused on an LSTM approach. Using over one year's worth of FNAL Linac data (2022), we have been able to train performant ML models. The results are encouraging, with most common Linac faults being identified, and some with actionable precursors.

Real-time Edge AI for Distributed Systems (READS)

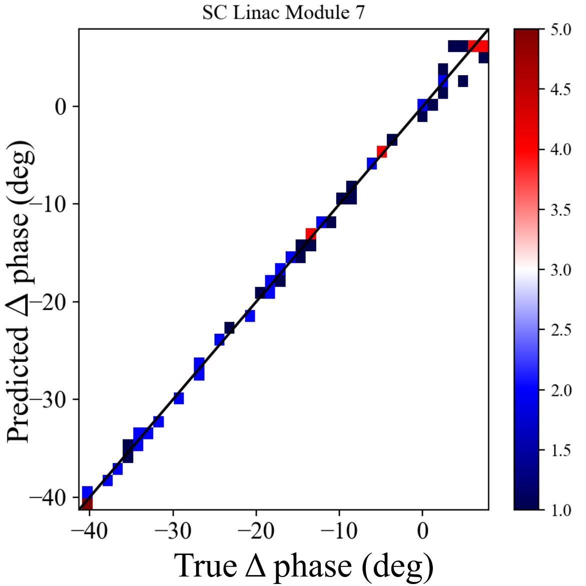
- Project 1: Main Injector Enclosure Beam Loss Disentangling
  - Infer real-time the machine origin of beam loss (Main Injector or Recycler) → upgrade machine protection, uptime and tuning
- Project 2: Mu2e Slow Spill Regulation
  - Improve the linearity of the Delivery Ring resonant extraction (Spill Duty Factor, SDF) → improve Mu2e experiment data collection

- Solution
  - Create and stream distributed readings from around the accelerator complex to perform near real-time inferences using fast FPGA hardware



Linac RF Optimization with ML

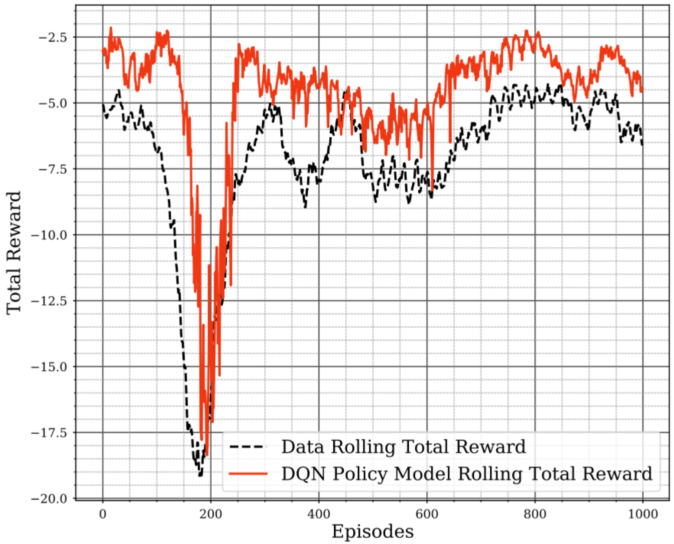
- Linac requires daily tuning of RF parameters to deliver stable beam energy with minimal particle loss
- Currently done manually: limited by expert availability and cannot optimize in multi-dimensional parameter space



- We are developing ML models that predict RF parameter settings to keep beam energy constant and minimize emittance for automated tuning
- Successful proof-of-concept test of single cavity phase regulation [link]
- Multi-cavity modeling promising, working on incorporating time-drift effect

GMPS AI (LDRD)

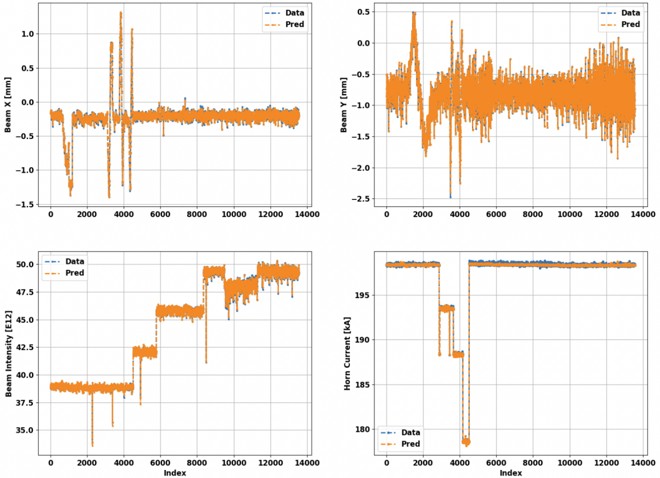
- The goal of the LDRD was to test the suitability of an FPGA-based “realtime” AI controller for the Booster Gradient Magnet Power Supply (GMPS).
- We chose to study reinforcement learning (RL) methods and use on-chip AI models – both are interesting challenges. RL models are notoriously data-hungry (can we produce enough data for training?), and an on-chip solution required modifications to HLS4ML in order to work on Intel-based FPGAs.



- The LDRD wrapped up in 2022 without full deployment. We are seeking additional funds to extend the project. However, we demonstrated the suitability of the FPGA platform and tested digital twins for training.
- See Phy. Rev. Accel. Beams 24, 104601

AI/ML for NuMI Beam Variable Predictions

- The goal of this project is to predict the NuMI beamline variables by taking account downstream muon monitor signals.
- These predictions give an independent measurement of the beam variables to monitor the quality of the beam delivery and also to detect the anomalies.
- In this approach, we use artificial neural networks to build a model to predict the proton beam position, the beam intensity and the horn current.



- Our results demonstrate the capability of developing useful ML applications for future beamlines such as DUNE
- This ML application can be used to reduce the neutrino flux systematics with the help of simulation studies

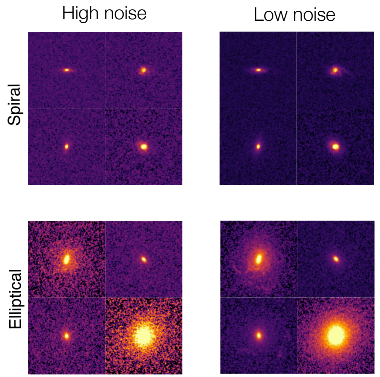
Deep Universal Domain Adaptation for cosmic Analysis

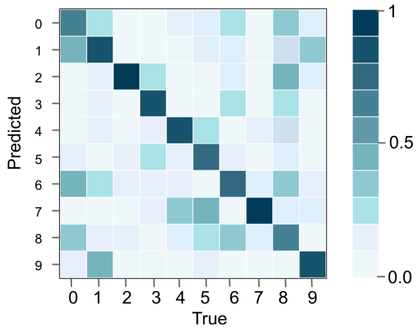
Deep Universal Domain Adaptation
Ciprijanovic, Lewis, Pedro, Madireddy, Nord, Perdue, Wild
(2022 in Neurips Workshop, 2023 in prep for journal)
CMS, Stealth SUSY search arXiv:2102.06976

- Goal:
  - **Adapt neural models** from training sets to observations.
- Problem:
  - **Training inevitably incurs bias** in neural models.
  - Training data sets (either simulated or observed) are inevitably different than new observational data.
- New Approach:
  - **Universal domain adaptation** is a new DA method that a) reduces the need hyperparameter tuning and b) reduces the requirement for overlap between training and observed data.
- Applications:
  - **Objects**: Strong Lenses, Spectra, Quasars, Galaxy Clusters
  - **Surveys**: DES, LSST, CMB-S4
  - **Connections**: particle inference

- Unsupervised domain adaptation from gradient reversal is used for Stealth SUSY background estimation

Example images of simulated galaxy morphologies with different levels of telescope noise.





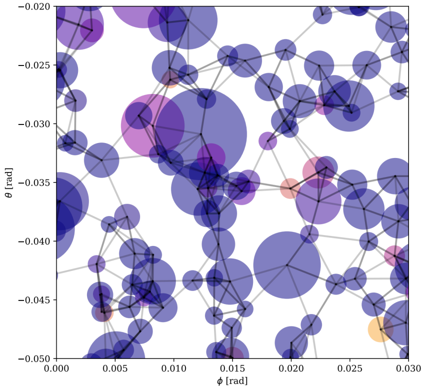
Confusion matrix for classification of galaxy types using DUDA

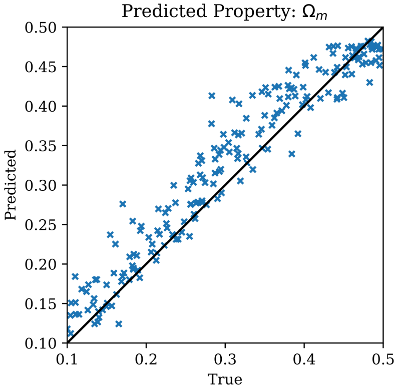
Self-Driving Telescopes

Spectroscopic Survey Optimization
Cranmer, Melchior, Nord, 2021 (Neurips workshop)

- Goal:
  - **Adaptively optimize** telescope survey scheduling
- Problem:
  - **Hand-crafted** observation schedules are **prohibitively expensive** and **don't adapt** to new info from environment or from data.
- New Approaches:
  - **Unsupervised Graph Neural Networks**: optimize an observation strategy to constrain cosmological parameters.
  - **Supervised Reinforcement Learning**: build a decision-making algorithm to prepare or adapt observations.
- Applications:
  - **Instruments**: Imaging, spectroscopic, interferometric
  - **Surveys**: Queue observations, DES, LSST, CMB-S4, +
  - **Connections**: accelerator tuning

A network of galaxies optimally selected for cosmic matter estimation.



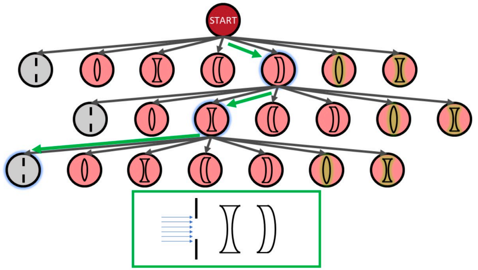


Optimized estimates of cosmic matter density for many different simulated universes.

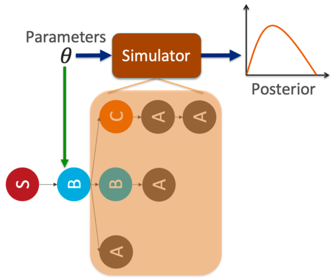
Automated Instrument Design

Optical System Design
Cohen (HS student) and Nord, 2023 (in prep.)

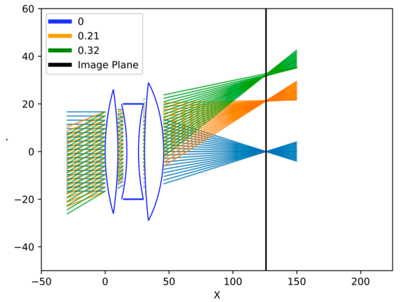
- Goal:
  - **Automate design** of telescope optics
- Problem:
  - **Humans run hand-crafted**, time-consuming simulations in **expensive** software to make guesses at optical system setup.
- New Approach:
  - Combine **binary trees** and **simulation-based Inference**
  - We can both **arrange optics** (with tree-based decision-making) and **predict optical element** shape parameters (with SBI).
  - We produce probabilistic outputs for optical systems so the human can make an informed decision.
- Applications:
  - **Instruments**: Imaging, spectroscopic, interferometric
  - **Surveys**: future survey instruments
  - **Connections**: accelerator design, symbolic regression



Schematic example of generating an optical system with our algo. Green arrows show optimized tree traversal.



Overview of algorithm. Tree produces optical system. Posteriors are of element shape parameters.



Example prediction for design of an optimized 3-element system.

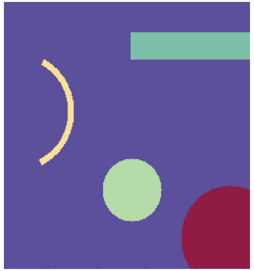
DeepBench - A simulation library for cosmology focu

- Motivations for DeepBench:
  - Beginner-friendly
  - Faster training convergence
  - Fills gap in benchmark dataset complexity

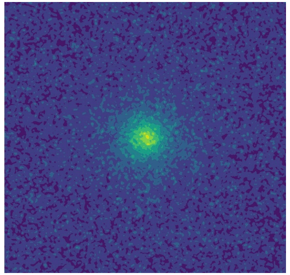
- Useful features:
  - Astronomical object profile simulations of varying complexity
  - Flexible parameter input requirements
  - Quick creation of benchmark dataset

Votberg, M., Lewis, A., Nord, B.

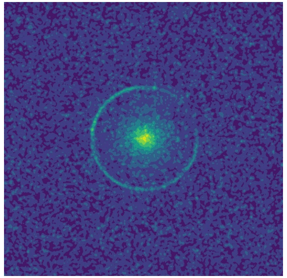
https://github.com/AeRabelais/DeepBench



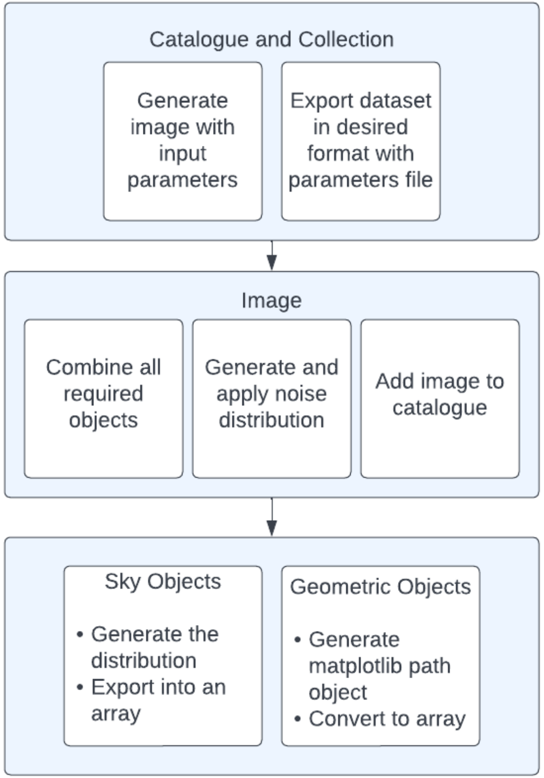
Geometric Image



Galaxy Image



Gravitational Lens Image



The code itself has 3 main pieces: Catalogue and Collection (manages all images and exports the final dataset), Image (composes the objects into a single file and adds noise if needed), and individual Objects (generate geometric shapes and astronomical objects).