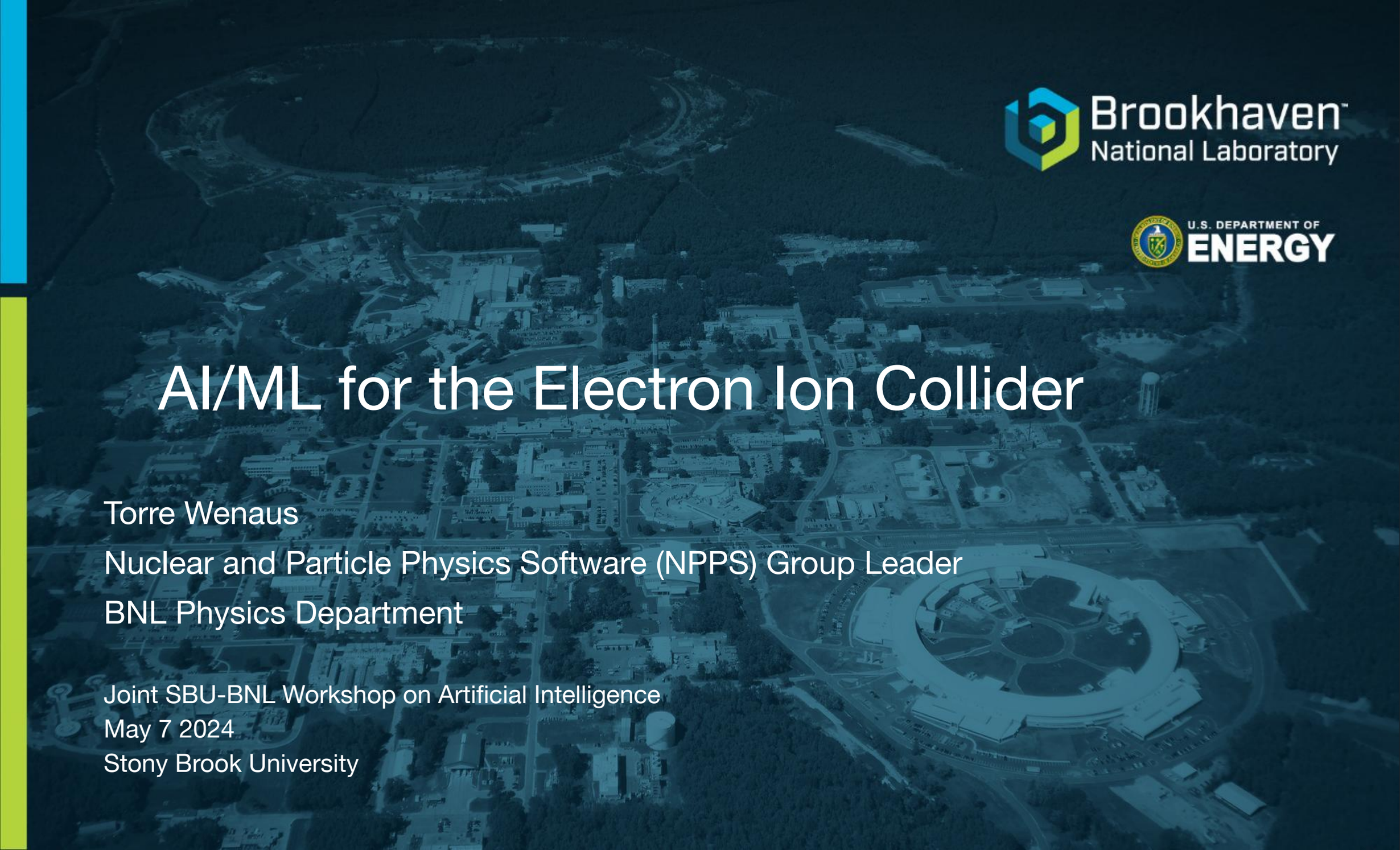


AI/ML for the Electron Ion Collider

Torre Wenaus

Nuclear and Particle Physics Software (NPPS) Group Leader

BNL Physics Department

Joint SBU-BNL Workshop on Artificial Intelligence
May 7 2024
Stony Brook University

Outline

- Brief(est) overview of EIC and ePIC
- AI(*) for the EIC
  - Current and Near Term
  - Long Term
- Local Collaboration
- Conclusion
- *Learn more*
- *Supplementary*

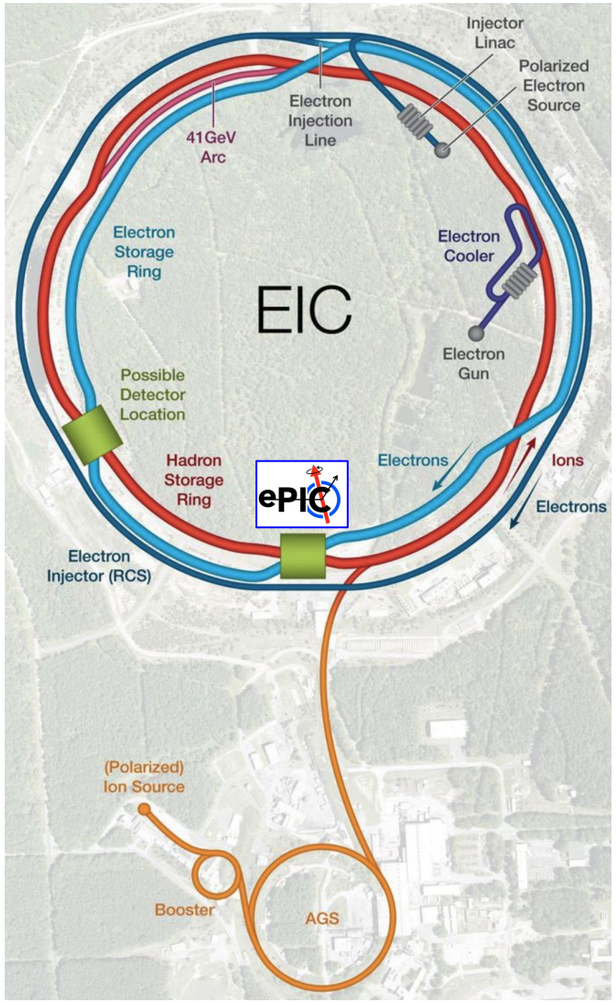
(*) For the purposes of this talk, AI == AI/ML

Not representative of all AI activity and planning in the EIC and ePIC community!

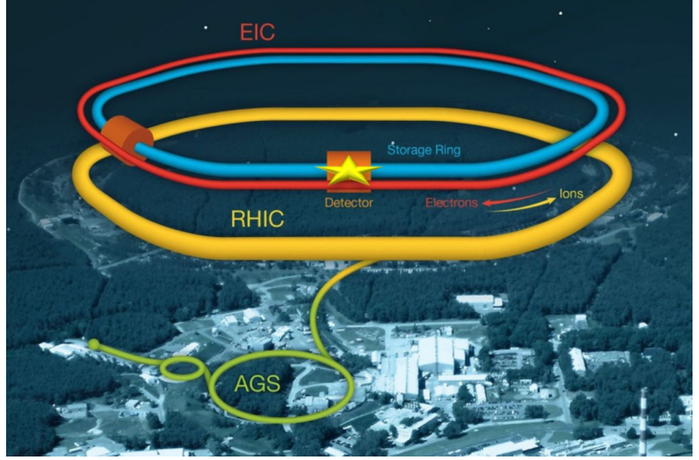
Drawing particularly on recent discussions at a BNL NPP Directorate ‘retreat’ on AI for EIC

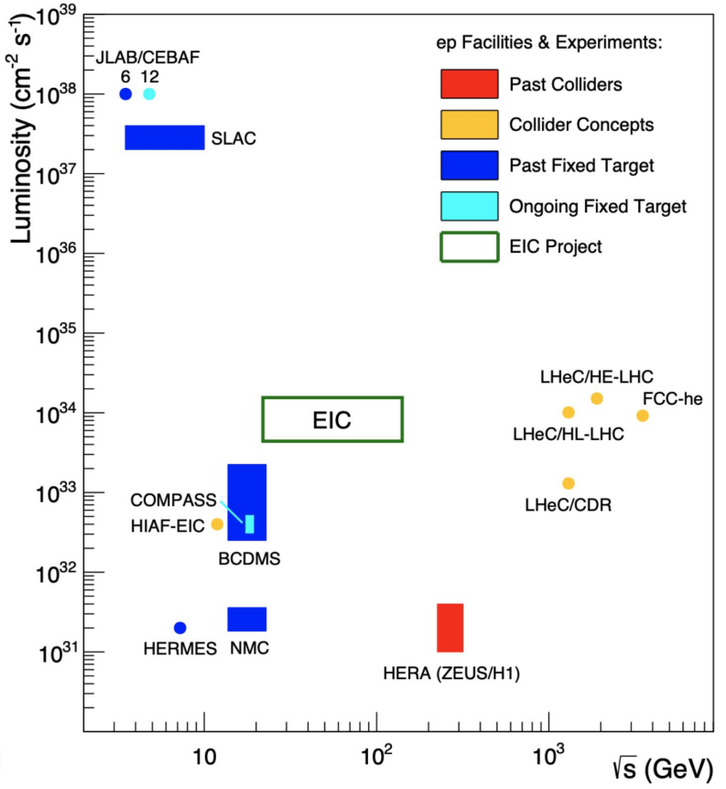
Covering experiment more than accelerator aspects, both rich in AI applications & potential

The Electron Ion Collider (EIC)



- RHIC's transformation into the EIC begins 2025 after sPHENIX/STAR physics program ends
- Physics datataking from ~2034
- It will be the first
  - electron-nucleus collider
  - high-luminosity electron-proton collider
  - e & p spin polarized collider
- Exploring the QCD frontier inside the nucleus
  - Nucleon structure - full 3D spatial and momentum structure, spin structure
  - Origin of nucleon mass
  - Precision study of proton spin
  - Emergent properties of a dense system of gluons
  - *Every collision event has physics interest*
- A unique project structure: two host labs, BNL and Jefferson Lab in Virginia
  - Close collaboration on all aspects: machine, detector, software and computing
- Together with a global community collaborating on the EIC



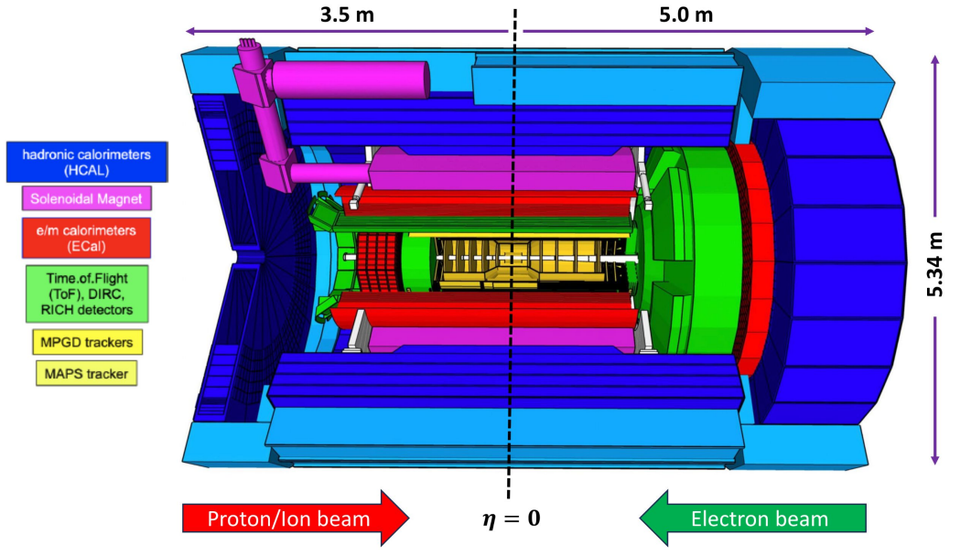


Brookhaven National Laboratory

The ePIC Detector at the EIC

170+ institutions, 24 countries, 500+ participants

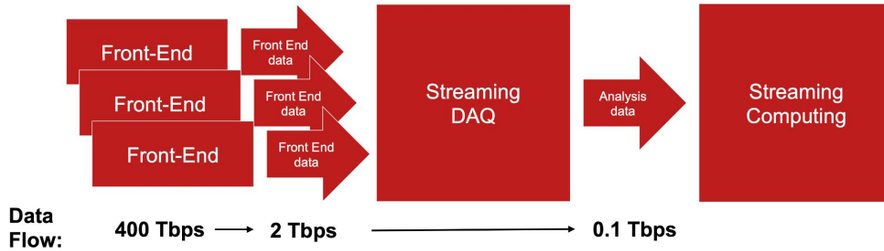
- DOE's EIC funding includes the ePIC detector, designed as the O(10m) 'electron microscope' probing e-p, e-ion collisions
  - including also far forward/backward detectors extending +/-~45m
- 17+ detector subsystems
  - charged particle tracking, vertex finding
  - particle identification
  - electromagnetic and hadronic calorimetry
  - precision requirements, many leading edge technologies
- A new 2.8m bore 1.7 tesla superconducting solenoid
- As hermetic a detector as possible
  - Asymmetric detector systems optimized for the different particle types and energies in the two directions
- Tightly integrated with the EIC machine and beamline
- *Streaming readout to capture every collision event*



Building a second EIC detector is a goal of the global EIC community to fully exploit the EIC's physics potential, how to fund it is TBD!

ePIC Streaming Readout: Maximizing Physics Reach

- While EIC luminosity is very high, the cross section is not
- Making it **tractable to read out ~100%** of the events
  - No trigger filter in the data acquisition
- It means reading out background as well as signal
  - For the benefit of a **complete, unbiased event sample**
  - Noise reduction, lossy compression applied to reduce storage/cost



- Data acquisition captures ~microsecond-wide '**time slices**' containing all detector data for hundreds of collision events
- Data is streamed to prompt reconstruction for an **early holistic view of the data**, its anomalies, calibration and analysis
  - Reconstruct, monitor/diagnose, calibrate, analyze as quickly as possible
  - Feeds directly into detector understanding, performance, timelines and quality of physics output
- A technique presenting many challenges from front-end electronics to the analysis
  - Making its first appearance in current/near term experiments e.g. sPHENIX at RHIC, and at the LHC
- ***With many opportunities for applying AI***

[ePIC streaming computing model report V1, Oct 2023]

AI for the EIC - Current and Near Term

- The 2023 Nuclear Science Advisory Committee Long Range Plan stated that "*EIC could be one of the first large-scale collider-based programs in which AI/ML is integrated from the start.*"
  - We've just left the starting gate and we've begun the work to achieve this!
- AI is already "*a key part of all software and computing working groups in ePIC*" - ePIC Spokesperson John LaJoie at AI4EIC last fall
  - ML-based algorithms supported and **used in the software framework**
  - **Applications** developing in fast calibration, streaming, particle identification, many others
  - Working on **centralized services** for training, model management, workflow integration
  - Monitoring integration of AI methods in monthly **production campaigns as progress metric**
  - **Data and analysis preservation**: key to leveraging rapidly evolving AI approaches
  - The first "**AI Challenge**" will shortly be presented to the collaboration (this year)
- EIC/ePIC community is bringing AI experience from their other activities. Local examples:
  - DUNE (FNAL Long baseline neutrino): AI to understand and quantify systematic uncertainties
  - sPHENIX (BNL RHIC): Real-time data reduction, generative AI for simulation and analysis
  - ATLAS (CERN LHC): large scale distributed AI workflows

AI for the EIC Today: Examples with Local Contributions

The AI/ML systematic bias problem

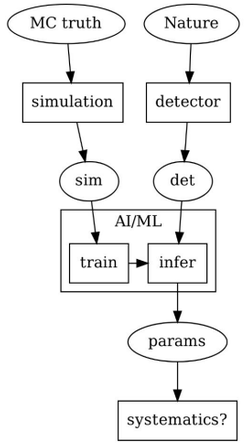
Typical AI/ML (+ conventional) processing chain ⟶

- **train on sim** and then **infer on det**
- "MC truth" ≠ "Nature", "simulation" ≠ "detector"

Cross-domain AI/ML inference is systematically biased.

- The bias can be "precisely wrong" and difficult/impossible to estimate with conventional analysis.
- Estimates based on the application of AI/ML inference to simulation ignores this cross-domain bias.

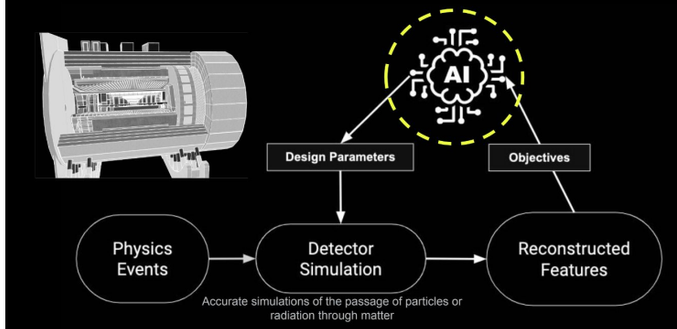
Need way to estimate systematics that is as precise as AI/ML and that considers the bias between simulation and real data.
⇒ need yet more AI/ML!



LS4GAN project, Brett Viren (BNL DUNE)

AI-Assisted Detector Design

The AI-assisted design embraces all the main steps of the sim/reco/analysis pipeline…



- Benefits from rapid turnaround time from simulations to analysis of high-level reconstructed observables
- The EIC SW stack offers multiple features that facilitate AI-assisted design (e.g., modularity of simulation, reconstruction, analysis, easy access to design parameters, automated checks, etc.)
- Leverages heterogeneous computing

Provide a framework for an holistic optimization of the sub-detector system
A complex problem with (i) multiple design parameters, driven by (ii) multiple objectives (e.g., detector response, physics-driven, costs) subject to (iii) constraints

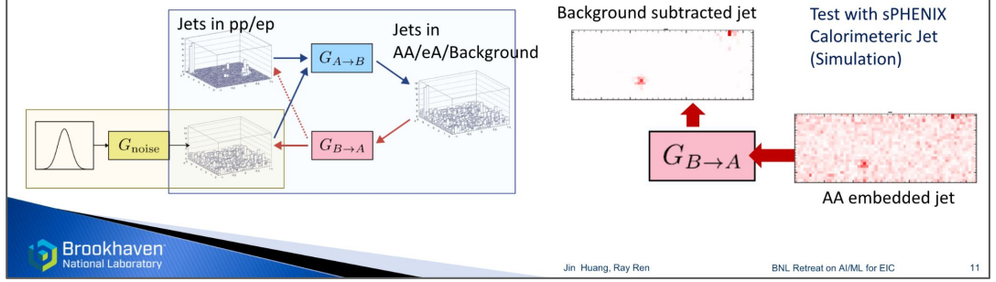
Those at EIC can be the first large-scale experiments ever realized with the assistance of AI

5

Cristiano Fanelli (CWM EIC)

Our approach: novel use of generative AI for analysis

- Last talk by Brett: cycle GAN used to bridge between simulation and reality
- It can also become an analysis tool to translate jets in pp/ep into jets in AA/eA/Background
- Self-supervised learning from either simulated jet embedding or real unpaired pp/AA real data.



Jin Huang (BNL eSPHENIX), Ray Ren (BNL CSI)

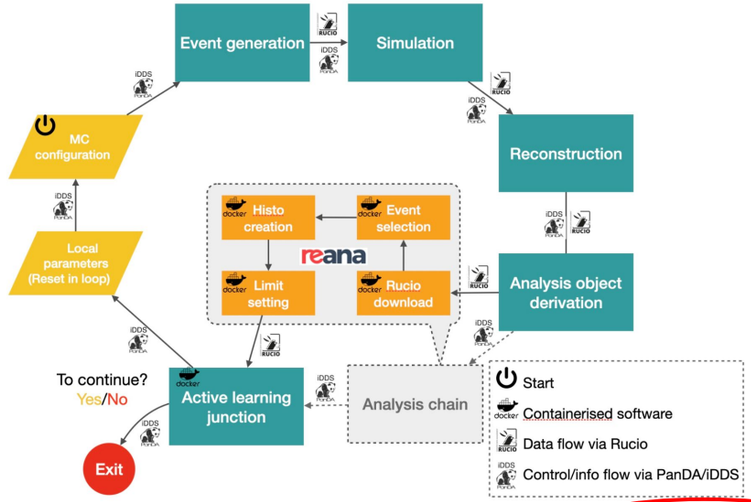


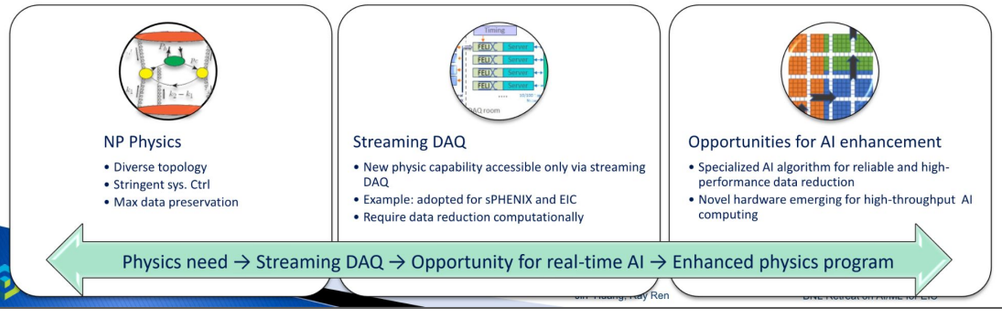
Figure 10: Bayesian optimisation based active learning with PanDA/iDDS and Rucio

Tadashi Maeno, Wen Guan (BNL ATLAS)

AI for the EIC Today: Local Examples 2

Real-time data reduction

- A few EIC subsystem has high noise/background rate that REQUIRE real-time data reduction computationally: dRICH, far detectors, calorimeters
- Our solutions specialized algorithm and hardware for efficient and high throughput real-time AI data reduction

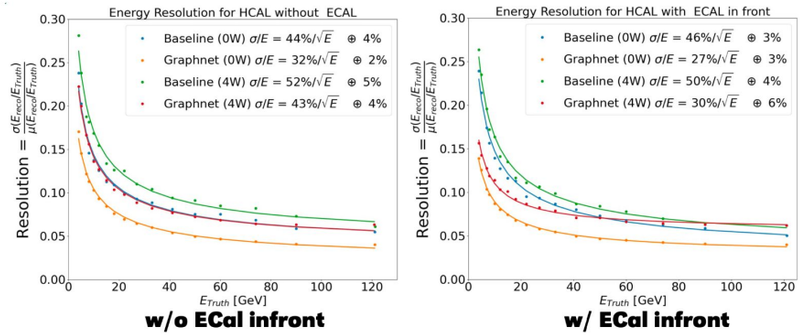


Jin Huang (BNL eSPHENIX)

Value Engineering through AI

thanks to Miguel Arratia Munoz UCR

Example: Optimization of ePIC forward HCal tower design
Software compensation using AI
→ eliminate expensive tungsten absorber plates
→ significant cost reduction in manufacturing



Brookhaven National Laboratory

Elke Aschenauer (BNL EIC)

Desired result: higher proton polarization

- What high-impact operational challenge can be addressed by MI/AI?
  ➔ Polarized protons.
- From the source to high energy RHIC experiments, 20% polarization is lost.
- Polarized luminosity for longitudinal collisions scales with P^4, i.e., a factor of 2 reduction!
- The proton polarization chain depends on a hose of delicate accelerator settings form Linac to the Booster, the AGS, and the RHIC ramp.
- Even 5% more polarization would be a significant achievement.

Brookhaven National Laboratory

Georg.Hoffstaetter@cornell.edu C-AD MAC 28 March 2024 3

Lucy Lin (BNL C-AD), Georg Hoffstaetter (Cornell)

AI for the EIC: Long Term

- Our recent BNL retreat included ‘blue sky’ discussions on the long term
- Within the context of a drumbeat of agency attention to foundation models, LLM technology
- Transient buzz, or lasting and building revolution? Planning for the latter
- EIC with its unbiased data sample encompassing deep emergent physics phenomena is fertile ground for exploring what we can learn
- Being ready requires 10 years of work, apace with the technology
- An R&D plan commencing now can deliver powerful capability over the decade
- We sketched out a possible timeline on two tracks, developing a ‘Large Particle Model’ and a cognizant integrated facility
  - Near term: 1-3 years
    - Complete R&D on using AI to draw trusted, quantified inferences from real data (ie continue the LS4GAN path)
    - Develop a prototype HENP LLM with current data (e.g. ATLAS, RHIC) and first generation feature extraction tools
  - Mid term: 3-5 years
    - Begin work on an integrated accelerator - experiment dataset and its AI instrumentation for a cognizant facility from accelerator to detector to analysis, targeting the EIC
  - Long term: 5-10 years
    - Documented analyses employing the HENP LLM (internal notes and/or peer reviewed publications)
    - Commission and deploy AI for EIC in parallel with machine and detector installation, commissioning and datataking

AI for the EIC: An Incomplete Sampling

Bold: local activity, current or near term plans

- Accelerator: **proton polarization**, **luminosity optimization**, **bunch merging**
- DAQ: **background/noise reduction**, **zero suppression**, **lossy compression**
- Controls: sub-second corrections/calibrations via AI-guided data prediction
- Conditions: smart conditions monitoring, problem prediction, anomaly detection
- Calibration/QA: fast calibration, **alignment**, real-time anomaly detection
- Operations: custom LLM chatbots, smart service status monitoring, log analysis
- Software: event generation, **fast simulation**, **reconstruction**, **detector design optimization**
- Distributed computing: **workflow/dataflow optimisation**, **resiliency**, **large scale distributed AI services**
- Accelerator-detector-DAQ-processing integration in a '**cognizant AI facility**'
- And throughout analysis
  - Physics object reconstruction, full-event analysis
  - **Understanding and quantifying uncertainties and systematics**
  - In the long term, insights from rapidly evolving AI technologies

Local Collaboration

- Local collaborating communities on AI for EIC, bringing together domain knowledge from NPP and CS AI expertise from CSI
  - EIC Project
  - EIC Group, BNL Physics Department
  - BNL Collider-Accelerator Department
  - Nuclear and Particle Physics Software Group, BNL Physics Department
  - Scientific Data and Computing Center (SDCC), BNL NPP/CSI
  - BNL Computational Science Initiative (CSI)

…within of course wider collaborations across EIC, NP, HEP, CS

I hope activities and plans I've mentioned help seed new ones!

Conclusion

- EIC aims to be the first large-scale collider-based program in which **AI is integrated from the start**
  - The ePIC experiment sees **AI as already a key tool** throughout
  - AI capable framework, applications developing throughout online, offline, analysis
  - ‘AI challenges’ within the collaboration will begin this year
- The EIC machine, detector(s), readout and analysis are **conducive to AI approaches**
  - **Integrated facility** from accelerator to detector to readout to analysis: ‘cognizant’ via AI
  - **Streaming readout** from front-end electronics to reconstruction/analysis in real time
    - Fast AI for on-the-fly data reduction, calibration, monitoring, reconstruction
  - A **complete unbiased data sample** dense with emergent physics phenomena
    - Will AI technologies 10 years from now be capable of deepening our physics insights?
    - Being ready to address the question requires R&D from today in a fast-moving field
- Locally, EIC AI benefits from **strong collaborations** among communities in this room, with much potential for more
  - RHIC experiments, HEP experiments (e.g. ATLAS, DUNE), BNL SDCC Computing Facility, BNL Computational Science Initiative (CSI), and further afield (DOE HEP-CCE, Leadership Computing Facilities)
  - Listening today to understand how we can add Stony Brook!

Thank you

Thank you to the team of BNL AI@EIC retreaters and others who contributed directly and indirectly to this talk!

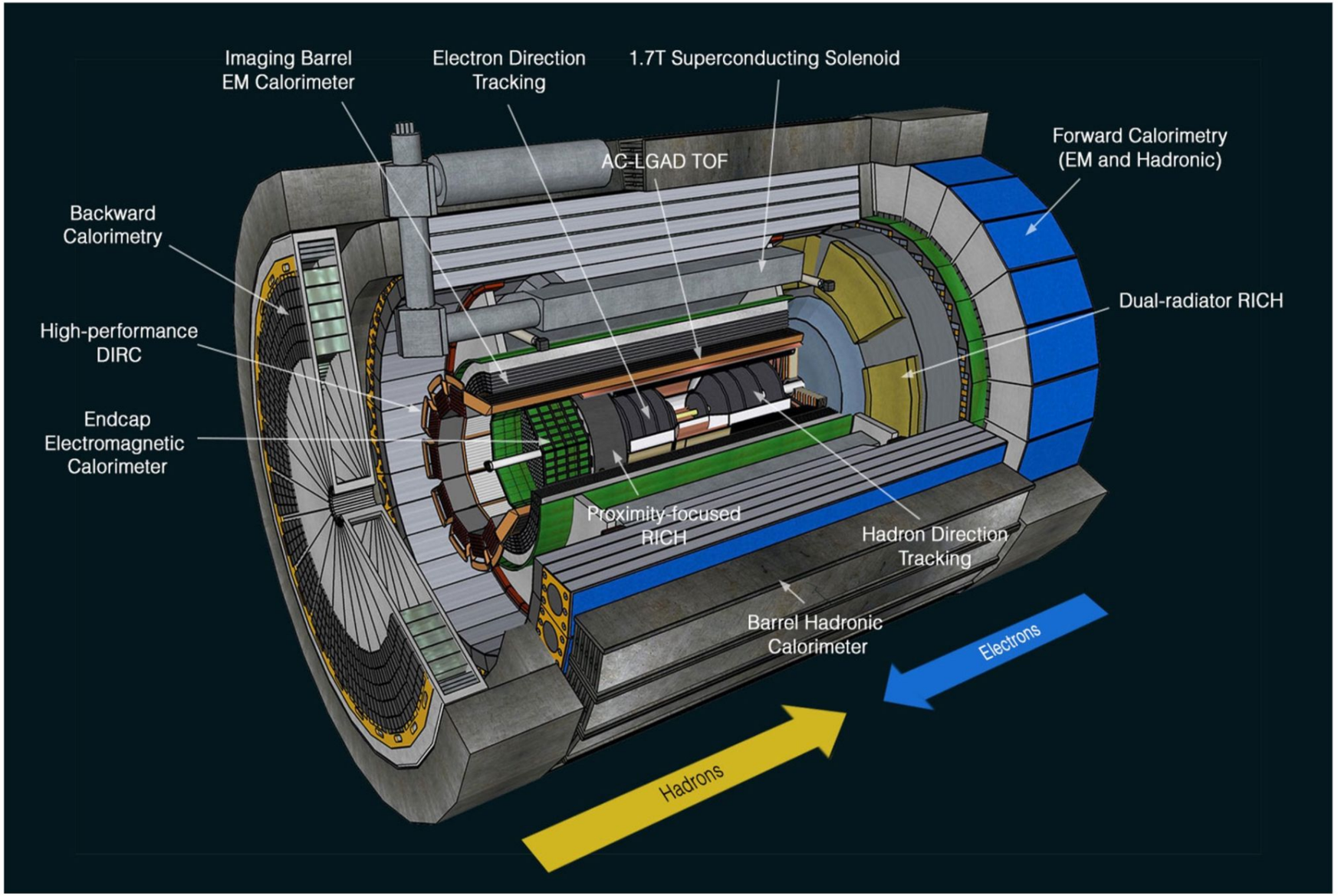
Elke Aschenauer, Kevin Brown, Ray Ren, Jamie Dunlop, Jin Huang, Brett Viren, Alexei Klimentov, Meifeng Lin, Shinjae Yoo, Adolfy Hoisie, Georg Hoffstaetter, Sergei Nagaitsev, Haiyan Gao, Hong Ma, Lucy Lin, Xiaofeng Gu, Jennefer Maldonado, Yuan Gao, Kolja Kauder, Frank Rathmann, Alexander Jentsch, John de Stefano, Shigeki Misawa, Dmitry Arkhipkin, and others

Learn More

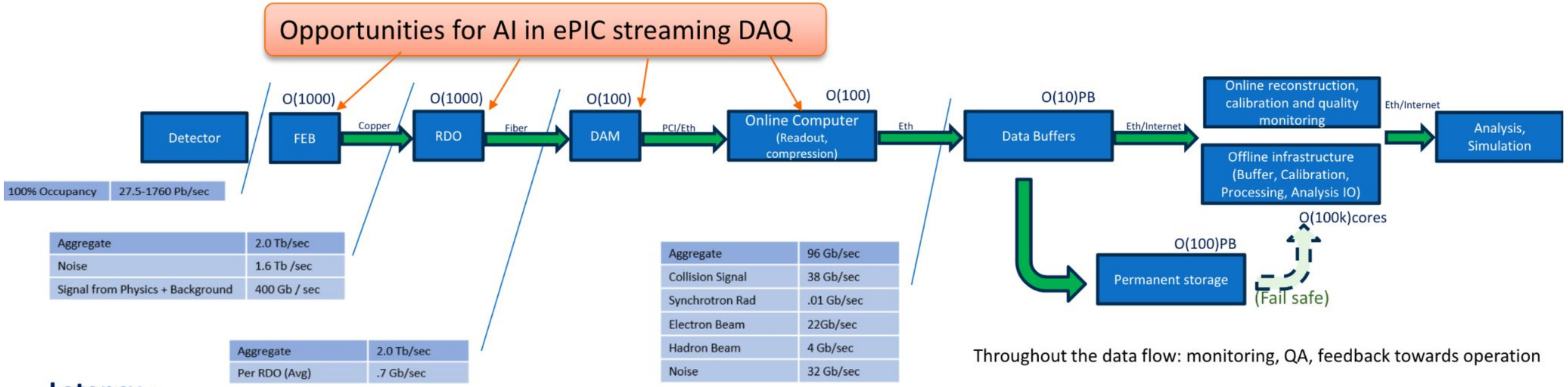
- EIC website
- Science Requirements and Detector Concepts for the Electron-Ion Collider: EIC Yellow Report
- ePIC Streaming Computing Model Report (first version, Oct 2023)
- AI4EIC 2023 Workshop (an ongoing series)

Supplementary

ePIC Central Detector



Opportunities for AI in ePIC streaming DAQ



Throughout the data flow: monitoring, QA, feedback towards operation

Latency :

0ns O(100)ns O(1)us O(10)us O(1)min O(1)min-O(1) day O(1)day-O(1)week

Possible facilities:

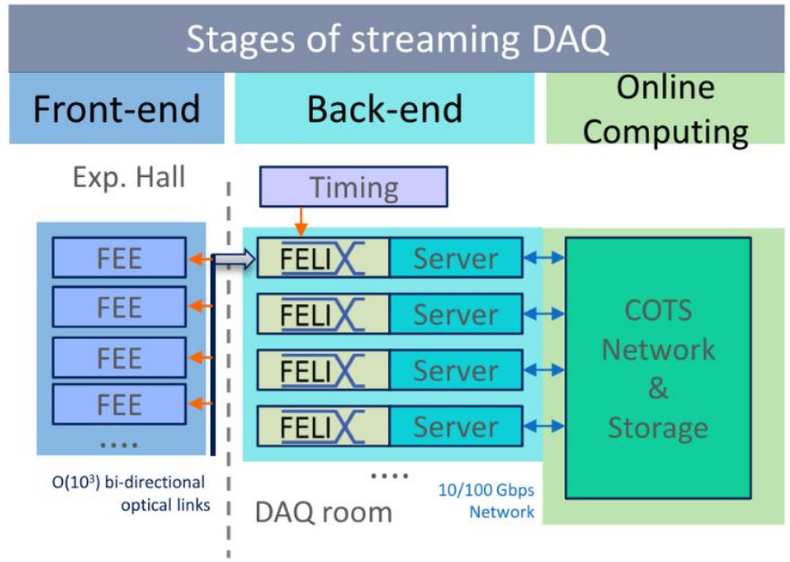
On detector On detector/rack DAQ room Host labs/Echelon 1 facility Remote resources

Reference:
- ePIC DAQ wiki: https://wiki.bnl.gov/EPIC/index.php?title=DAQ
- ECCE computing plan, *Nucl.Instrum.Meth.A* 1047 (2023) 167859

Brookhaven National Laboratory

AI in Streaming Readout DAQ

- Main challenge: data reduction
  - Traditional DAQ: triggering was the main method of data reduction, assisted by high level triggering/reconstruction, compression
  - Streaming DAQ need to reduce data computationally: zero-suppression, feature building, lossy compression
- Opportunities for Real-time AI
  - Emphasize on reliable data reduction, applicable at each stages of streaming DAQ: Front-end electronics, Readout Back-end , Online computing
  - Data quality monitoring, fast calibration/reconstruction/ feedback
    - Has many AI application too
    - Not focus of this talk, nonetheless important for NP experiments



The ePIC Computing Model: Streaming DAQ to Global Processing

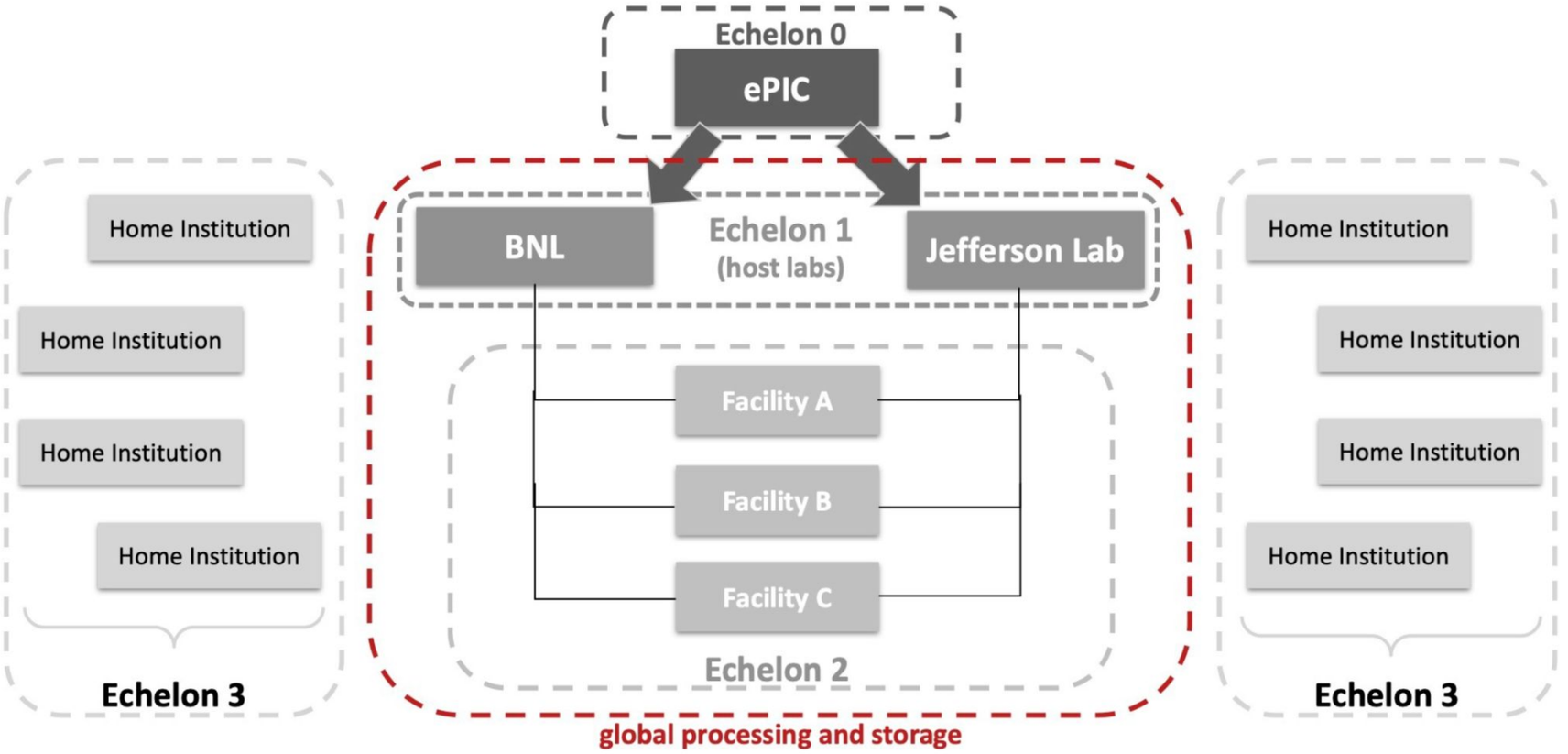
Four tiers:

Echelon 0:
ePIC experiment

Echelon 1:
Two host lab facilities each with full data archive

Echelon 2:
Global processing and data management centers

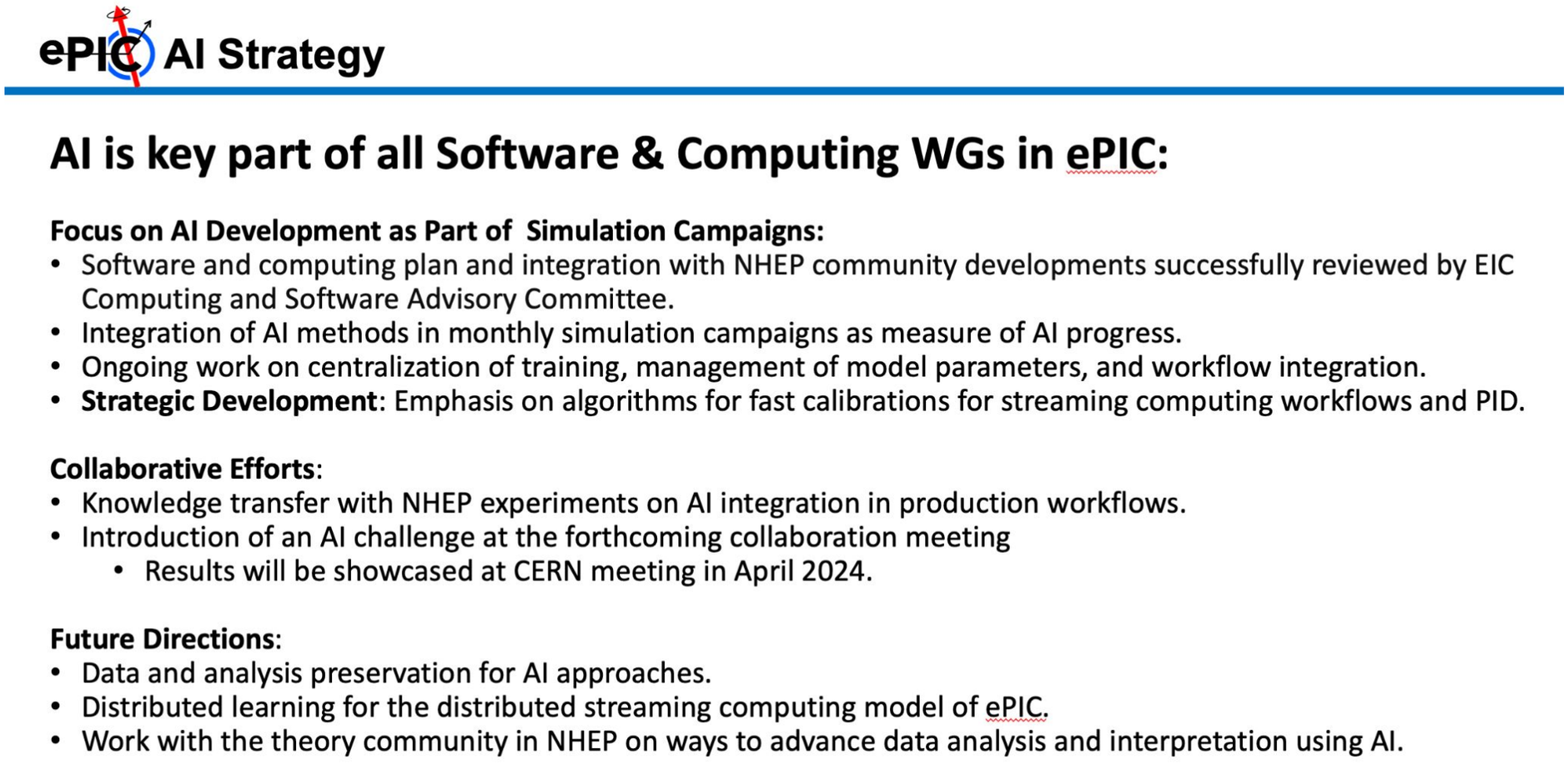
Echelon 3:
Home institutes: ePIC physicists doing analysis



E1 + E2 + E3 seen as a ‘web’ or ‘mesh’ rather than a hierarchy, drawing on LHC experience

ePIC AI Strategy

ePIC Spokesperson John LaJoie at the AI4EIC workshop last November



AI model challenge in Accelerator based Nuclear and Particle Physics

Critical Opportunity:

- Experiments at future accelerators such as the Electron Ion Collider (EIC) will employ data streaming systems that preserve virtually all the data such that a **fully unbiased study** of the data sample, together with accelerator data, can be made.
- These massive datasets, rich in the complex physics embedded within, are an ideal basis to draw on the **techniques of the AI LLM revolution to bring a transformative change** in deriving scientific insights from experimental HENP data.
- **We have the opportunity to build a fully cognizant facility from accelerator to detector and analysis.**
- **BNL is ideal for this work as the only US laboratory hosting multiple user facilities across different science domains, and is home to the only collider in the US**
  - **RHIC, US ATLAS and the EIC comprise some of the largest HENP datasets available today and in the future.**

Expected Impacts:

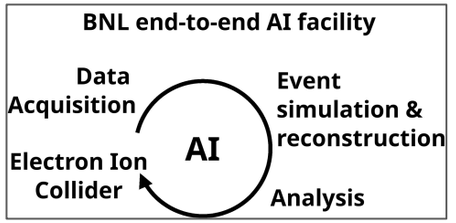
- Experimental HENP complexes such as the EIC are multi billion dollar enterprises that warrant application of the most sophisticated techniques to **maximize their discovery potential, teasing out the greatest possible science return**
- In economic impact,
  - AI-driven efficiency improvements will yield the same data in much less time, **reducing operations and energy costs**.
  - Can reduce the compute and storage intensive demands of HENP analysis, **saving compute and energy costs**.
    - Cost saving example: Improving the signal to background in recorded EIC data by 10% through the use of AI (intelligent DAQ) would lead to a $300k/year saving for archiving media.
  - Techniques developed for accelerators should be **readily transferable to medical and industrial applications**
- An ideal training ground for the nation's AI workforce, ensuring **continued US leadership in this critical new technology**

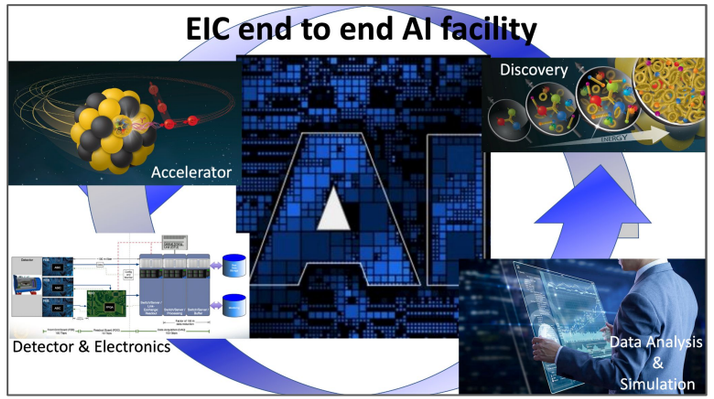
Required R&D:

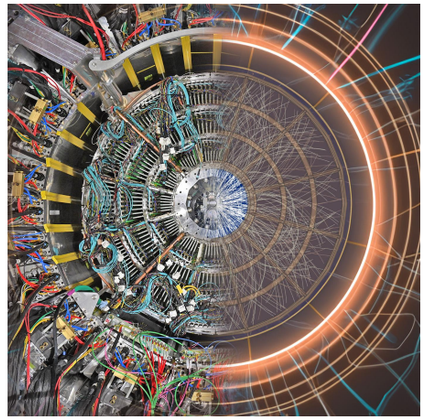
- AI dataset integration of accelerator, experiment and calibration/QA systems, e.g. for real-time optimization of luminosity and background; **optimization in both directions between the machine and the experiment**
- Create a **HENP LLM** (HENP data elements replacing 'words'), and develop training and feature extraction techniques
- Develop techniques to **enhance trustworthiness in building and using AI**

Timeline:

- Near term: 1-3 years
  - Complete R&D on using AI to draw trusted, quantified inferences from real data ***[ie continue the LS4GAN path]***
  - Develop a **prototype HENP LLM with current data** (e.g. ATLAS, RHIC) and first generation feature extraction tools
- Mid term: 3-5 years
  - Begin work on an integrated accelerator - experiment dataset and its AI instrumentation for a cognizant facility from accelerator to detector to analysis, targeting the EIC
- Long term: 5-10 years
  - Documented analyses employing the HENP LLM (internal notes and/or peer reviewed publications)
  - Commission and deploy AI for EIC in parallel with machine and detector installation, commissioning and datataking







From detector to AI in sPHENIX