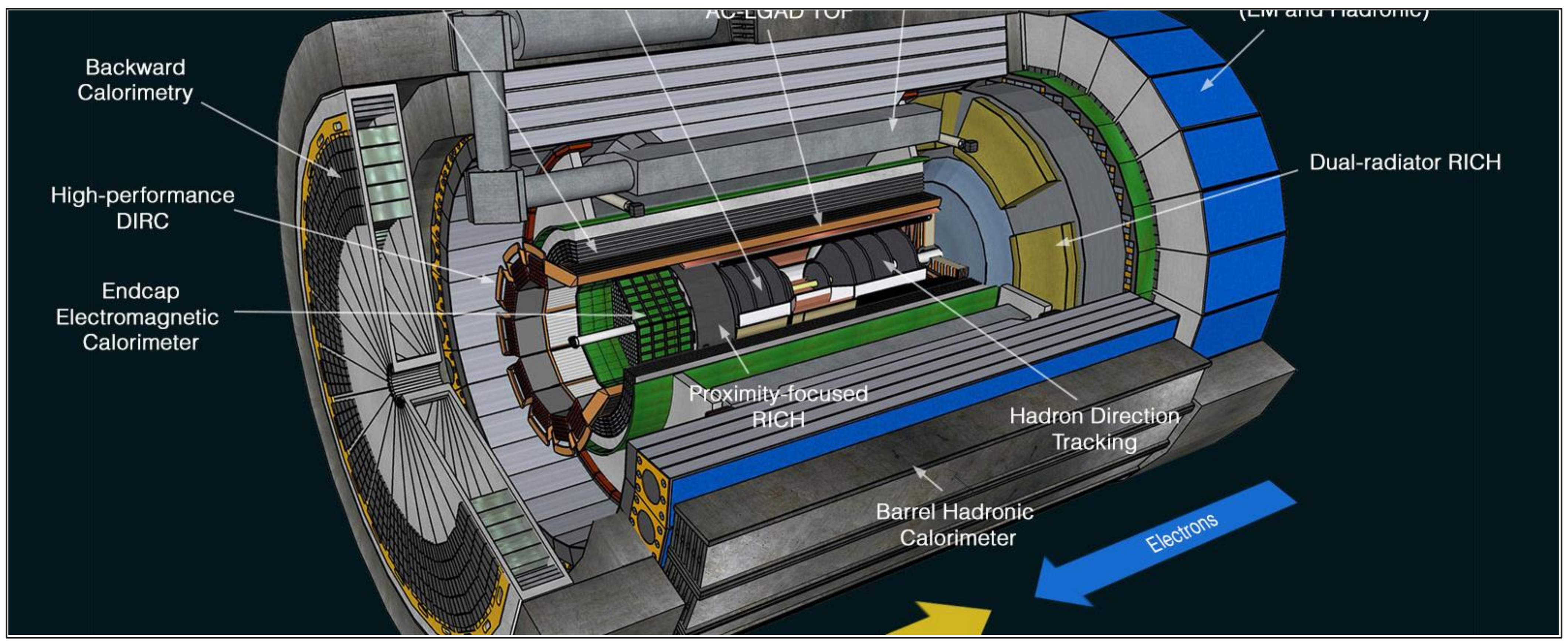
# Agentic AI for Real-Time Expedited Discovery from High-Complexity EIC Data Streams

*Purdue University (Lead) · Fermi National Accelerator Laboratory · Los Alamos National Laboratory · MIT · New Jersey Institute of Technology*

W. Xie (PI, Purdue) · N. Tran, J. Mitrevski (Fermilab) · M. Liu, X. Li (LANL) · P. Harris (MIT) · D. Yu (NJIT) · M. Liu, R. Zhang (Purdue)



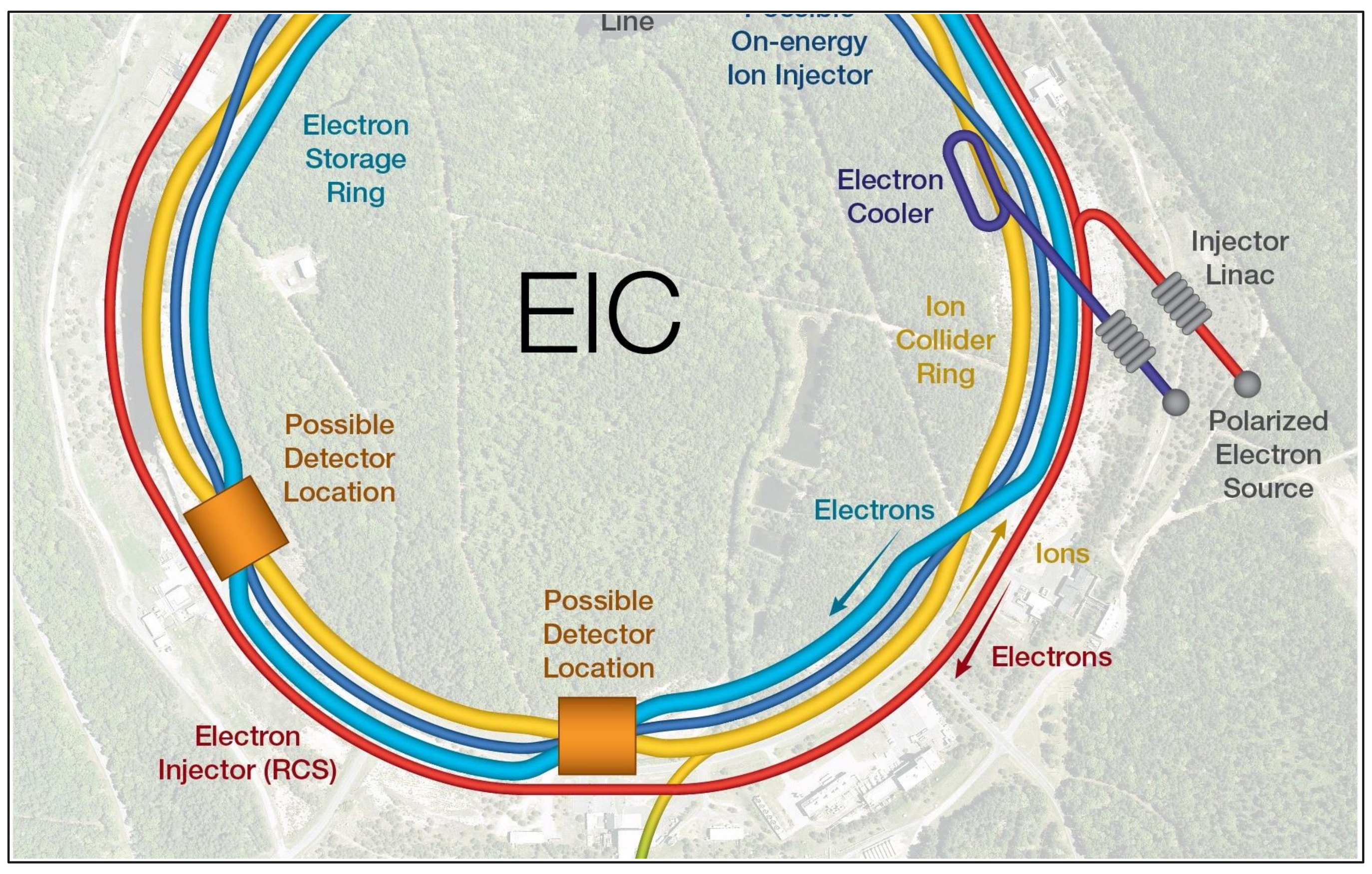




*Fig. 1 — Left: ePIC detector cutaway. Right: the planned EIC accelerator complex at Brookhaven National Laboratory, showing the electron and ion storage rings and the two possible detector locations. The innermost silicon-pixel tracking layers are where synchrotron and beam-gas background first saturate the data stream — the point of intervention targeted by this proposal.*

## Motivation

At the future Electron-Ion Collider, synchrotron and beam-gas backgrounds in the innermost silicon pixel layers will exceed GHz rates — orders of magnitude above the true DIS signal, which sits at hundreds of kHz. Left unmitigated, this saturates readout bandwidth, data movement, storage, and downstream reconstruction, restricting the physics reach of the experiment.

### *Key Question*

*Can hardware-optimized sparse AI models significantly reduce EIC beam-gas backgrounds and thus downstream data volumes while preserving downstream physics performance within the experimental bandwidth constraints? Furthermore, can a bounded agent-assisted workflow measurably shorten the path from model training to an optimized hardware-integrated implementation?*

## Approach

**01** **Benchmarks & evaluation framework:** Establish realistic EIC benchmarks — datasets, baselines, and evaluation protocols — for high-complexity silicon-pixel streams under synchrotron/beam background.

**02** **Sparse representation & low-latency models:** Develop physics-preserving sparse-data models, from sensor level to physics analysis, that suppress or compress background while retaining tracking/vertexing utility.

**03** **Agentic co-optimization & deployment:** Test bounded agent-assisted workflow components and demonstrate measurable AI advantage on one integrated deployment path.

### Sensor-to-Physics Pipeline

| Silicon pixel hits (signal + background) | Sparse chip encoder (set transformer) | Cross-layer GNN (hit-retention scores) | Cleaned hit stream (compressed / filtered) | Tracking & physics (vertexing, analysis) |
|---|---|---|---|---|

*Agentic loop — constrained quantization/compression sweeps · synthesis-report parsing · physics-aware regression testing · feeds back into model & deployment selection.*

## Timeline

**MONTHS 1–2**
Finalize use case, benchmark spec & dataset splits, non-AI baselines.

**MONTH 3**
Benchmark v1 release; GPU/FPGA/hybrid deployment screening.

**MONTHS 4–5**
Train & compare sparse models; down-select primary model.

**MONTH 6**
Integrate agentic pipeline with performance, robustness, deployment profile towards candidate architecture.

### Performance Metrics

**≥10×** background/data-volume reduction (≥2× best non-AI baseline, matched retention)

**≥95%** signal retention, ≤2% downstream performance loss

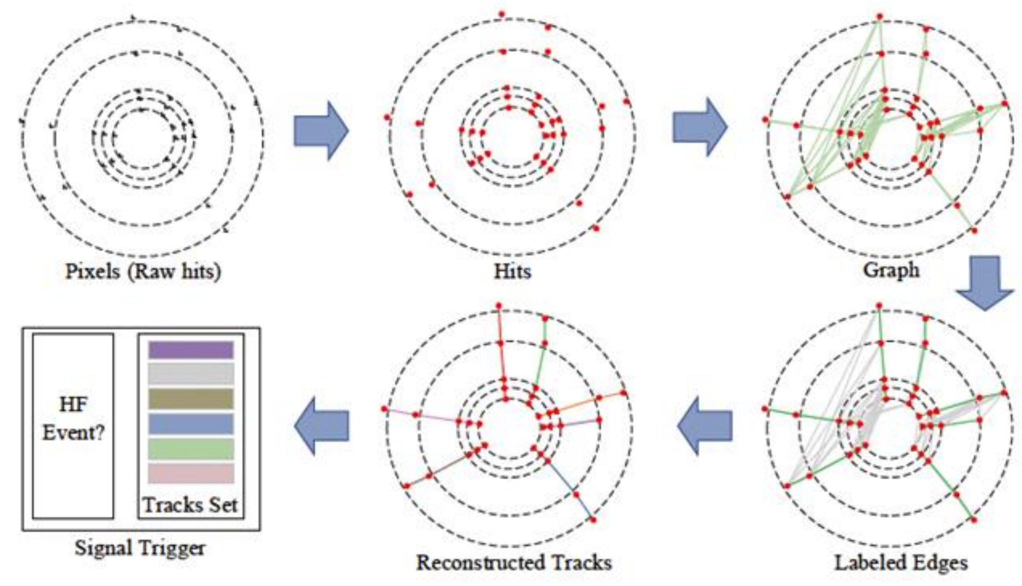
**Sustained** benchmark stream rate within latency & resource budget

**Scaling** advantage over best conventional baseline per unit compute

**≥30%** faster iteration or ≥3× more design points via agentic workflow

## Prior work

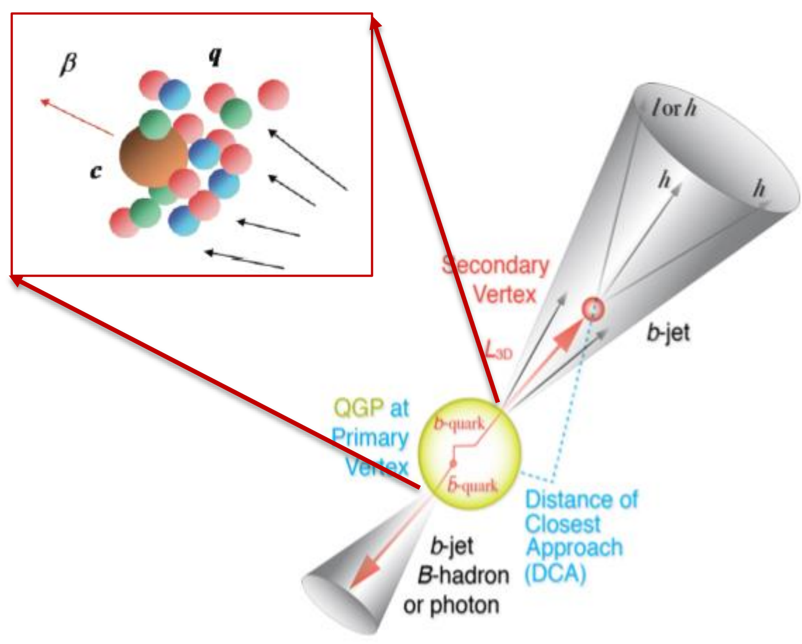
*Conceptual flow for GNN based tracking*



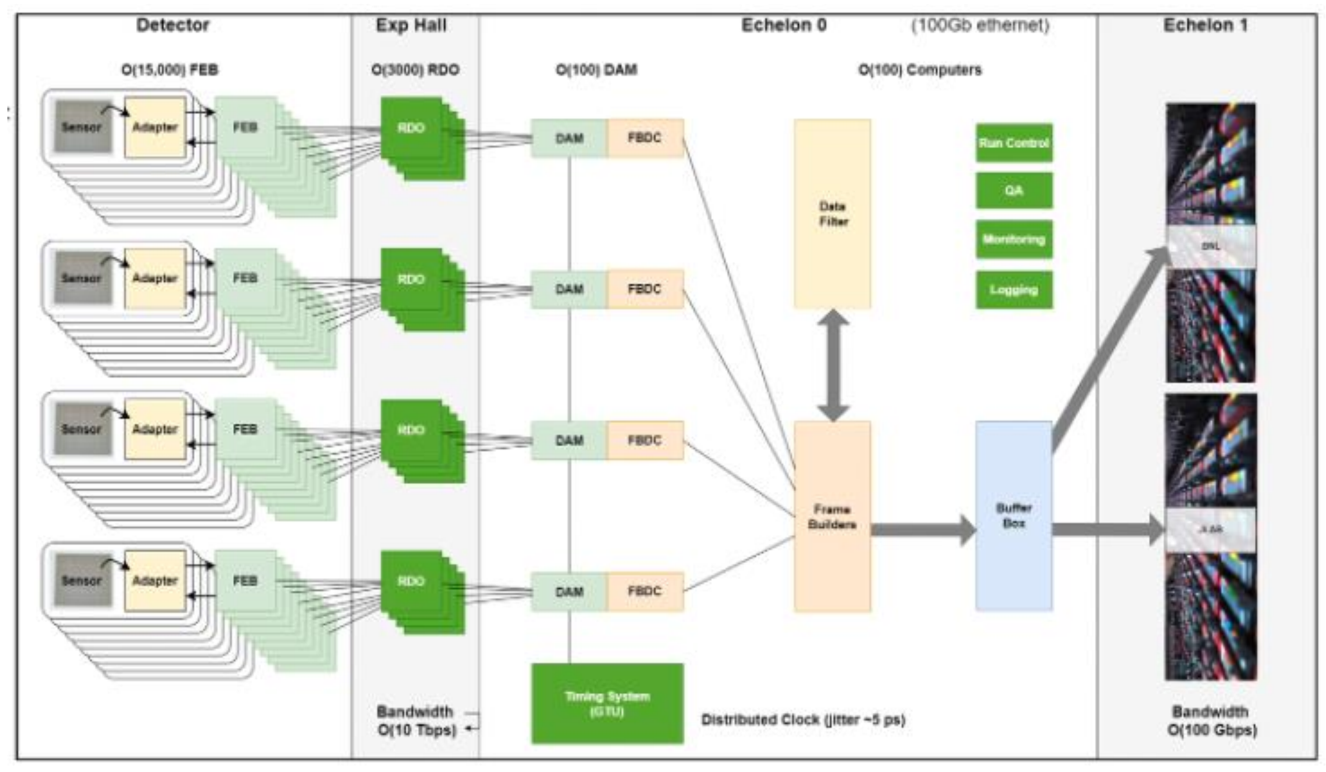

*Initial sPHENIX model comparison*

| | with track radius estimation | | | without track radius estimation | | |
|---|---|---|---|---|---|---|
| Model | # Parameters | Accuracy | AUC | # Parameters | Accuracy | AUC |
| Set Transformer | 299,266 | 86.40% | 91.92% | 298,882 | 72.04% | 78.92% |
| GarNet | 284,210 | 86.22% | 91.81% | 284,066 | 72.59% | 79.61% |
| PN+SAGPool | 780,934 | 86.25% | 92.91% | 780,678 | 69.22% | 77.18% |
| BGN-ST | 363,426 | **87.56%** | **93.22%** | 363.170 | **74.13%** | **81.81%** |

## Experiment Dataflow




*Exploring hardware architectures for streaming AI acceleration*




### Related Work

Kvapil et al., "Intelligent experiments through real-time AI," PoS ICHEP2024 (2025) 1033 [arXiv:2501.04845] — FPGA/hls4ml real-time AI for sPHENIX & EIC, the deployment foundation this proposal extends to sensor-level background suppression.

FERMILAB-POSTER-26-0282-CSAID
This poster has been authored with support of Fermi Forward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

Contact: wxie@purdue.edu