

AI4EIC Activities



Cristiano Fanelli, Tanja Horn

Summer 2025 Joint EICUG/ePIC Collaboration Meeting

AI in Nuclear Physics – Grand Challenges*

❖ Harness the physics program of the Electron Ion Collider (EIC)

- AI/ML will help guarantee maximum science output from the EIC



Next generation QCD experiments like EIC can integrate AI from the start

❖ Realize the science potential of FRIB

- A variety of AI/ML tools will be developed to address specific needs including beam generation, event characterization, detector response, experiment optimization, and data analysis

❖ Event Reconstruction in Nuclear Physics

- AI techniques for reconstruction of tracks in time projection chambers at FRIB, and for heavy ion collisions

❖ Improve Tracking Algorithms

- AI/ML to significantly improve tracking at all NP accelerator facilities

❖ Particle Identification

- AI/ML to complement existing Monte Carlo methods for PID
- Gamma-Ray Energy Tracking Array (GRETA): AI/ML to reconstruct the pat of multiple gamma rays from measured interaction positions and deposited energies

Community activities / perspectives related to EIC

- 2020 ASCAC Subcommittee on AI for Science
- EIC Yellow Report (Nucl. Phys. A **1026** (2022) 12247)
- 2023 NSAC Long-Range Plan
- AI4EIC (C. Allaire, C. Fanelli, ..., T. Horn, et al., Comp. & Softw. For Big Science **8** (2024) 1, 5)

AI/ML is Ubiquitous

Assistance



AI-Powered POS



Entertainment



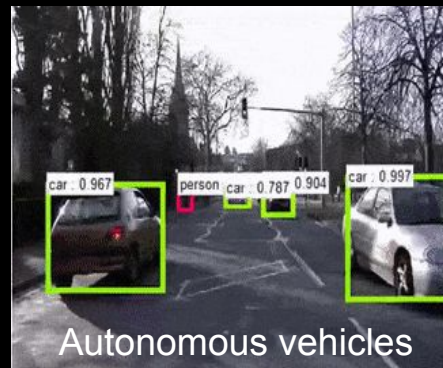
Health



Security



Home automation



Autonomous vehicles

Generative AI

Gen AI is revolutionizing how we work, communicate, and innovate across nearly every field. Can create new content—text, images, code, or even data

The AI4EIC WG

- AI/ML is becoming ubiquitous also in NP. As shown in M. Diefenthaler's report on the ePIC SW & Computing WG, AI is playing a key role in empowering data processing, analysis and enabling autonomous experimentation and control.

Artificial Intelligence

<https://www.eicug.org/content/wg.html>

The mission of the Artificial Intelligence (AI) Working Group is to develop and maintain connections to the data science community, and engaging with the rapidly evolving AI/ML (Machine Learning) toolset that may impact the realization of the EIC science mission. The AI will organize events to educate and assist the EIC community in utilizing AI/ML techniques in the area of detector design and controls, simulations, data readout and analysis, and theory and phenomenology. The AI will also serve as an [entry point to AI applications](#).

- The entry point, [eic.ai](#), is structured in such a way to track all AI4EIC events (workshops, hackathons, tutorials, meetings). Projects and AI/ML software repositories are included in a resource-hub webpage. The website contains also a living review of AI-related papers for the EIC science and also features spotlights on impactful works.
- **AI is a rapidly evolving field**. AI4EIC acts as both a catalyst and an incubator for ideas, demonstrating how cutting-edge AI techniques can be applied to advance EIC science. E.g.,
 - How can generative AI contribute to the EIC science?
 - How do we deal with uncertainty quantification?

Generative AI for EIC: Cherenkov

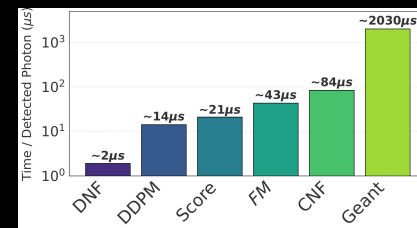
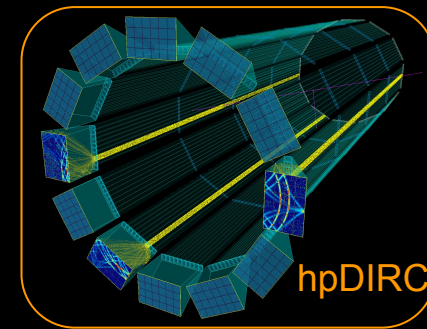
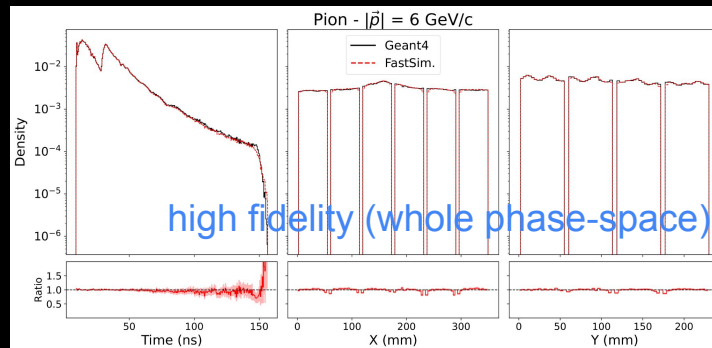
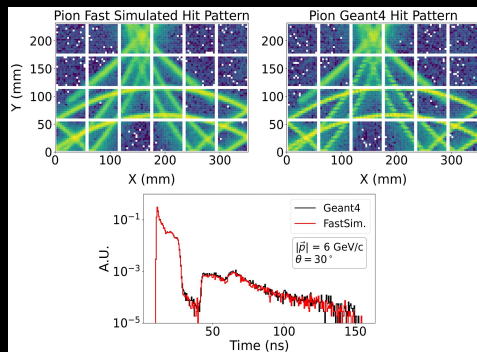
Ongoing collaboration
with the ePIC DIRC WG

1. High-Fidelity Fast Simulation:

Developed generative models capable of producing photon hit distributions with fidelity comparable to Geant4, but at a fraction of the computational cost—critical given the expense of tracking optical photons through complex geometries.

J. Giroux, M. Martinez, CF "Generative Models for Fast Simulation of Cherenkov Detectors at the Electron-Ion Collider." arXiv:2504.19042 (2025). — submitted Machine Learning: Science and Technology [link]

Flow-based



2. Enhanced Particle Identification:

Achieved competitive PID performance across the full detector phase space, with significantly reduced computational time— $O(10\mu\text{s}/\text{track})$ —compared to traditional methods

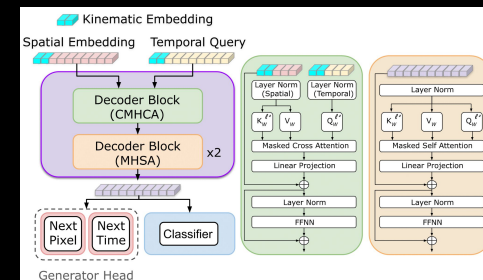
CF, J. Giroux, and J. Stevens. "Deep (er) reconstruction of imaging Cherenkov detectors with swin transformers and normalizing flow models." Machine Learning: Science and Technology 6.1 (2025): 015028. [link]

3. Towards Foundation Models for Cherenkov Detectors:

Recently introduced a unified model architecture capable of performing both reconstruction and fast simulation, enabling simultaneous achievement of (1) and (2) within a single framework.

J. Giroux, CF, "Towards Foundation Models for Experimental Readout Systems Combining Discrete and Continuous Data." arXiv:2505.08736 (2025). [link]

Transformer-based



data as sequences of tokens

<https://github.com/orqs/ai4eic/repositories>

Generative AI for EIC: RAG Agent

What is Retrieval Augmented Generation (RAG)?

- Access up to date information without explicitly training of LLM.
- Reduce “Hallucination” of LLM.
- Grounding LLM to truth to increase reliability by providing citations.

(source code) <https://github.com/ai4eic/EIC-RAG-Project>

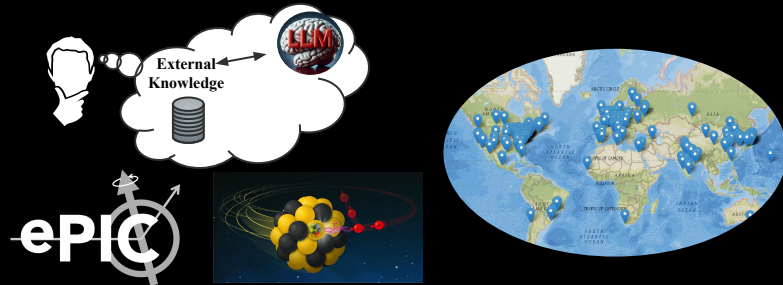
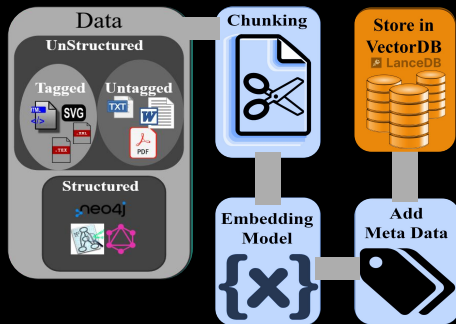
(agent) <https://rogs4eic-ai4eic.streamlit.app/RAG-ChatBot>

Why need RAG for Large Scale Physics Experiments?

- EIC large scale experiment (e.g., EICUG ~1,500 users, ePIC 180 institutions)
- Regular updates to documents, Wiki etc; Tot document size ~ scale of experiment
- Newbies may take months to get to know the full experimental details

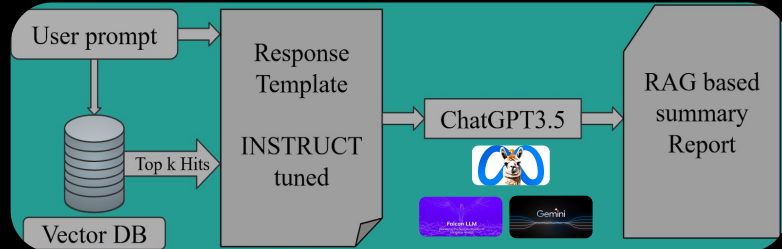
“Ingestion” of data

- Creation of vectorized knowledge base
- Every node below influence RAG performance
- 200+ recent arXiv papers on EIC (since 2021)



“Inference”

- Given a prompt compute similarity index to most similar vectors in VectorDB
- Use LLM to further narrow down and summarize the finding

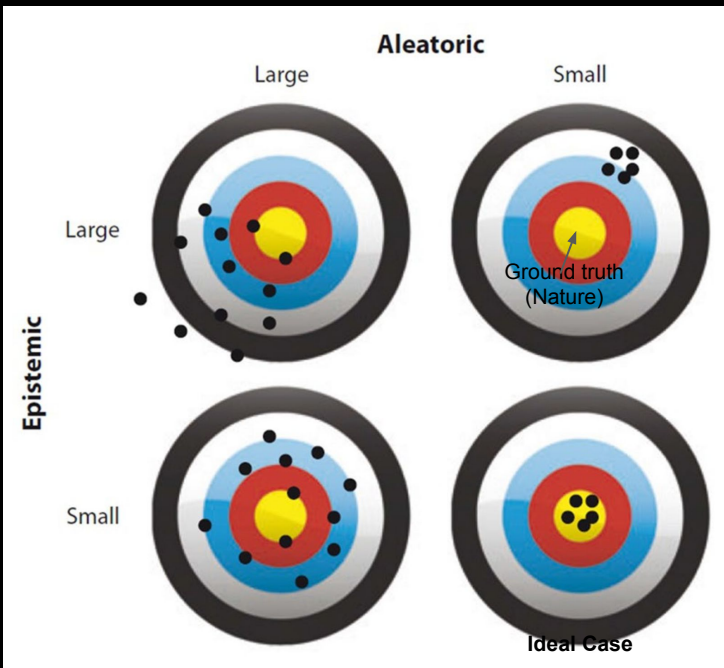


- AI4EIC Collaboration with Ramaiah University, Data Science (India)
 - **B.Sc. thesis by Tina J Jat (supervised by Tapasi Ghosh, in collaboration with AI4EIC)** – This project advances prior OpenAI-based work by developing a fully open-source RAG pipeline for the EIC, using LLaMA 3.2/3.3 and mxbai-embed-large. With a local vector database, it ensures data privacy, avoids external sharing, and remains cost-effective.
- Now EIC RAG works with open source models within olamma. We are working on moving towards agentic workflow
- **Topical Meeting:** a topical meeting will take place between July 30, 2025, 11am EST (<https://indico.bnl.gov/event/28948/>)
- **Workfest@AI4EIC2025:** A workfest on AI agents for EIC will take place during the AI4EIC workshop at MIT, October 27-30, 2025.

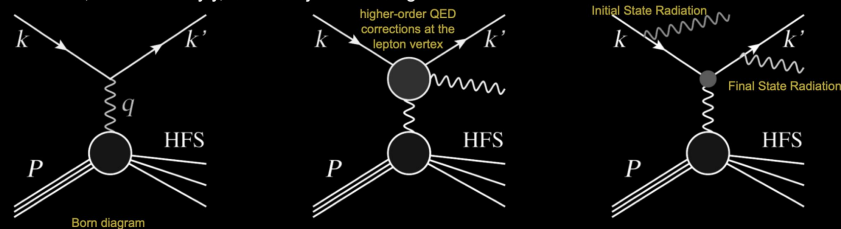
(source code) <https://github.com/ai4eic/EIC-RAG-Project>

(agent) <https://raqs4eic-ai4eic.streamlit.app/RAG-ChatBot>

ML-enabled Kinematics Reconstruction (with UQ)



Deep Inelastic Scattering (DIS) is governed by the four-momentum transfer squared of the exchanged boson Q^2 , the inelasticity y , and the Bjorken scaling variable x .

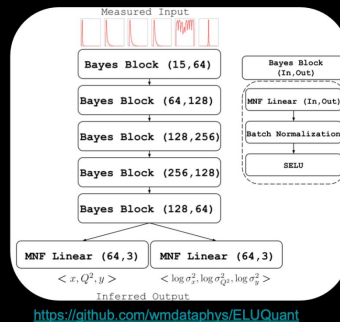


These kinematic variables are related via $Q^2 = s \cdot x \cdot y$, where s is the square of the center-of-mass energy.

$$s = (k + P)^2, \quad Q^2 = -q^2, \quad y = \frac{q \cdot P}{k \cdot P}, \quad \text{and} \quad x = Q^2 / (s y). \quad \text{DIS Kinematics}$$

CF, J. Giroux 2024 Mach. Learn.: Sci. Technol. 5 015017

Event-Level Uncertainty Quantification



<https://github.com/wmdataphys/ELUQuant>

$$\mathcal{L}_{Tot.} = \mathcal{L}_{Reg.} + \gamma \mathcal{L}_{Phys.} + \beta \mathcal{L}_{NF.}$$

Learn the Posterior over the weights

$$\mathcal{L}_{MNF} = \mathbb{E}_{q(\mathbf{W}, \mathbf{x}_T)} [-KL(q(\mathbf{W}|\mathbf{z}_T) \| p(\mathbf{W})) + \log r(\mathbf{z}_T|\mathbf{W}) - \log q(\mathbf{z}_T)]$$

Access epistemic uncertainty through sampling MNF [1] layers

Learn the regression transformation

$$\mathcal{L}_{Reg.} = \frac{1}{N} \sum_i \sum_j \frac{1}{2} (e^{-s_j} \|\mathbf{v}_j - \hat{\mathbf{v}}_j\|^2 + s_j), \quad s_j = \log \sigma_j^2$$

Access aleatoric as a function of regressed output [2]

Constrain the physics

$$\mathcal{L}_{Phys.} = \frac{1}{N} \sum_i \log \hat{Q}_i^2 - (\log s_i + \log \hat{x}_i + \log \hat{y}_i)$$

CF, J. Giroux Mach. Learn.: Sci. Technol. 5 015017 (2024)

See foundational work by:

M. Diefenthaler et al., EPJ C 82.11 (2022): 1064; Arratia et al., NIM-A 1025 (2022): 166164

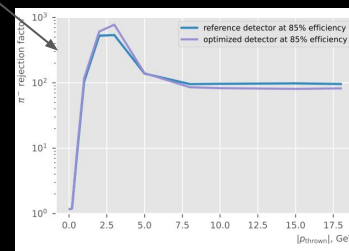
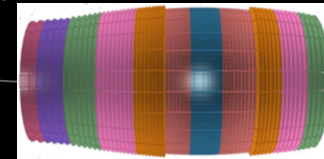
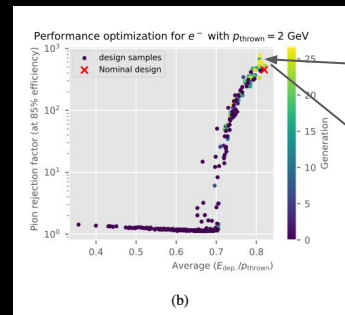
AI for Detector Design

AI for Design used for several sub-detector systems

(right) Performance Optimization for a Scintillating Glass EM Calorimeter for Detector 2

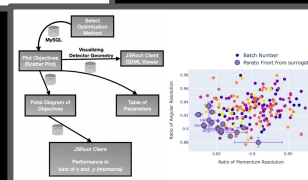
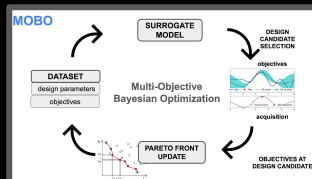
- Beyond pure geometrical optimization
- 2-objectives (multi-objective) optimization
- Genetic algorithms and Bayesian optimization

Performance Optimization for a Scintillating Glass EM Calorimeter for Detector 2

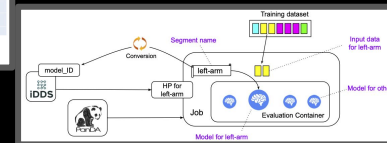


J. Crafts, R. Fatemi, Horn, D. Kalinkin, JINST 19 C05049 (AI4EIC proceeding)

A project to develop a framework for a scalable and distributed AI-assisted detector design and optimization for high complexity problems is also underway



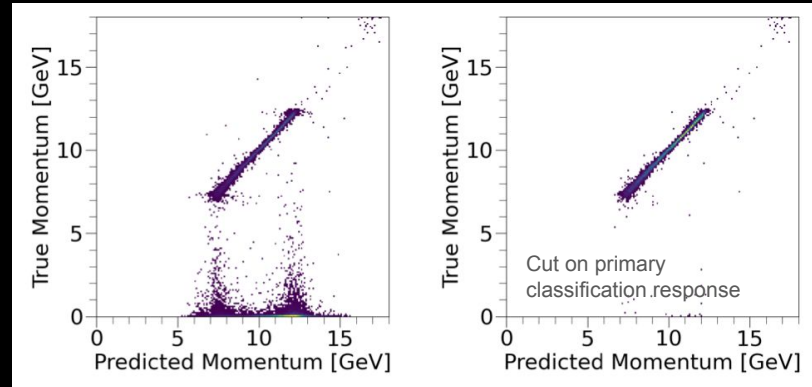
<https://ai4eicdetopt.pythonanywhere.com/>



- Advance SOTA of MOO
- Development of suite of data science tools for interactive navigation of Pareto front
- Leverage and adapt cutting-edge workload management systems

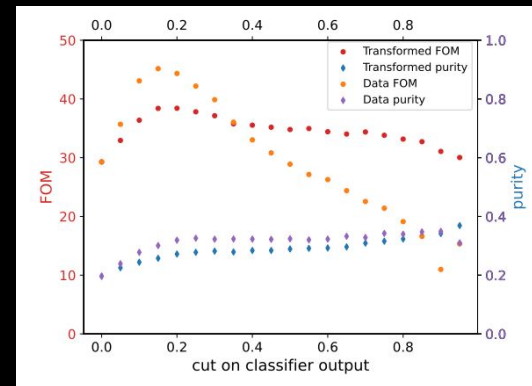
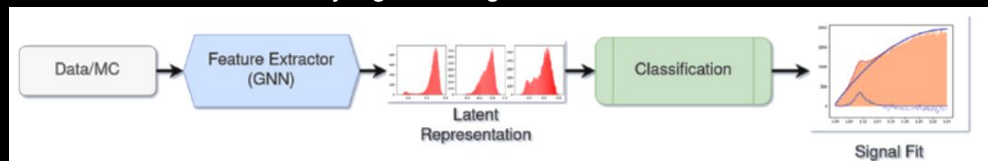
AI for Tracking and Event Classification

- Quasi-real photoproduction measurements involve tracking scattered electrons at small angles relative to the beamline in a backward electron tagger
- Due to high flux of electrons from bremsstrahlung process, the tracking detector systems experience high occupancy
- GNN-based Object Condensation methods use for track building:
 - Avoid “looping” over high order of hit combinations



S. Gardner et al 2024 JINST 19 C05052 (AI4EIC proceeding)

- Normalizing Flows for domain adaptation when identifying Λ hyperon events
- Used GNN to classify signal/background events

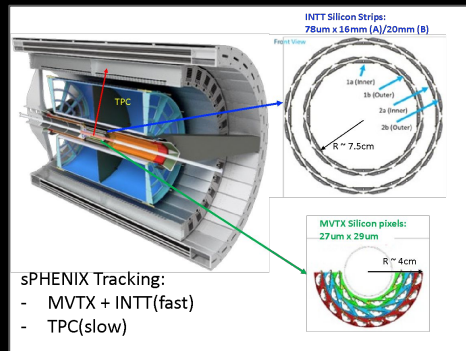


R. Kelleher et al 2024 JINST 19.06 C06020 (AI4EIC proceeding)

AI applications for real time event processing

ML on FPGA

- Fast Data Processing and Autonomous Detector Control for sPHENIX and Future EIC Detectors



Identify D/B hadrons with real-time ML

- Topology of D/B decays
- Monitor collision vertex
- Feedback for improvement

The challenges:

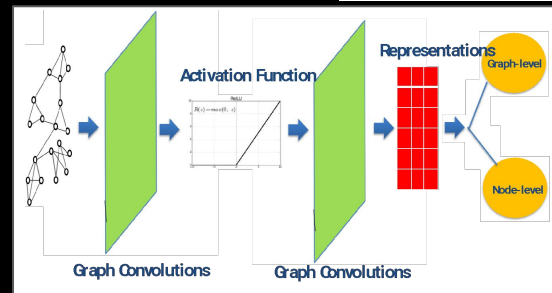
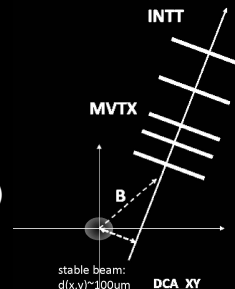
Very high p+p collision rate: ~3MHz

Low rate of rare signals: ~150Hz (beauty for eg)

Limited DAQ trigger bandwidth: ~15 kHz

(or 0.5% of p+p collisions)

No effective conventional triggers available

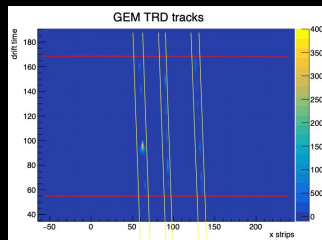


J. Kvapil et al, JINST 19.02 (2024): C02066.

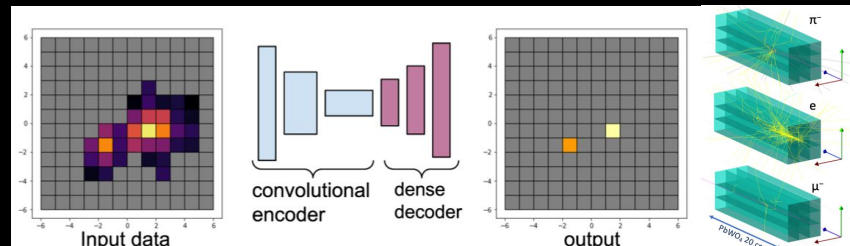
- Machine learning on FPGA for event selection at JLab

AI algorithms on FPGA:

- GNN for pattern recognition
- DNN/LSTM track fitting
- DNN for PID
- VAE for calorimetry



GEM TRD



Calo reconstruction

F. Barbosa et al, IEEE Trans. Nucl. Sci. 70 (2023) 6, 960-965

S. Furlotov, JINST 17 (2022) 06, C06009 (AI4EIC proceeding)

See other contributions at the joint session between SRO XII and AI4EIC, <https://indico.bnl.gov/event/24286/>

DOE Recognition of AI4EIC

- The AI Working Group within the EICUG has gained notable recognition in recent DOE Funding Opportunity Announcements, underscoring the growing impact of our efforts on supporting the future of EIC science with AI/ML
- Highlighted in two consecutive FOAs DE-FOA-0002875 (2023) and DE-FOA-0003845 (2025) both reference the AI4EIC workshops, emphasizing **our role in addressing how AI/ML might contribute to advance the research, design, and operation of the future EIC.**
- Recognition reflects the DOE's alignment with EICUG's vision to integrate AI/ML with experimental data analysis and high-performance computing in nuclear physics.

DE-FOA-0002875 (deadline: Jan 11, 2023; award ceiling \$2M/year)
Artificial Intelligence and Machine Learning For Autonomous Optimization and Control of Accelerators and Detectors

NP held a one-day roundtable on “Machine Learning and Artificial Intelligence for NP Accelerator Facilities” on January 30, 2020, with focus on discussing opportunities in AI/ML for improving efficiencies of accelerator operations of NP facilities. Additional information on the workshop and copies of presentations can be found at: <https://science.osti.gov/np/Research/ai>. An NP community workshop at TJNAF in March of 2020 considered priority research opportunities in AI/ML. Also, an NP community computational science workshop in September 2022 was held to identify future directions in computational nuclear physics by a combination of high-performance computing combined with AI/ML and experimental data. And in October of 2022, an AI for EIC workshop was held to address how AI might contribute to advance research, design and operation of the future EIC. More information on these workshops and meetings can be found in the Reference section below.

DE-FOA-0003845 (deadline: Jan 14, 2025; lab award ceiling \$3.5M/year)
Artificial Intelligence and Machine Learning Applied to Nuclear Science and Technology

2022 identified future directions in computational nuclear physics by a combination of high-performance computing combined with AI/ML and experimental data. And in October 2022 and October 2023, AI for EIC workshops were held to address how AI might contribute to advance research, design and operation of the future EIC. More information on these workshops and meetings can be found in the Reference section below.

These mentions affirm our trajectory and provide momentum for continued cross-disciplinary collaboration and future funding opportunities.

AI4EIC Workshops

average 200+ participants

Past Editions

- 5 workshops organized or co-organized by the AI4EIC WG. In the last 12 months we contributed to organize the joint session with the Streaming Readout at the Streaming Readout Workshop SRO-XII held in Tokyo Dec 2024, and we are currently organizing the 4th AI4EIC workshop that will be held at MIT supported by the Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)

Streaming Readout Workshop SRO-XII +AI4EIC (2024)

Dec 2-4, 2024
University of Tokyo
Asia/Tokyo timezone

Enter your search term

Overview

Timetable

Contribution List

Registration

Participant List

Final announcement

Organization Committee

Zoom connection

Code of Conduct

Travel information

Accommodation

Lunch

Social Event

SRO XI

SRO X

AI4EIC 2023

This meeting brings together DAQ specialist and experimentalist from all over the world, to discuss the learning experience from existing streaming DAQ system and collaborate on future Streaming DAQ system at many facilities and experiments and in particularly the EIC.

This SRO XII edition will be held in Tokyo, Japan, from 12/2 to 12/4. The University of Tokyo will host the workshop.

At this time, we will have a joint session between SRO and AI4EIC to discuss the development and implementation of AI/ML based technologies in the streaming readout and DAQ.

The topics to be discussed in this workshop are:

- streaming DAQ and experiences at many facilities
- real-time calibration and data processing in SRO and heterogeneous computing
- application of AI/ML technologies
- ASICs, FPGAs, Data Aggregation, new challenges for SRO
- establishment of work plans for the future SRO system

We will provide zoom connection to allow the remote participation. However, we encourage in-person participation to have deep discussion.

The in-person registration is closed now.

Remote participation is welcome and please proceed with registration form if you attend remotely.

MIT/IAIFI (2025)

Artificial Intelligence for the Electron Ion Collider (AI4EIC) 2025

Oct 27-29, 2025
Aachen/New York time zone

Enter your search term

Artificial Intelligence for the Electron Ion Collider (2025)

Overview

Call for Abstracts

Registration

Participant List

Scientific Organizing Committee

Local Organizing Committee

Transportation

Lodging

JINST Proceedings AI4EIC2025

Conveners

The 4th AI4EIC Workshop will take place at MIT, Wong Auditorium (Tang Center), in Boston. Organized in collaboration with the AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI), the event will feature in-person participation, with live streaming available for remote attendees.

The previous three AI4EIC workshops each drew an average of 200 participants and fostered meaningful discussions on the full range of AI/ML applications for the EIC—including accelerator and detector design, theory, and analysis—resulting in published proceedings and a community paper (<https://eic.ai>).

Proceedings will be published in the [Journal of Instrumentation](#). The abstract submission (1 page max, figures can be included) is now open.

In this MIT workshop, we will delve deeper into the active and emerging applications of AI/ML within the EIC community, with a focus on ongoing efforts related to the ePIC experiment and beyond.

Scientific Organizing Committee

40 contributions overall, with 1:4 on AI/ML in SRO.

<https://eic.ai/workshops>

CFNS (2021)

Workshop: AI4EIC: Experimental Applications of Artificial Intelligence for the Electron Ion Collider

Dec 1-5, 2021

Artificial Intelligence for the Electron Ion Collider

W&M (2022)

2nd workshop on Artificial Intelligence for the Electron Ion Collider

Nov 18-19, 2022

Workshop: AI4EIC: Experimental Applications of Artificial Intelligence for the Electron Ion Collider

CUA (2023)

AI4EIC 2023 Annual Workshop

November 10, 2023 to November 11, 2023

Workshop: AI4EIC: Experimental Applications of Artificial Intelligence for the Electron Ion Collider

2025 AI4EIC Workshop @MIT/IAIFI

October 27-30, 2025



Scientific Organizing Committee

Abhay Deshpande, Stony Brook, BNL

Ben Nachman, Stanford

Cris Fanelli, W&M

David Lawrence, JLab

Malachi Schram, JLab

Marco Battaglieri, INFN

Mike Williams, MIT

Or Hen, MIT

Phiala Shanahan, MIT

Tanja Horn, CUA

Torre Wenaus, BNL

Conveners

AI/ML for ePIC and Beyond

- Wenliang (Bill) Li (Mississippi State University), Rachel Montgomery (U. of Glasgow), Anselm Vossen (Duke University)

Calibration, Monitoring, and Experimental Control in Streaming Environments

- Yeonju Go (Brookhaven National Lab), Torri Jeske (Jefferson Lab)

AI/ML for Accelerators

- Kevin Brown (Brookhaven National Lab), Brahim Mustapha (Argonne National Lab)

AI/ML for Data Analysis and Theory

- Brandon Kriesten (Argonne National Laboratory), Simonetta Liuti (University of Virginia)

Trends in Data Science

- Philip Harris (Massachusetts Institute of Technology), Yaohang Li (Old Dominion University)

AI/ML in Production, Distributed ML

- Kishan Rajput (Jefferson Lab), Ray Ren (Brookhaven National Lab)

- **2 sessions per day** - AI4EIC 2025 will feature invited talks and selected abstract contributions.
- **Abstract submission** (1 page max, figures can be included) is now open.
- **Proceedings** will be published in the *Journal of Instrumentation*.
- **Workfest**: the workshop will host satellite events such as a focused workfest on AI agents for EIC applications, organized by the AI4EIC working group.

indico.bnl.gov/event/AI4EIC2025

<p>Using machine learning for particle identification in ALICE lukasz Kamil Graczykowski et al 2022 <i>JINST</i> 17 C07016</p> <p>▼ Open abstract View article PDF</p>
<p>Artificial Intelligence for imaging Cherenkov detectors at the EIC C. Fanelli and A. Mahmood 2022 <i>JINST</i> 17 C07011</p> <p>▼ Open abstract View article PDF</p>
<p>High performance FPGA embedded system for machine learning based tracking T. Xuan et al 2022 <i>JINST</i> 17 C07003</p> <p>▼ Open abstract View article PDF</p>
<p>Machine learning on FPGA for event selection S. Furtleov et al 2022 <i>JINST</i> 17 C06009</p> <p>▼ Open abstract View article PDF</p>
<p>Design of detectors at the electron ion collider with artificial intelligence C. Fanelli 2022 <i>JINST</i> 17 C04038</p> <p>▼ Open abstract View article PDF</p>
<p>AI for Experimental Controls at Jefferson Lab T. Jeske et al 2022 <i>JINST</i> 17 C03043</p> <p>▼ Open abstract View article PDF</p>
<p>Frontiers in computing for artificial intelligence T.S. Humble et al 2022 <i>JINST</i> 17 C03037</p> <p>▼ Open abstract View article PDF</p>
<p>Machine learning for track reconstruction at the LHC L.-G. Gagnon 2022 <i>JINST</i> 17 C02026</p> <p>▼ Open abstract View article PDF</p>
<p>Accelerator and detector control for the EIC with machine learning T. Britton and B. Nachman 2022 <i>JINST</i> 17 C02022</p> <p>▼ Open abstract View article PDF</p>
<p>EIC detector overview D.W. Higinbotham 2022 <i>JINST</i> 17 C02018</p> <p>▼ Open abstract View article PDF</p>

<p>R. Kelleher, A. Vossen Normalizing flows for domain adaptation when identifying A hyperon events (AI4EIC proceeding, <i>JINST</i> 19 C06020)</p>
<p>G. Matousek, A. Vossen Photon classification with Gradient Boosted Trees at CLAS12 (AI4EIC proceeding, <i>JINST</i> 19 C06006)</p>
<p>S. Gardner, R. Tyson, D. Glazier and K. Livingston Object condensation for track building in a backward electron tagger at the EIC (AI4EIC proceeding, <i>JINST</i> 19 C05052)</p>
<p>G. Gavalion on behalf of the CLAS12 Collaboration Real-time charged track reconstruction for CLAS12 (AI4EIC proceeding, <i>JINST</i> 19 C05050)</p>
<p>J. Craft et al Performance optimization for a scintillating glass electromagnetic calorimeter at the EIC (AI4EIC proceeding, <i>JINST</i> 19 C05049)</p>
<p>K. Suresh et al Towards a RAG-based Summarization Agent for the Electron-Ion Collider (AI4EIC proceeding, <i>JINST</i> 19 C07006)</p>
<p>C. Fanelli et al Physics Event Classification Using Large Language Models (AI4EIC proceeding, <i>JINST</i> 19 C07011)</p>
<p>M. Diefenthaler et al AI-Assisted Detector Design for the EIC (AID(2)E) (AI4EIC proceeding, <i>JINST</i> 19 C07001 2024)</p>
<p>T. Jeske et al. Hydra: Computer Vision for Data Quality Monitoring (AI4EIC proceeding, <i>JINST</i> 19 C12005 2024)</p>
<p>D. Lawrence et al. ML-based Calibration and Control of the GlueX Central Drift Chamber (AI4EIC proceeding, <i>JINST</i> 19 C11012 2024)</p>
<p>M. Karwowska et al. Particle identification with machine learning from incomplete data in the ALICE experiment (AI4EIC proceeding, <i>JINST</i> 19 C07013 2024)</p>

Artificial Intelligence for the Electron Ion Collider (AI4EIC)

Review | Open access | Published: 15 February 2024

Volume 8, article number 5, (2024) [Cite this article](#)

Download PDF

✓ You have full access to this [open access](#) article

C. Allaire, R. Ammendola, E.-C. Aschenauer, M. Balandat, M. Battaglieri, J. Bernauer, M. Bondi, N. Branson, T. Britton, A. Butter, I. Chahrouh, P. Chatagnon, E. Cisbani, E. W. Cline, S. Dash, C. Dean, W. Deconinck, A. Deshpande, M. Diefenthaler, R. Ent, C. Fanelli , M. Finger, M. Finger Jr., E. Fol, ... P. Zurita [Show authors](#)

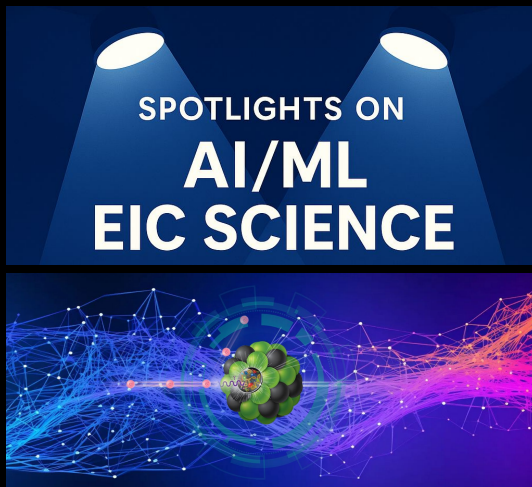
14 citations (INSPIRE-HEP)

AI4EIC 2025 Abstract submission now open
(1 page max, figures can be included)

<https://indico.bnl.gov/event/28082/abstracts/>

Spotlight on AI/ML for EIC Science

- This spotlight features a curated selection of the most influential papers per year that showcase cutting-edge applications of AI and ML to EIC science.
- These works represent milestones in the integration of modern computational techniques into experimental and theoretical efforts within the EIC community. We highlight these papers for their innovation, impact, and potential to guide future research in EIC.
- A sub-committee within AI4EIC is appointed to select the spotlights



Full list can be found here:
<https://eic.ai/spotlight>

ARTIFICIAL INTELLIGENCE FOR THE ELECTRON ION COLLIDER

PAPER SPOTLIGHT

2025

J. Y. Araz, V. Mikuni, F. Ringer, N. Sato, F. Torales Acosta, R. Whitehill

PHYSICS LETTERS B (2025): 139694.

Point cloud-based diffusion models for the Electron-Ion Collider

This [paper](#) uses machine learning to advance full event generation at the EIC through a novel diffusion model that combines point cloud representation with transformer modules to generate all particle species while preserving physical conservation laws. The ML-based approach significantly outperforms previous generative models across all evaluation metrics and demonstrates potential for broader applications in collider phenomenology

R. Milton, S. J. Paul, B. Schmookler et al.

NIM-A, 170613

Design and simulation of a SiPM-on-tile ZDC for the future EIC, and its performance with graph neural networks

This [NIM A article](#) presents a novel Zero-Degree Calorimeter design and uses Graph Neural Networks to dramatically improve its energy/angle resolution. The GNN-based reconstruction meets or exceeds the stringent requirements from the EIC Yellow Report, showcasing how cutting-edge deep learning can optimize calorimeter performance even with complex detector geometries (an important

Impact of AI4EIC Initiatives: Examples

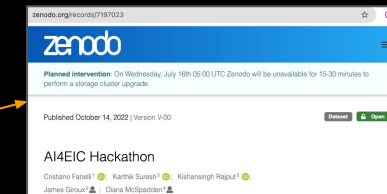
The seeds planted through AI4EIC initiatives continue to bear fruit over the years, giving rise to follow-up research and publications

AI4EIC hackathon
(ML for Cherenkov
detectors)

(2022)

dataset

CHEP
proceeding



AI4EIC Hackathon: PID with the ePIC dRICH

Cristiano Fanelli^{1,2,,} James Giroux³, Diana McSpadden⁴, Kishansingh Rajput⁵, Karthik Suresh⁶, Evaristo Cibusa⁷, Wouter Deconinck⁸, Eric Walter⁹, Andrea Bressan², Markus Diefenthaler², Tanja Horn², and Torrey Wenaus¹*

¹Brookhaven National Lab, Upton, NY, 11973, USA
²Catholic University of America, Washington, DC, 20064, USA
³INFN Trieste and University of Trieste, Department of Physics, Trieste, 34127, Italy
⁴Istituto Superiore di Sanità, INFN, viale Regina Elena 299, Rome, 00161, Italy
⁵Thomas Jefferson National Accelerator Center, Newport News, VA, 23606, USA
⁶University of Manitoba, Physics, Winnipeg, MB, R3T 2N6, Canada
⁷William and Mary, Williamsburg, VA, 23185, USA

(2024)

Inspired development of
ML algorithms for
Cherenkov detectors

(2025)

[Mach. Learn.: Sci. Technol. 6 015028](#)

AI4EIC hackathon
(LLM for EIC
Science)

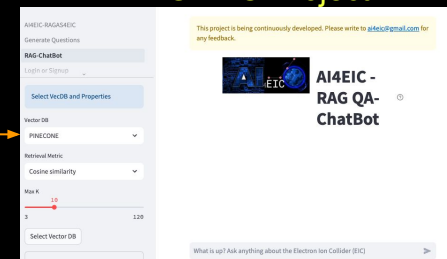
(2023)

JINST
proceeding



(2024)

RAG4EIC Project



JINST 19.07 (2024): C07006.
<https://github.com/ai4eic/EIC-RAG-Project>

Summary

- **Active Engagement and Events**

- Overall organized four AI4EIC workshop (+tutorials and hackathons), plus a joint event with Streaming Readout in December 2024; and also contributing to joint workshops with APS DNP
- AI4EIC workshops have become a pivotal venue in the NP community, recognized by federal agencies in their Funding Opportunity Announcements.
- A workshop at MIT/IAIFI is upcoming (October 27-30, 2025) <http://indico.bnl.gov/event/AI4EIC2025>

- **Successful Hackathon Events**

- AI4EIC Hackathons have achieved remarkable success, fostering innovation and collaboration and new research papers. Software infrastructure developed by AI4EIC utilized by ePIC hackathon.

- **Promoting AI/ML Literacy**

- Continued efforts to enhance AI/ML literacy within the EIC community, resulting in:
 - Editor of 23 JINST proceedings for the first and third workshops. More proceedings to come this year.
 - 2 hackathons documented in peer-reviewed papers.
 - 1 community paper on Computing and Software for Big Science published in 2024 with 14 citations.
 - Other AI4EIC contributions: APS DNP workshops; lecture series for summer school students (HUGS)

- **Topical Meetings and Innovative Projects**

- Covering cutting-edge topics and projects actively involving the AI4EIC community.
 - See, e.g., the repository on RAG-based summarization and upcoming meeting on July 30.

<https://eic.ai/>

Backup

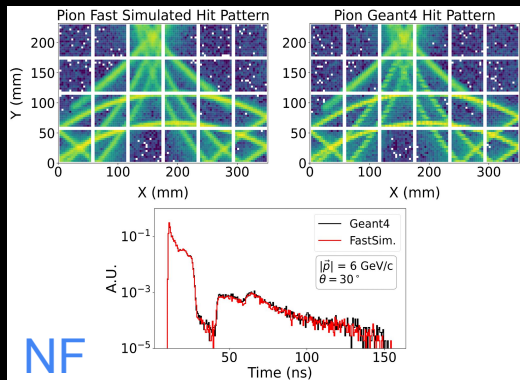
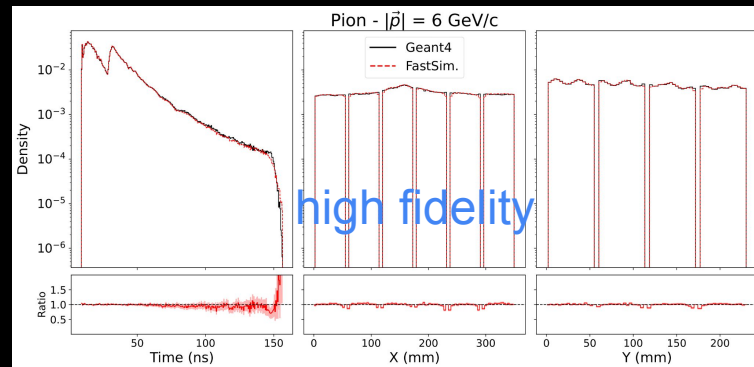


Fast Simulation at EIC – hpDIRC

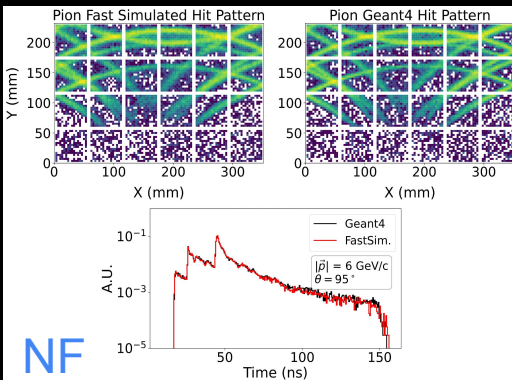
Architectures: Normalizing Flows (NF), Continuous Normalizing Flows (CNF), Conditional Flow Matching (CFM), Denoising Diffusion Probabilistic Models (DDPM), Score Based Generative Models (SB)

- **Suite of SOTA Generative Models** – Compare modern SOTA generative algorithms in the space of DIRC simulation
- **Hit-Level Learning** – Model conditioned on kinematic parameters ($|p|, \theta$)
- **Agnostic to Photon Yield** – Ensure model independence from photon yield
- **Abstract away Fixed Input Size** – Address limitations with discrete distributions; data preprocessing transform DIRC readout (row, col) to (x,y) in mm and uniformly smear over PMT pixels

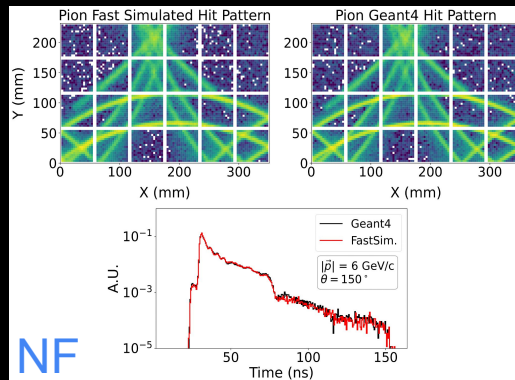
J. Giroux, James, M. Martinez, and CF. "Generative Models for Fast Simulation of Cherenkov Detectors at the Electron-Ion Collider."
arXiv:2504.19042 (2025).



NF



NF



NF

Simulation is fast - $O(0.5)\mu\text{s}$ per hit (effective)

(hpDIRC standalone sim)

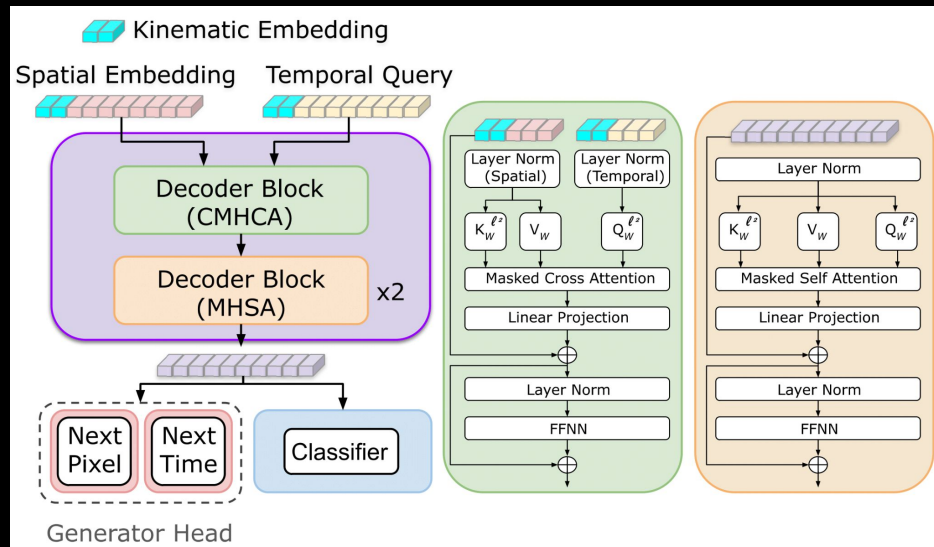
Foundation Models - hpDIRC

- Foundation Models capable of generalizing to multiple task
 - Pre-trained backbone structure (transformer based)
- *Fine-tune* to different tasks
 - Generation
 - Classification
 - Noise Filtering
- Represent hits in *tokenized* space

spatial $\rightarrow \{|\vec{p}|, \theta, \text{SOS}_p, p_1, \dots, p_n, \text{EOS}_p\}$

time $\rightarrow \{|\vec{p}|, \theta, \text{SOS}_t, t_1, \dots, t_n, \text{EOS}_t\}$

J. Giroux and C Fanelli "Towards Foundation Models for Experimental Readout Systems Combining Discrete and Continuous Data." *arXiv:2505.08736* (2025).

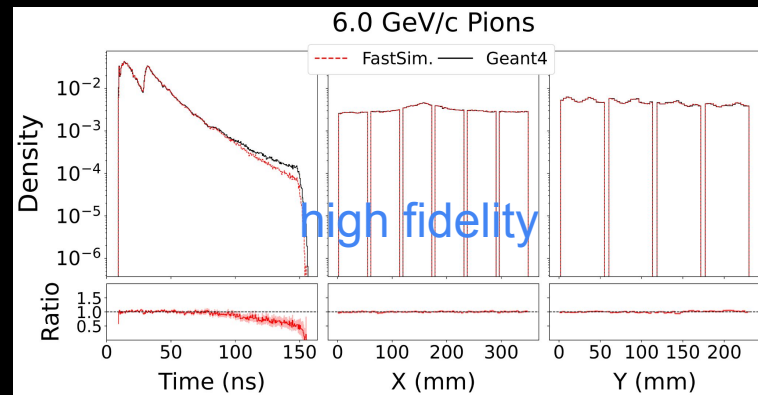


[Github](#)

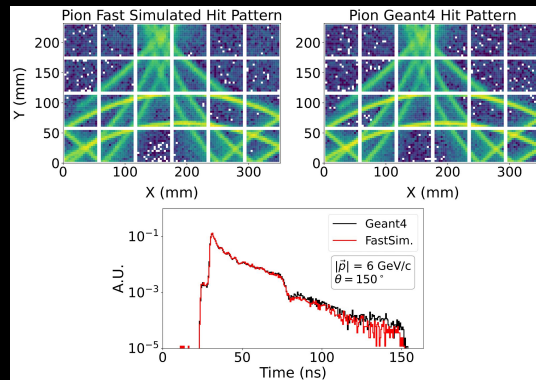
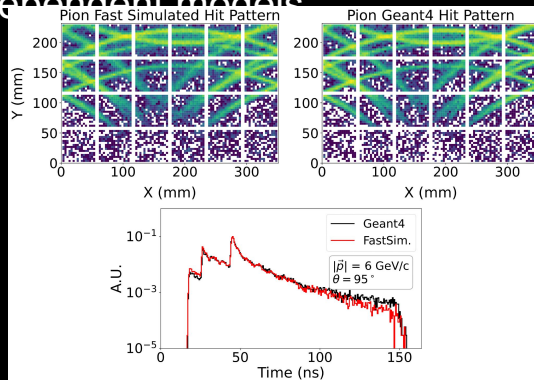
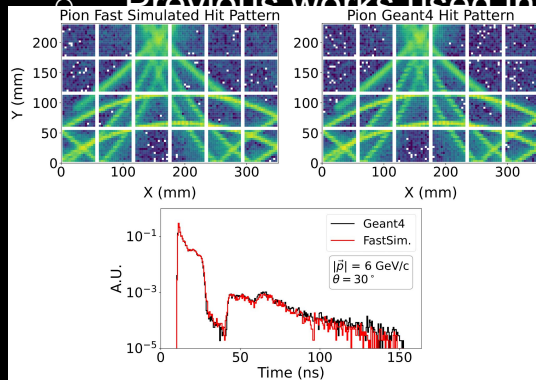
All code is open source and pre-trained models are provided.

Foundation Models - hpDIRC

- Fast simulation through *next token* prediction
- Directly learns variability in photon yield
 - Model conditioned on kinematic parameters ($|p|$, θ)
 - No external modeling of photon yield required
- Class conditional (particle type) generation through a *Mixture of Experts* (MoE)



Previous works used independent models

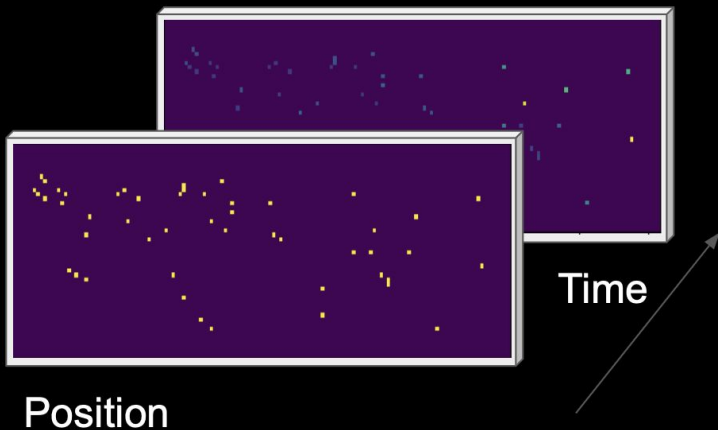


Simulation is fast - $O(0.02)s$ per track (effective)

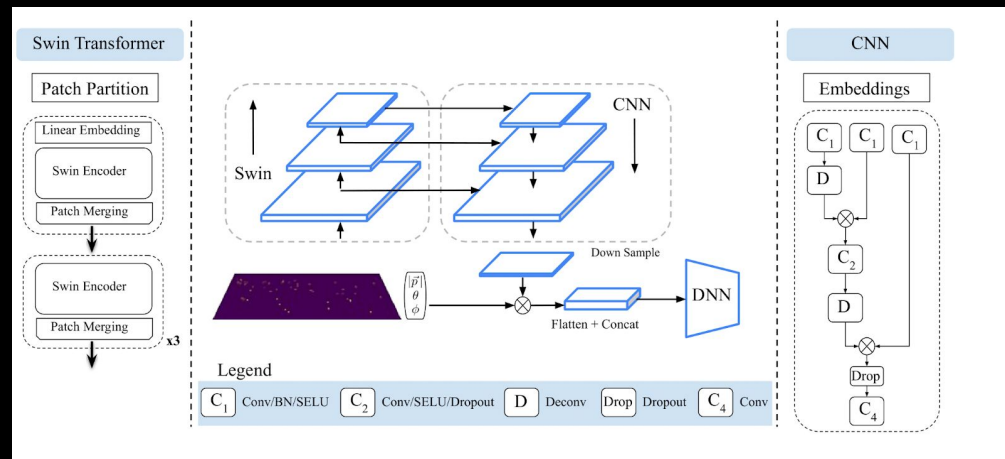
(hpDIRC standalone sim)

Deep(er)RICH: Transformer-based PID

CF, J. Giroux, J. Stevens. "Deep(er)RICH"
Machine Learning: Science and Technology 6.1 (2025): 015028.



- Individual tracks do form “images” in optical boxes
 - Sparse point representations
- Possibility of overlapping hits
 - Same x,y - different times
 - Construct these as images as FIFO
 - Tends to be low percentage of overlap



- Hierarchical Vision Transformer (Swin) - encoder style feature extraction
 - Windowed attention - higher throughput
- Combine information through CNN - utilize skip connections for different resolutions
- Inject kinematics as concatenated information to DNN

Takeaways

- Simulation

- Order of magnitude faster than Geant4 - we have shown our algorithms (not FM) run extremely well on CPU
- Simulation is easily usable by users without GPU - PID should use GPU for efficiency
- Possibility to enable time-imaging - we can simulate PDF's on the fly on GPU

- PID

- Increased performance shown at GlueX
- Also shown increased performance at hpDIRC (very preliminary)
- Compute wise - Geometric LUT is cheap - but so are we on GPU - mainly depends if we can outperform

- Foundation Model

- Everything under one architecture - bulk of model remains identical - changes in final layer
- More computationally intensive than previous models for simulation - requires GPU
- PID is still very cheap and fast
- Possibility for noise filtering - possibility of applications to high rate environments such as in the dRICH

RAG4EIC

Methods to evaluate RAG's performance

- [RAG Assessment Score \(RAGAS\)](#)
- LLM as Judge^[1]

1. Faithfulness –

$$\text{Faithfulness score} = \frac{|\text{Number of claims that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

2. Answer relevance –

$$\text{Answer relevance} = \frac{|\text{\#claims correctly responded}|}{|\text{\#claims in given text}|}$$

3. Context relevance –

$$\text{Context relevance} = \frac{|\text{\# relevant unique refs}|}{|\text{\# refs retrieved}|}$$

Methods to improve RAG architecture

- Better chunking strategies. LaTeX Splitter, TWikiSplitter
- Metadata based filtering.
- Response Template fine-tuning. INSTRUCT tuning
- Model fine-tuning. Computationally costly.

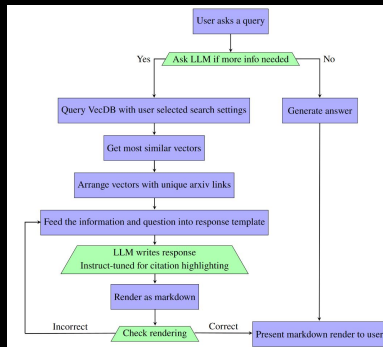
- Towards Multimodal Agents: A multimodal AI assist in interpreting visual data, cross-referencing documentation, providing actionable insights in real time...

Question What is the pixel size and expected momentum resolution at 2GeV from a MAPS ITS3 tracking layer

Response

- The pixel size of the MAPS ITS3 tracking layer is 10 μm
<ARXIV_ID><http://arxiv.org/abs/2108.08314v1> </ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2203.07626v2> </ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2302.01447v1> </ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2305.15593v1> </ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2210.08609v2> </ARXIV_ID>
- The expected momentum resolution at 2 GeV from a MAPS ITS3 tracking layer is 2.5% to 15% of the initial momentum <ARXIV_ID><http://arxiv.org/abs/2305.15593v1> </ARXIV_ID>
<ARXIV_ID><http://arxiv.org/abs/2210.08609v2> </ARXIV_ID>

Faithfulness	Answer relevance	Context relevance
100%	50%	70%



Question What are the dimensions of MAPS detector used in EIC?

Response

- The MAPS detector used in the EIC has dimensions ranging from 3.67 cm to 43 cm, with a sensor size of 10x10 cm <ARXIV_ID><http://arxiv.org/abs/2203.07626v2>, <http://arxiv.org/abs/2302.01447v1>, <http://arxiv.org/abs/2210.04879v2>, <http://arxiv.org/abs/2305.02079v2>, <http://arxiv.org/abs/2302.01447v1>
- The ePIC MAPS electron endcap detector has a geometry parameter listed, and the ePIC MAPS hadron endcap detector also has geometry parameters listed <ARXIV_ID>
<http://arxiv.org/abs/2203.07626v2>, <http://arxiv.org/abs/2305.15593v1>, <http://arxiv.org/abs/2305.15593v1>, <http://arxiv.org/abs/2305.15593v1>, <http://arxiv.org/abs/2305.15593v1>
- The ePIC MAPS vertex detector design can reach as low as 0.05, and the imaging layers of MAPS sensors provide excellent position resolution <ARXIV_ID>
<http://arxiv.org/abs/2203.07626v2>, <http://arxiv.org/abs/2305.15593v1>
- The MAPS sensor used in the EIC is based on low-mass, low-power consumption MAPS in 65 nm technology <ARXIV_ID><http://arxiv.org/abs/2203.07626v2>
- The MAPS sensor used in the EIC is based on low-mass, low-power consumption MAPS in 65 nm technology <ARXIV_ID><http://arxiv.org/abs/2203.07626v2>

# of Hallucination	Faithfulness	Answer relevance	Context relevance	Answer correctness
3/50	26%	62%	100%	Qualitative

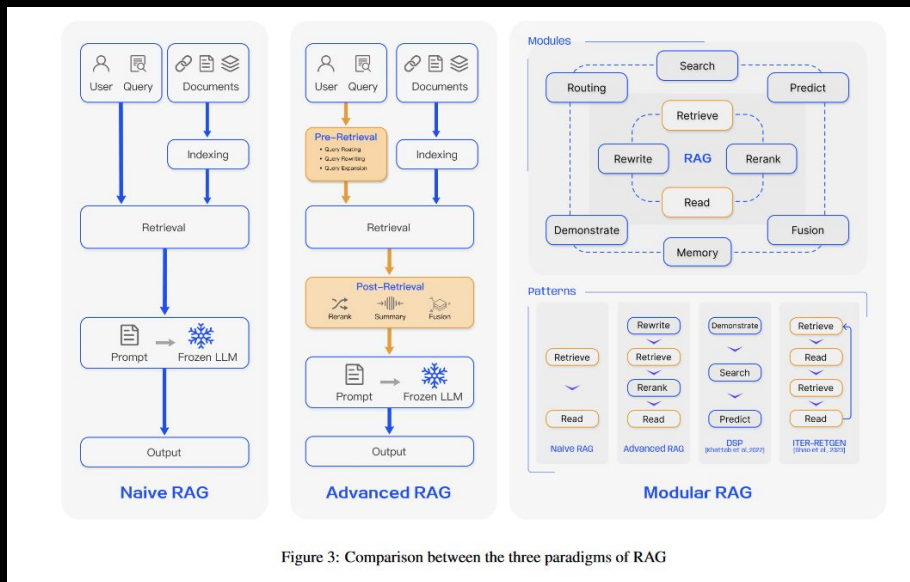
Various RAG methods

- **Naive RAG:**

- Simple & Direct process
- No refinement or optimization
- High throughput but accurate for complex queries.
- Use Case: Straight forward Q&A like shift info or look up tables

- **Advanced RAG:**

- Adds Pre- and Post Retrieval accuracy (RAG4EIC)
- Pre Retrieval contains Routing logic, Query rewriting and Query expansion
- Post Retrieval contains Reranking, summary and fusion strategy for cohesive response
- Use Case: Document Databases



- **Modular RAG**

- Highly flexible and customizable with multiple interchangeable modules.
- Enhanced by **Graph RAG** for structured, relation-aware retrieval.
- Introduces the idea of Agentic workflows
- Use Case: Complex reasoning and decision making tasks.

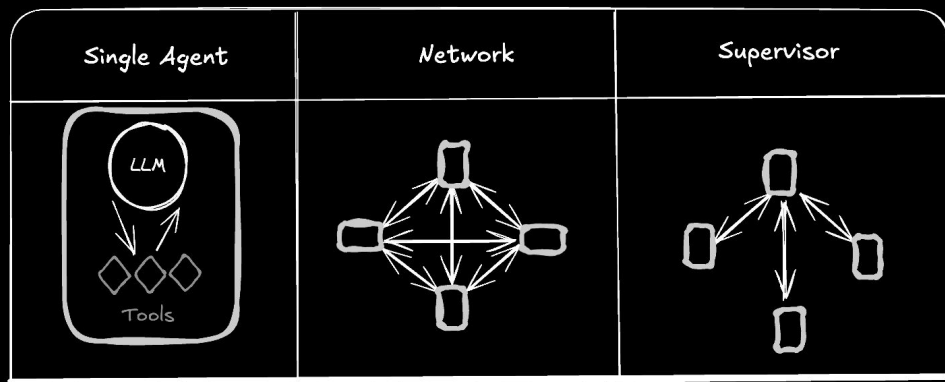
Better than RAG... (Agentic RAG)

- **Single Agent**

- A single LLM agent orchestrates tasks and leverages external tools for assistance.
- Similar to Modular RAG where multiple tools can be coupled to a Single LLM.

- **Network**

- Multiple peer agents (Fine-tuned LLM) can communicate and collaborate using shared/specific tools. The entire workflow is usually optimized through RL
- However, need to have specialized agents and model I/O bottleneck (with poor resource). Since any agent can collaborate with any other agent
- Use case: During data taking agents monitor different sub detector systems and share insights in real time to raise alarm or report on data quality.

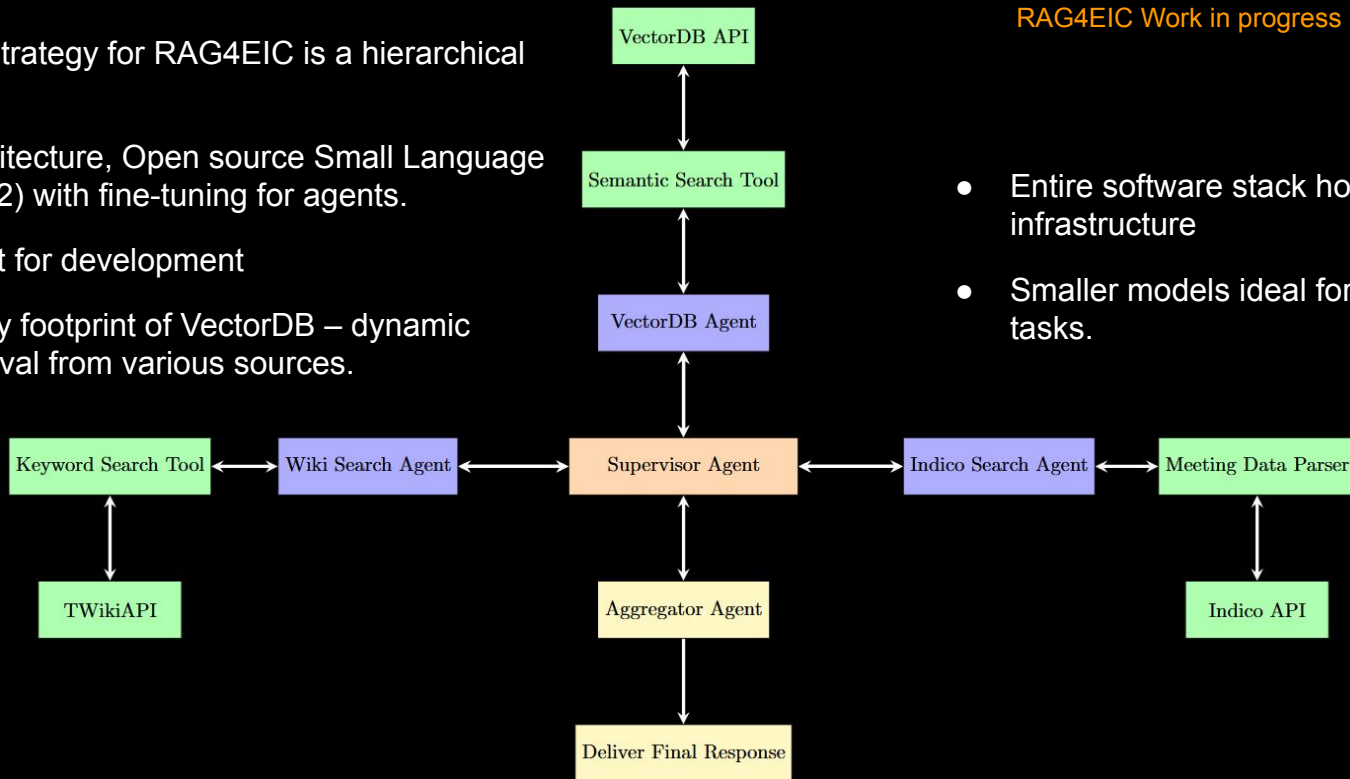


- **Supervisor Agent**

- A central supervisor agent oversees and delegates tasks to specialized agent.
- Optimizes workflow, with minimal dependency when adding new agents
- Lower I/O bottleneck compared to Network setups but have higher task latency
- Use case: Coordinating tasks like beam alignment and detector status.

Better than RAG.. (Agentic RAG)

- The envisioned strategy for RAG4EIC is a hierarchical agent
- Hierarchical architecture, Open source Small Language Models (Llama3.2) with fine-tuning for agents.
- Langchain Toolkit for development
- Reduced Memory footprint of VectorDB – dynamic information retrieval from various sources.



RAG4EIC Work in progress (K. Suresh et al)

- Entire software stack housed within the infrastructure
- Smaller models ideal for specialized tasks.

Also an Ongoing HEP Effort

chATLAS

An AI Assistant for the
ATLAS Collaboration

DANIEL MURNANE, GABRIEL FACINI,
RUNZE LI & CARY RANDAZZO

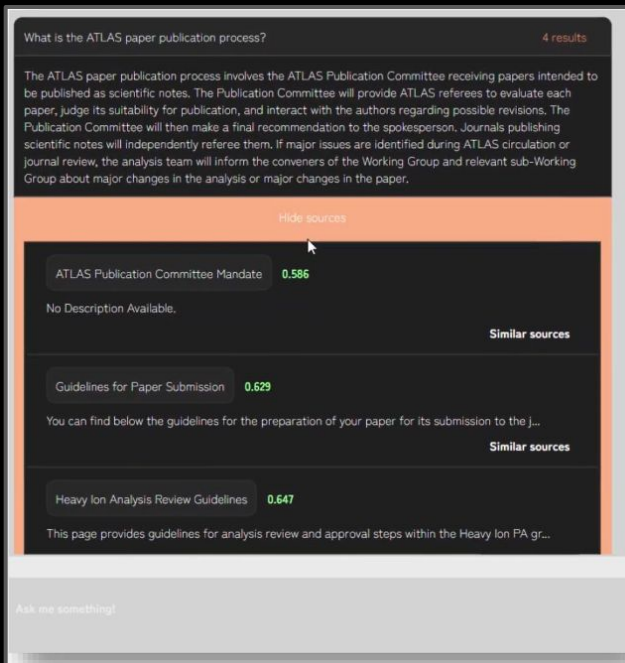
AI4EIC WORKSHOP
NOVEMBER 30, 2023



BERKELEY LAB

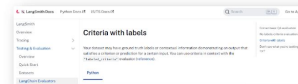


LOUISIANA TECH
UNIVERSITY



OPEN QUESTIONS & CONCLUSION

- How to avoid hallucinations? Probably "GPT-5 / Q* / AGI" will make this hurdle irrelevant
- How to best "censor" politically incorrect responses (e.g. which analysis team is the best?)
- How to measure the quality of responses – LangSmith AI-assisted evaluators?
- What is the best dataset to gather for fine-tuning?
- How to anonymize email threads and discussion forums?



We are having a lot of fun building this thing from scratch, but if there was an open-source scientific community framework for AI Assistants, it would be even more fun!

Talk at AI4EIC-2023

An AI Assistant for the ATLAS Collaboration

The ongoing AI4EIC Project:

Retrieval Augmented Generation using LLM pipeline

Ingestion

- Load data (**semantic**) both structured and unstructured from sources
- Split data in small repetitive chunks – **text corpus**
- **Embed** these chunks into a vector space using an embedding model
- Store these vectors in a database for retrieval later.

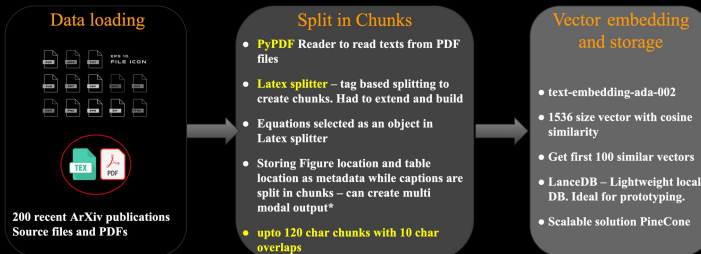


Inference

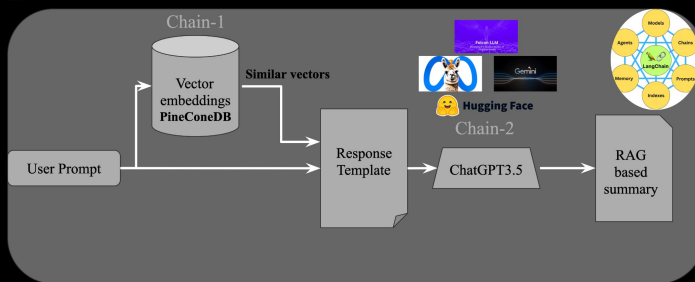
- Given a prompt, compute similarity index. Select the most closest vectors
- Choose a response template. Embed the vectors along with input prompt and feed into LLM
- Evaluate the response
 - Model fine-tuning
 - Build metrics for context tuning
- Multi Modal Output

A pilot study – Ingestion

- The Objective – Summarization tool with “relevant citations” for within EIC.
- Framework built has to be scalable and fairly automated



The Generic Workflow of building Traditional RAG



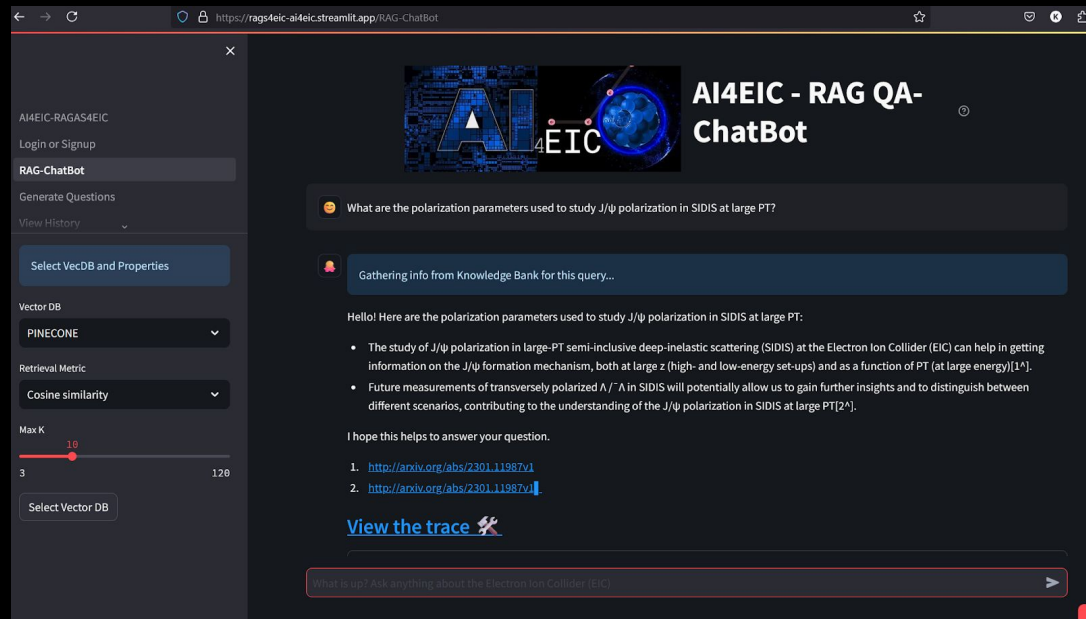
Talk at AI4EIC-2023

A Large Language Model-based Assistant for the Electron Ion Collider

The ongoing AI4EIC Project:

RAG application and evaluation

Repository: <https://github.com/ai4eic/EIC-RAG-Project>
 Web-Application: <https://rags4eic-ai4eic.streamlit.app/>



[Check out the full trace of the pipeline in langsmith](#)

Classic Metrics

Metric Name	Definition	Score
Output Renderer Frequency	Frequency of correctly rendered output response in the markdown format	78.0% ± 5.8%
Claim Recognition Rate	$CRR = \frac{\text{Number of answered claims}}{\text{Total number of claims in question}}$	96.4% ± 3.4%
Claim Accuracy Rate	$CAR = \frac{\text{Number of correctly answered claims for a question}}{\text{Total number of recognized claims in question}}$	88.9% ± 8.3%
Source Citation Frequency	$SCF = \frac{\text{Number of time question's source cited}}{\text{Total queries}}$	85.3% ± 5.0%
Hallucination Frequency	$HF = \frac{\text{Number of hallucinations}}{\text{Total queries}}$	2% ± 2%

RAGAS – LLM based evaluation

Metric Name	Definition	Score
Faithfulness	Fraction of correctly rendered output response in the markdown format	87.4% ± 5.5%
Context Relevancy	Relevancy of the retrieved context to the question and generated answer	61.4% ± 4.3%
Context Entity Recall	$CER = \frac{\text{claims in context} \times \text{claims in ground truth}}{\text{claims in ground truth}}$	98.7% ± 1.2%
Answer Relevancy	Measures relevancy of the generated answer to the retrieved context and the question	77.2% ± 2.3%
Answer Correctness	Measures correctness of the generated answer to the idea response	72.3% ± 2.4%

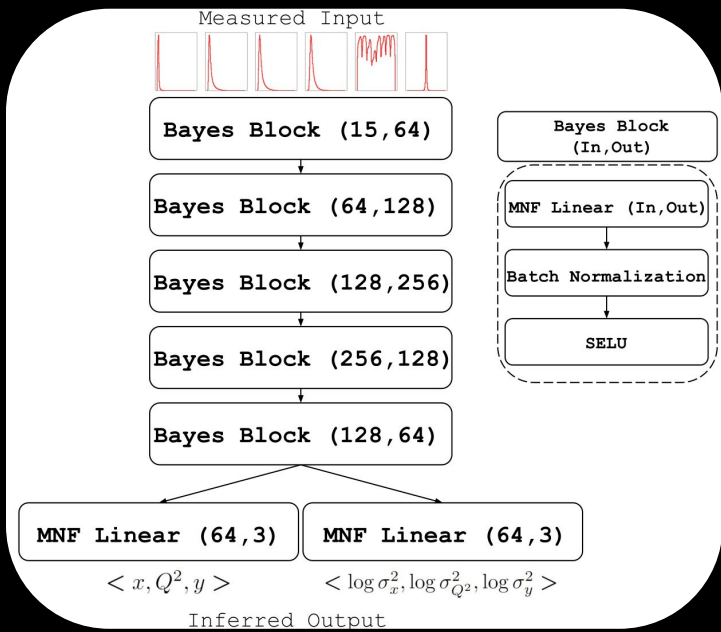
Talk at AI4EIC-2023

A Large Language Model-based Assistant for the Electron Ion Collider

ELUQuant: Event-level UQ

CF, J. Giroux 2024 Mach. Learn.: Sci. Technol. 5 015017

Event-Level Uncertainty Quantification



<https://github.com/wmdataphys/ELUQuant>

$$\mathcal{L}_{Tot.} = \mathcal{L}_{Reg.} + \gamma \mathcal{L}_{Phys.} + \beta \mathcal{L}_{NF.}$$

Learn the Posterior over the weights

$$\mathcal{L}_{MNF.} = \mathbb{E}_{q(\mathbf{W}, \mathbf{z}_T)} [-KL(q(\mathbf{W}|\mathbf{z}_{T_f})||p(\mathbf{W})) + \log r(\mathbf{z}_{T_f}|\mathbf{W}) - \log q(\mathbf{z}_{T_f})]$$

Access epistemic uncertainty through sampling MNF [1] layers

Learn the regression transformation

$$\mathcal{L}_{Reg.} = \frac{1}{N} \sum_i \sum_j \frac{1}{2} (e^{-s_j} \|\mathbf{v}_j - \hat{\mathbf{v}}_j\|^2 + s_j), \quad s_j = \log \sigma_j^2$$

Access aleatoric as a function of regressed output [2]

Constrain the physics

$$\mathcal{L}_{Phys.} = \frac{1}{N} \sum_i \log \hat{Q}_i^2 - (\log s_i + \log \hat{x}_i + \log \hat{y}_i)$$

Aleatoric (ELUQ) vs RMS (other)

Y Bin	DA Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.147955	0.061922	0.057942
(0.2, 0.5)	0.134833	0.075418	0.061706
(0.1, 0.2)	0.145530	0.097903	0.071238
(0.05, 0.1)	0.175290	0.132783	0.082945
(0.01, 0.05)	0.252723	0.184589	0.115453

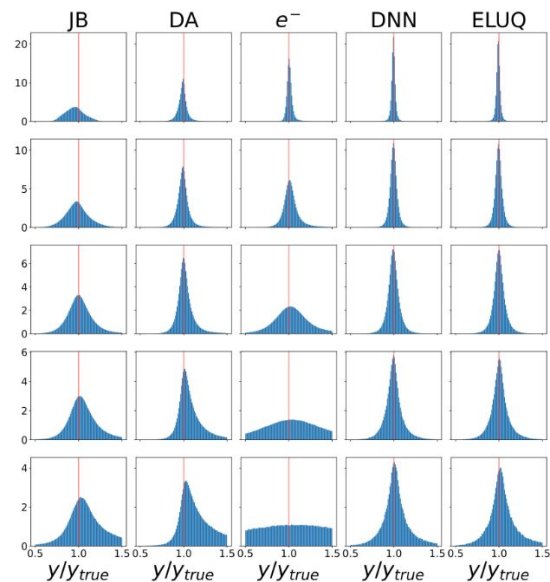
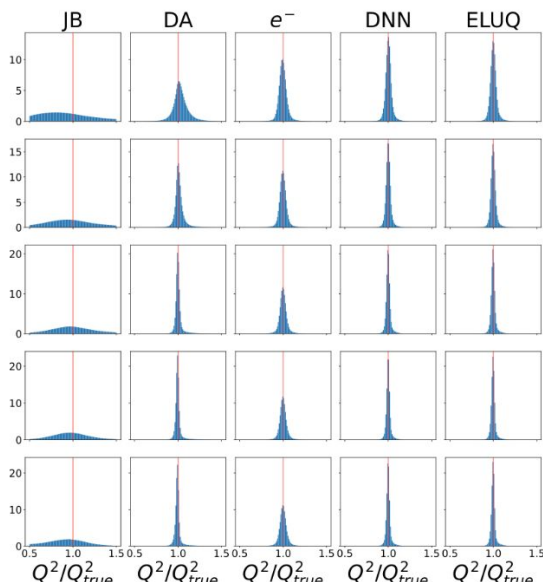
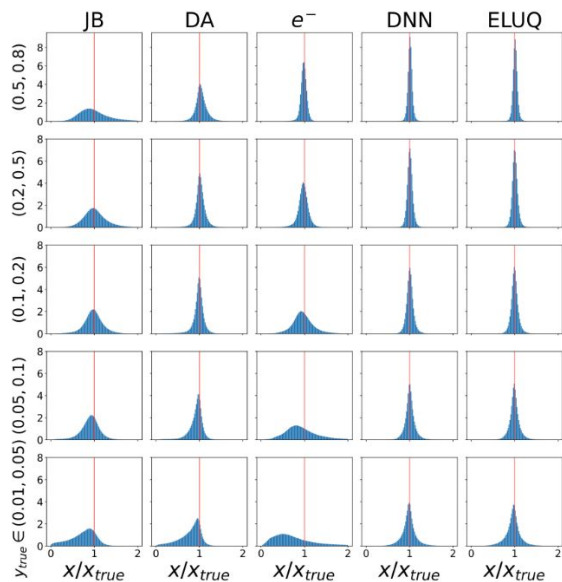
Table 2: Aleatoric RMS Comparisons - X

Y Bin	e Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.056694	0.044052	0.041349
(0.2, 0.5)	0.055787	0.037505	0.032280
(0.1, 0.2)	0.054219	0.033230	0.029640
(0.05, 0.1)	0.053403	0.032501	0.029411
(0.01, 0.05)	0.053470	0.032139	0.029431

Table 3: Aleatoric RMS Comparison - Q2

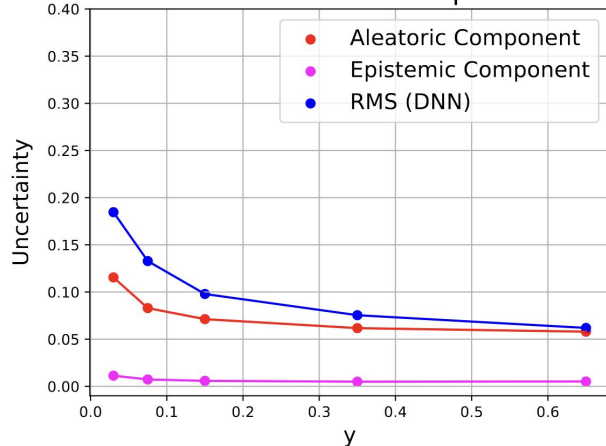
Y Bin	DA Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.060537	0.031194	0.034643
(0.2, 0.5)	0.082115	0.053126	0.044249
(0.1, 0.2)	0.098631	0.078143	0.061840
(0.05, 0.1)	0.127276	0.109309	0.078276
(0.01, 0.05)	0.158493	0.147391	0.120546

Table 4: Aleatoric RMS Comparison Y

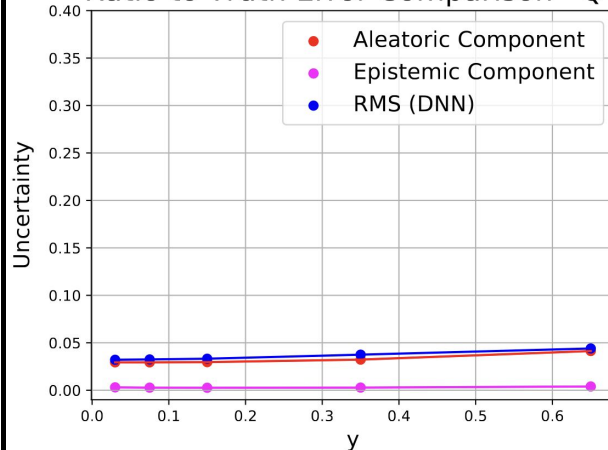


ELUQ vs DNN

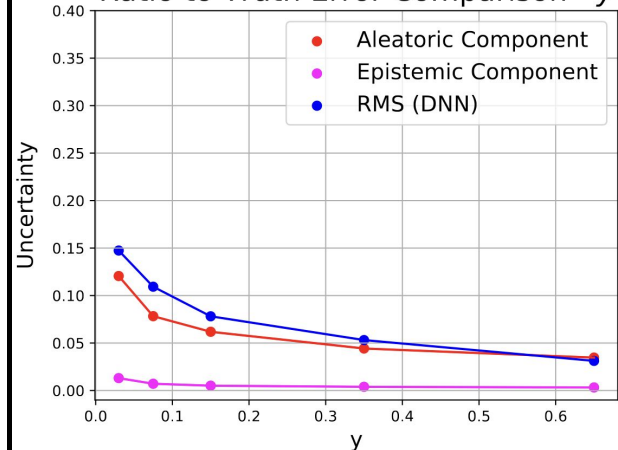
Ratio to Truth Error Comparison - x



Ratio to Truth Error Comparison - Q^2

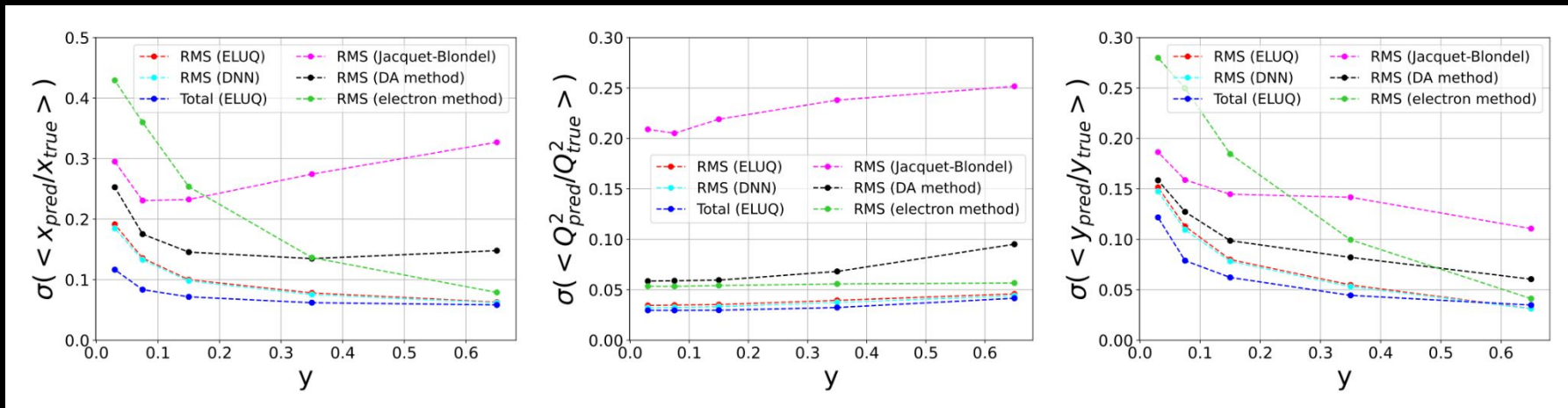


Ratio to Truth Error Comparison - y



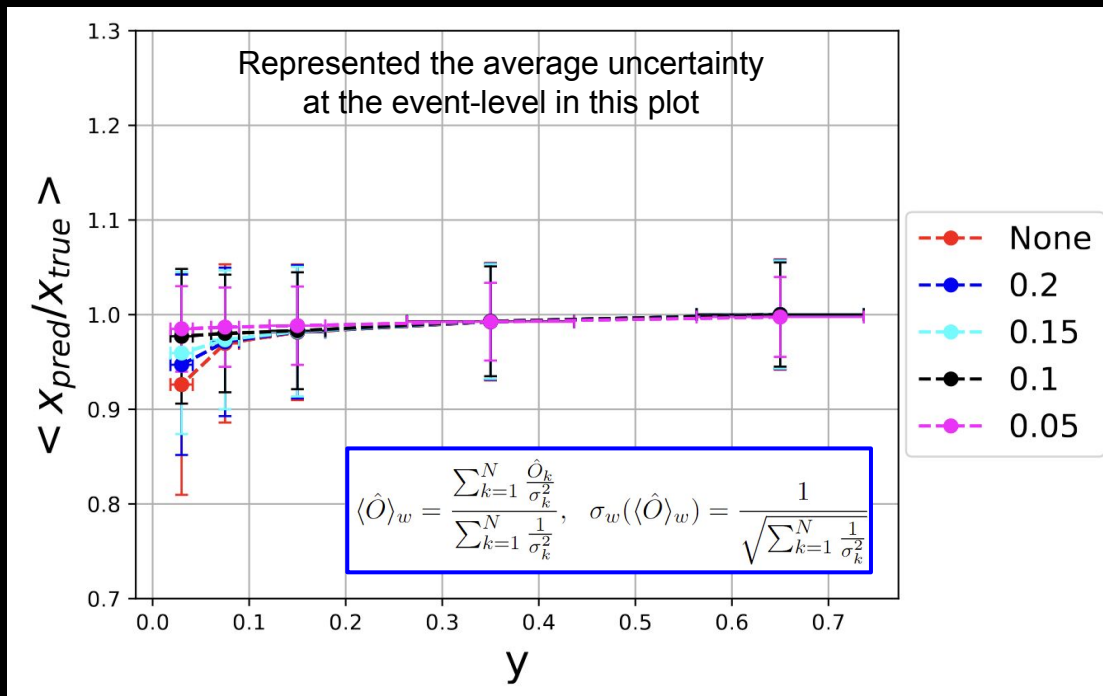
- The RMS (MNF) roughly coincide with that of DNN as seen previously
- The RMS (DNN) for x and y is larger at low y given the distributions are broader
- The epistemic is systematically smaller than aleatoric component.
- At large y , for x and y the total uncertainty (epistemic+aleatoric) close to RMS of DNN

All Methods Compared



- At low y , the RMS are typically larger due to “broader” distributions
- DNN and MNF have smaller RMS over the whole y range compared to other methods (this was also the finding of [NIM-A 1025 \(2022\): 166164](#))
 - “Our method outperforms other methods over a wide kinematics range”
 - “The RMS resolution for y and x increase at lower y , even for the DNN reconstruction. ... This results ... may be attributed to further acceptance, noise, or resolution effects that deteriorates the measurement of the HFS”

Leveraging Event-Level Information

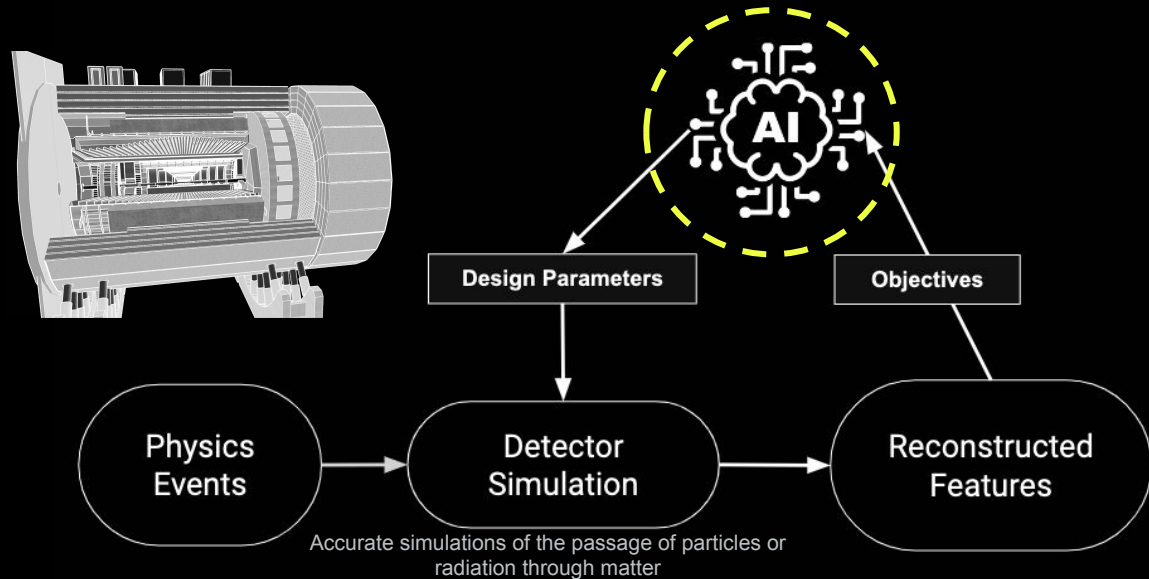


- Removing events with large relative event-level uncertainty (with respect to the network prediction) improve the ratio to truth and reduce inaccuracy
- Notice these cuts do not use any information at the ground truth level
- Bonus: this could be sensitive to anomalies.

— In the plot above events with at least one among x, Q^2, y with a relative uncertainty larger than a threshold are removed —

AI-assisted Detector Design

Compute-intensive simulation pipelines



- 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

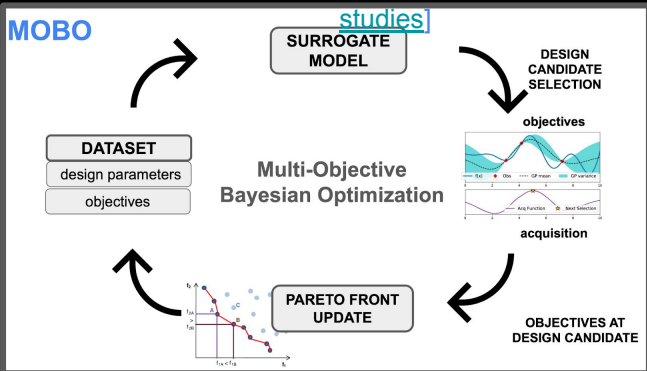
Provide a framework for an holistic optimization of the sub-detector system
A complex problem with (i) **multiple design parameters**, driven by (ii) **multiple objectives** (e.g., detector response, physics-driven, costs) subject to (iii) **constraints**

AI-assisted Detector Design @EIC

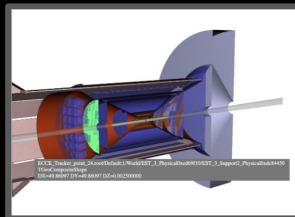


M. Diefenthaler et al (AID2E Collaboration) 2024 JINST 19 C07001

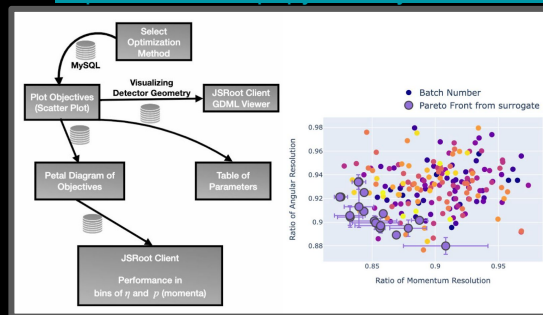
[\[Link to complexity studies\]](#)



(i) Advance state of the art MOO complexity to accommodate a large number of objectives and explore usage of physics-inspired approaches

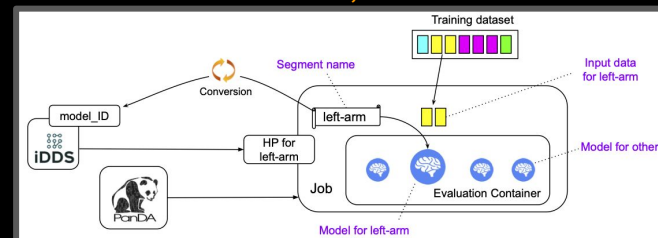


<https://ai4eicdetopt.pythonanywhere.com/>



(ii) Development of suite of data science tools for interactive navigation of Pareto front (multi-dim design with multiple objectives)

Distributed, scalable!



(iii) Leverage cutting-edge workload management systems capable of operating at massive data and handle complex workflows

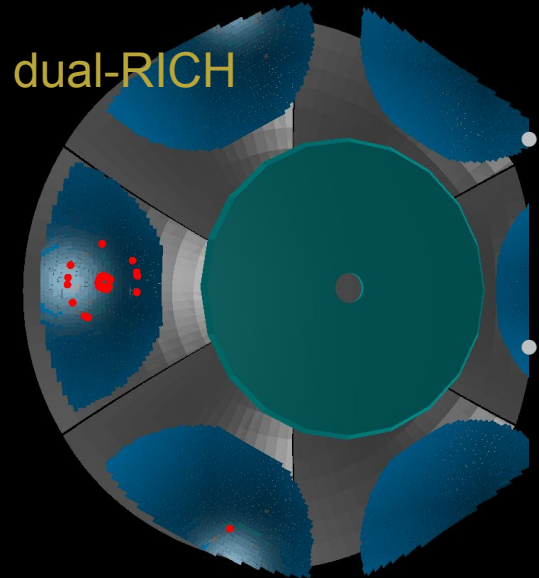


py ∞

- Examining solutions on the Pareto front of EIC detectors at different values of the budget can have great cost benefits
- A fractional improvement in the objectives translates to a more efficient use of beam time which will make up a majority of the cost of the EIC over its lifetime

AID2E Applications

- Considering all the constraints as ePIC is in the process of finalizing engineering designs, we can select those sub-detectors that still have tunable parameters



dual-RICH

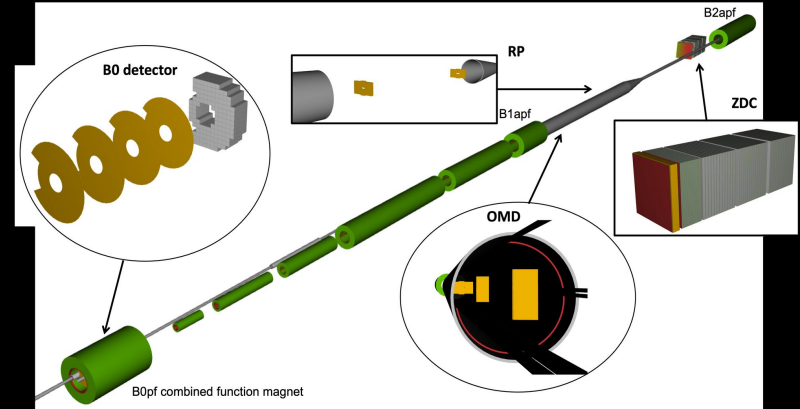
Mirror, sensor placement, gas, mirror material (lower costs material)...

PID performance, costs, ...

E. Cisbani *et al* 2020 *JINST* **15** P05009

Far-Forward

- B_0 magnetic field map, distance between layers, central location of tracker*
- Momentum resolution, acceptance*

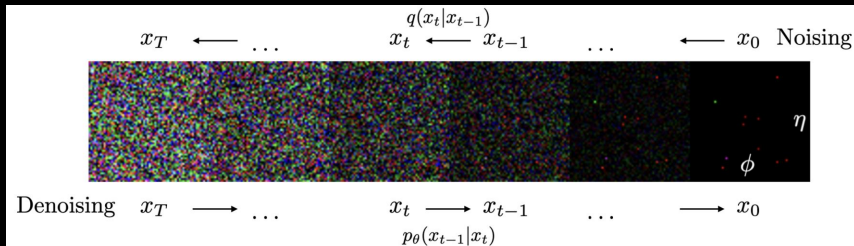


- AID2E framework can support the design of a possible detector-2; it is also being used for other compute-intensive tasks, such as alignment and calibration

Diffusion Models

"Diffusion model approach to simulating electron-proton scattering events"

- **Denoising Probabilistic Diffusion Models (DDPM)**



- Image based representation of (NC) DIS events
 - Center-of-mass energy representative of EIC $s = (105 \text{ GeV})^2$
 - Uniform pixelization of Cylindrical detector
- Pixel intensity represented as rescaled particle momentum
- PID represented through different channels (analogous to RGB image)
- Reconstruct high-level observables from generated images

