Artificial Intelligence and Machine Learning Activities and Next Steps at ePIC





<u>Asking ChatGPT:</u>

What Artificial Intelligence and Machine Learning activities are currently done at the Electron Ion Collider and what are the next steps?

The Electron Ion Collider (EIC) is a proposed next-generation particle accelerator that is being designed to collide electrons with ions in order to study the structure of matter at the quantum level. The EIC is not yet built, and as such it is not currently possible to say what specific artificial intelligence (AI) or machine learning (ML) activities are being carried out at the facility.

However, it is likely that AI and ML techniques will be used in various aspects of the EIC's operations, such as data analysis, control and automation systems, and detector design and optimization. Some potential applications of AI and ML in the EIC might include:

- Developing algorithms for analyzing and interpreting the large volumes of data that will be generated by the EIC's experiments.
- Designing machine learning models to identify and classify particles produced in the EIC's collisions.
- Developing AI-based control systems for optimizing the performance of the EIC's accelerator and detectors.

It is not possible to say what the next steps will be for AI and ML at the EIC until the facility is built and begins operations. However, these are some potential areas where AI and ML could play a role in the EIC's research and operations.

This is not completely true :) see next slide. Of course it remains hard to anticipate exactly how AI/ML along with the computing landscape will evolve in the next 10 years



\$

<u>AI in the EIC Schedule</u>



ChatGPT can only fetch data prior to the year 2021 as its training stopped in the year 2021

We had two workshops, last one was in October 2022...



- Design, simulations, reconstruction and analysis with AI/ML are activities that are already ongoing
- Al-based Control (that presumably will start slightly before with accelerator) and SRO are next steps



[1] https://eic.ai/workshops

<u>Outline</u>

- Modern features of ePIC SW favorable for AI/ML
- Ongoing activities
 - Detector design
 - (Simulation)/Reconstruction/PID/Analysis
 - Al4EIC + hackathon example
- Next steps
 - Tutorial/documentation
 - AI/ML in SRO and for Control
 - Desiderata (a non-exhaustive list)
- Conclusions

"The ePIC detector could be (one of) the first large-scale detector to be designed with the assistance of AI"

"Streaming Readout could allow rapid turnaround of physics data and integration of AI/ML tools"

"In ~2030, the ePIC detector could be one of the first largely automated experiments where AI will be largely applied for control."



<u>AI-favorable features of ePIC SW stack</u>

- Design:
 - Geometry implementation via data source makes transparent the coupling of AI to the software stack design parameters
 - Modularity of geometry description
 - Automated features (checking overlaps)
- More in general:
 - Code repository, continuous integration, containerization
 - Open, simple, self-descriptive data formats (flat data model in general allows flexibility for AI/ML applications)
 - Support for truth information
- Use of HEP-supported packages (e.g., ACTS, includes ONNX plugin)
 - JANA2 includes an integrated Python interface

From YR: Physics WG Detector Requirements

Hermetic and multi-purpose detector

- Inclusive, semi-inclusive, exclusive measurements
- Jets, heavy quark
- Diffractive and tagging

Requirements are needed to fulfill the EIC physics program

_		Noman	alatura	Tracking			Electrons and Photons			π/K/p PID		HCAL		Muono	
η	Nomenciature			Min p _T	Resolution	Allowed X/X ₀	Si-Vertex	Min E	Resolutio n σ _E /E	PID	p-Range (GeV/c)	Separation	Min E	$\begin{array}{c} \text{Resolution} \\ \sigma_{\text{E}}/\text{E} \end{array}$	Muons
6.9 — -5.8			low-Q ² tagger		δθ/θ < 1.5%; 10 ⁻⁶ < Q ² < 10 ⁻² GeV ²										
		Auxiliary													
4.5 — -4.0	* p	Detectors	Instrumentation to separate charged particles from γ												
4.0 — -3.5														~50%/√E+6%	
3.5 — -3.0									2%/√E+ (1.3)%						
3.0 — -2.5					σ _p /p ~ 0.1%×p+2.0%		σ _{xy} ~30µm/p _T +		(1-5)76						
2.5 — -2.0			Backwards Detectors		2		чорт				≤ 7 GeV/c			~45%/√E+6%	
2.0 — -1.5					$\sigma_p/p \sim 0.05\% \times p{+}1.0\%$		σ _{xy} ~30µm/p⊤+ 20µm		70/1/=+						
1.5 — -1.0									(1-3)%	π suppression					
1.0 — -0.5										up to 1:104					
0.5 - 0.0		Control		100 MeV/ π		50/	σ _{xyz} ~ 20 μm.	50					500		
0.0 - 0.5		Detector	Barrel	σ _p /p ~ (r _p /p ~ 0.05%×p+0.5%	~5% or less	or s d₀(z) ~ d₀(rφ) ~ 20/p⊤ GeV μm + 5 μm	MeV			≤ 10 GeV/c	≥ 3σ	~500 MeV	~85%/√E+7%	Useful for bkg, improve resolution
0.5 — 1.0				135 MeV K					(10-12)%/		≤ 15 GeV/c				
1.0 — 1.5											≤ 30 GeV/c				
1.5 - 2.0					σ₂/n ~ 0.05%×n+1.0%		σ~30um/pτ+		√E+(1-3)%						
2.0 - 2.5			Forward Detectors			20µm			3σ e/π	≤ 50 GeV/c					
2.5 - 3.0					$\sigma_{\rm p}/{\rm p} \sim 0.1\% \times {\rm p+2.0\%}$		σ _{xy} ~30μm/p _T +				< 30 GeV/c			~35%/√E	
3.0 - 3.5							40μm σ _{xy} ~30μm/p _T +				< 45 GoV/c				
3.5 - 4.0	∱e Au Det		Instrumentation to separate charged				60µm				3 40 Gev/c				
4.0 - 4.5		Auxiliary Detectors													
			particles nom 7												
>6.2					Gintrinsic(1/1)/1t1 < 1%:							-			
			Proton Spectrometer		Acceptance:										
					0.2< pT <1.2 GeV/c										



[4] R. Abdul Khalek, et al. "EIC yellow report." Nuclear Physics A 1026 (2022): 122447

<u>AI-assisted Design</u>

- Majority of the technology choices have been made, and we expect to work mainly with continuous parameters.
- Examples: plenty of degrees of freedom to optimize in the central (e.g., tracking, dRICH) and in the far-forward/backward systems



Al-assisted tools (with human guidance), allow to achieve design goals (physics and detector performance, costs)

Multi-Objective Optimization Detector response Physics gains Costs

- Learning the optimal compromises between competing objectives
- Holistic Optimization of ePIC
- Accelerate design process, understand tradeoffs, while optimizing costs





[5] CF et al., Nucl.Instrum.Meth.A 1047 (2023) 167748 [6] K. Suresh,interactive navigation of Pareto front, <u>https://ai4eicdetopt.pythonanywhere.com/</u>

Sim/Reco/Ana: ongoing activities/examples



efferson Lab

[7] Examples taken from AI4EIC workshop https://indico.bnl.gov/e/AI4EIC

...and many more activities

<u>Tutorials/</u>

<u>Documentation</u>

- Example from AI4EIC hackathon
 - PID with dRICH of ePIC using full simulations based on the ePIC SW stack
- Tutorial/documentation for the hackathon:
 - Documentation (problem description and dataset): <u>https://doi.org/10.5281/zenodo.7197023</u>
 - Provided example training scripts in both PyTorch and TensorFlow for a "plug and play" experience (focus on model development and insert it into our pipeline and have the training/evaluation code)
 - Provided detailed description on how to utilize the framework, AWS environments, deal with resource contention, and perform distributed training for more advanced users



The best solutions were all Machine Learning/Deep Learning-based, they were quite original, and they outperformed solutions based on classical cut-based approaches (followed by some teams). While this is only a first step towards deeply learning the identification of particles reconstructed with the d-RICH, these exploratory studies clearly indicates the potential of ML/DL approaches for reconstruction and PID.



Training Events	1.5 Million Events	With Magnetic Field ($\sim 1.5T$)			
Momentum	15 GeV/c	at Interaction Point $(0, 0, 0)$			
Theta θ	20°	at Interaction Point $(0, 0, 0)$			
Phi ϕ	0°	at Interaction Point $(0, 0, 0)$			
Training Events	3 Million Events	With Magnetic Field ($\sim 1.5T$)			
Momentum	15-20GeV/c	at Interaction Point $(0, 0, 0)$			
Theta θ	$15 - 16^{\circ}$	at Interaction Point $(0, 0, 0)$			
Phi d	$0 - 5^{\circ}$	at Interaction Point (0, 0, 0)			

Problem	Threshold		
Number	Accuracy		
Problem 1	94%		
Problem 2	86%		
Problem 3	80%		

*Problem 3: addition of noisy hits

Best solutions:

- JINR team: CatBoost
- Jets team: 2D CNN

https://ai4eichackathon.pythonanywhere.com/leaderboard

Moving forward

- Add machine learning libraries to the EIC environment
- Tutorials/Documentations
- Study flexible and efficient algorithms for SRO (work with stream of data robust against potential change in experimental conditions)
- Large-scale AI/ML architectures
 - Considerations on scalability and specific infrastructure need
 - ML lifecycle;
 - Distributed training:
 - may become necessary in AI pipelines working with big data: training time exponentially increases, scalability cumbersome, other limitation factors (e.g., algorithm computational complexity outpaces the main memory)
 - ...and more





- The EIC community is active on AI/ML (not included are also funded AI activities for EIC):
 - The number of AI/ML activities in ePIC is anticipated to grow in the next few months (e.g., reconstruction, PID) with more data from our simulation campaigns
 - In the long-term, AI/ML will likely permeate and contribute to multiple aspects of near real-time analyses
- Need support for interdisciplinary research and develop multi-disciplinary workforce:
 - It would be advantageous to have a more coherent workforce on AI/ML (e.g., AI sub-WG)
 - Educational/training activities simply necessary. This is a discussion that already started in the Computing and Software WG and the idea is to actively begin this year.
- Transitioning from prototyping to deployment in production environments
- Not enough emphasis on algorithmic development
 - e.g., we do not know what is the best ML/DL that we can do with Cherenkov detectors, though there are clear hints it is worth continuing these studies (event-level reco)
- Robustness, explainability, also very important cross-cutting themes for applications in our field.





[1] AI4EIC, https://eic.ai

[2] NP LRP, Hot and Cold QCD Town Hall

https://indico.mit.edu/event/538/contributions/1194/attachments/527/903/NSAC-LRP_MIT_2022_3.pdf

[3] ePIC SW Infrastructure Review, "AI/ML Synergy", https://indico.bnl.gov/event/16676/contributions/66948/attachments/42874/71986/Slides%20-AI_ML%20syn ergy.pdf

[4] EIC Yellow Report, R. Abdul Khalek, et al. "EIC yellow report." Nuclear Physics A 1026 (2022): 122447

[5] Al-assisted Optimization of the ECCE Tracking System at the Electron Ion Collider, Nucl.Instrum.Meth.A 1047 (2023) 167748

[6] K. Suresh, interactive navigation of Pareto front, <u>https://ai4eicdetopt.pythonanywhere.com/</u>

[7] 2nd workshop on AI for EIC, <u>https://indico.bnl.gov/e/AI4EIC</u>

[8] Computational Nuclear Physics and AI/ML Workshop, 6-7 Sep 2022

[9] P. Bedaque, et al. "Al for nuclear physics." The European Physical Journal A 57.3 (2021): 1-27.



Backup

Hot & Cold QC Community



1:00	EIC Science: ep Reactions	Hatta Yoshitaka 🥝
	Stata Center, 32-123, MIT	11:00 - 11:15
	EIC Science: eA Reactions	Anna Stasto 🥝
	Stata Center, 32-123, MIT	11:15 - 11:30
	EIC Theory Workshop Summary	Iain Stewart 🥝
	Stata Center, 32-123, MIT	11:30 - 11:33
	Discussion	
	Stata Center, 32-123, MIT	11:33 - 11:45
	Lattice theory for Hot and Cold QCD	Martha Constantinou 🥔
2:00	Stata Center, 32-123, MIT	11:45 - 12:05
	Machine Learning and Artificial Intelligence Applications for QCD (exp)	Cristiano Fanelli 🥝
	Stata Center, 32-123, MIT	12:05 - 12:20
	Machine Learning and Artificial Intelligence Applications for QCD (theory)	Phiala Shanahan 🥝
	Stata Center, 32-123, MIT	12:20 - 12:35
	Discussion	
	Stata Center, 32-123, MIT	12:35 - 12:50

Recommendation 2: EIC Project

We recommend the expeditious completion of the EIC as the highest priority for facility construction.

The Electron-Ion Colider (EIC) is a powerful and versatile new accelerator facility, capable of coliding high-energy beams ranging from heavy ions to polarized light ions and protons with high-energy polarized electron beams. In the 2015 Long Range Plan the EIC was put forward as the highest priority for new facility construction and the expeditious completion remains a top priority for the nuclear physics community. The EIC, accompanied by the general-purpose large-acceptance detector, ePIC, will be a discovery machine that addresses fundamental questions such as the origin of mass and spin of the proton as well as probing dense gluon systems in nuclei. It will allow for the exploration of new landscapes in OCD, permitting the "tomography", or high-resolution multidimensional mapping of the quark and gluon components inside of nucleons and nuclei. Realizing the EIC will keep the U.S. on the frontiers of nuclear physics and accelerator science and technology.

- Building on the recent EIC project CD-1 approval, the community-led Yellow-Report, and detector proposals, the QCD research community is committed to continue the development and timely realization of the EIC and its first detector, ePIC. We recommend supporting the growth of a diverse and active research workforce for the ePIC collaboration, in support of the expeditious realization of the first EIC detector.
- We recommend new investments to establish a national EIC theory alliance to enhance and broaden the theory
 community needed for advancing EIC science and the experimental program. This theory alliance will contribute
 to a diverse workforce through a competitive national EIC theory fellow program and tenure-track bridge
 positions, including appointments at minority serving institutions.

Recommendation 2: EIC Project



Recommendation 4: Computing

High-performance and high-throughput computing are essential to advance nuclear physics at the experimental and theory frontiers. Increased investments in computational nuclear physics will facilitate discoveries and capitalize on previous investments.

- We recommend increased investments for software and algorithm development, including in AI/ML, by strengthening and expanding programs and partnerships, such as the DOE SciDAC and NSF CSSI and AI institutes.
- We recommend increased support for dedicated high-performance and high-throughput mid-scale computational hardware and high-capacity data systems, as well as expanding access to leadership computing facilities.
- Advanced computing is an interdisciplinary field. We recommend establishing programs to support the development and retention of a diverse multi-disciplinary workforce in high-performance computing and AI/ML.

Recommendation 4: Computing





2 plenary talks on AI/ML (theory and experiment)

dRICH reconstruction

- Indirect Ray Tracing (IRT)
 - The basic idea is that, given tracking information and RICH PMT hits, the Cherenkov-photon emission angle can be reconstructed.
 - The distribution of observed photon angles is compared to the expected angle for each particle type and the most likely particle type is determined.
 - Fast, non computationally intensive. Lowest accuracy compared to other methods in this slide.
- Direct Ray Tracing (DRT)
 - Simulates a PMT hit pattern based on the track kinematics and particle hypothesis
 - Construct likelihood by comparing "PDF" to the observed hit pattern
- Event-level algorithm (EVT)
 - Motivation: two close tracks can produce misidentification
 - Builds upon DRT. Improvement by looking at each event as a whole rather than individual tracks
 - \blacksquare \rightarrow sum over all tracks in the event







16

AI/ML in SRO

The development of streaming readout (SRO) for the NP driven by research initiatives:

- Streaming Grand Challenge [1] and the facility for "Innovation in Nuclear Data Readout and Analysis" (INDRA) at JLab
- BNL LDRD "High Throughput Advanced Data Acquisition for eRHIC, Particle Physics and Cosmology Experiments"
- PHENIX, STAR and sPHENIX (BNL), • KM3NeT(INFN), BDX (JLAB) and CBM (FAIR)



SRO for next generation electron scattering [2]

ML deployed on stream of real data

CLAS + EPSCI @JLab



Hierarchical clustering VS traditional clustering of energy deposited by photons; Al robust against variations in experimental conditions* (uncalibrated data in SRO)

Courtesy of M. Battaglieri (JLab)

Many active projects regarding SRO at JLab: INDRA/ASTRA [3], AIEC (AI for Experimental Control) [4], Hydra (Online monitoring) [5], SRO with ML on FPGA [6]

[1] A. Boehnlein, R. Ent, and R. Yoshida, Grand Challenge in Readout and Analysis for Femtoscale Science, 2018 [2] F. Ameli, et al., Streaming readout for next generation electron scattering experiments, Eur. Phys. J. Plus, 2022 [3] M. Diefenthaler et al., Diefenthaler, Markus, et al. Evaluation & Development of Algorithms & Techniques for Streaming Detector Readout. No. 2020-LDRD-LD2014. 2020. [4] T. Jeske, et al. "AI for Experimental Controls at Jefferson Lab." JINST 17.03 (2022): C03043. — AI4EIC proceedings [5] T. Britton, B. Nachman. "Accelerator and detector control for the EIC with machine learning." JINST 17.02 (2022): C02022. — Al4EIC proceedings [6] S. Furletov et al., Machine learning on FPGA for event selection - Al4EIC proceedings

CLAS12 SRO setup

TriDAS SR back end

The CLAS12 Forward Tagger, JLab



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

Intelligent Experiment Through Real-Time AI (DOE FOA funded 2022-2023)

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)



Collaboration of NP, HEP and CS: LANL, MIT, FNAL, NJIT, ORNL, UNT, CCNU



Courtesy of Ming Liu (LANL)

No effective conventional triggers available



[1] Huang, Yi, et al. "Efficient Data Compression for 3D Sparse TPC via Bicephalous Convolutional Autoencoder." 2021 20th IEEE (ICMLA). IEEE, 2021. [2] F. Fahim, et al., "HLS4ML" arXiv:2103.05579 (2021)

Unfolding and "data-driven" learning

Courtesy of B. Nachman Unfolding



OmniFold [1]

Using ML for differential cross section measurements (OmniFold and otherwise). These tools for recent measurements with DIS from HERA data and the same tools could be used at the EIC.



A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, and J. Thaler "OmniFold: A Method to Simultaneously Unfold All Observables" <u>Phys.</u> <u>Rev. Lett. 124.</u> 182001 2020 Courtesy of M. Arratia (UCR), B. Nachman Lepton-jet correlation in DIS at H1 [1]



- First example of ML-assisted unfolding (MultiFold method): enables simultaneous and unbinned unfolding in high dimensions.
- This development will allow us to do unbinned cross-section measurements
- Similarly, this could be applied at EIC

[1] V. Andreev et al. (H1 Collaboration), "Measurement of Lepton-Jet Correlation in Deep-Inelastic Scattering with the H1 Detector Using Machine Learning for Unfolding" <u>Phys. Rev. Lett. 128, 132002</u> In the "opposite" direction, it could be exciting thinking about data-driven learning that relies less on simulations, with tools like, e.g., one-class classification / anomaly-detection [1] and weak supervision / topic modeling [2].

Flux+Mutability [1]





Same architecture applied to n/ $\!\gamma$ showers reconstruction in GlueX and BSM dijet signatures at LHC

Leitmotif in AI/ML

Courtesy of B. Nachman (LBNL)

Uncertainty Quantification

 $\begin{array}{c} \text{statistical (aleatoric) / systematic (epistemic)} \\ \text{decrease with more events} \\ & \text{model bias} \\ \hline \\ \text{Precision / Optimality: } NN(x) \neq \frac{P_{true}(x|S+B)}{P_{true}(x|B)} \\ & \text{precision / Optimality: } NN(x) \neq \frac{P_{true}(x|S+B)}{p_{true}(x|B)} \\ & \text{inaccurate training data} \\ NN(x)|_{P_{true}=P_{train}} \neq \frac{P_{true}(x|S+B)}{P_{true}(x|B)} \\ & \text{model/optimization flexibility} \\ & \text{Statistical uncertainty} \\ \end{array}$

limited prediction statistics



 $p_{\text{prediction}}(x) \neq p_{\text{true}}(x)$

inaccurate prediction data

used to estimate the p-value is not accurate."

inference/uncertainty-aware approaches



 B. Nachman, "UQ for ML Applied to Data Analysis", talk at <u>AI4EIC Meeting on Uncertainty Quantification</u>
 B. Nachman, *How to achieve optimality and account for* <u>uncertainty</u>, arXiv:1909.03081

Courtesy of M. Williams (MIT/IAIFI)



- The Lipschitz constant of the map between the input and output space represented by a neural network is a natural metric for assessing the robustness of the model.
- This new method constrains the Lipschitz constant of dense DL models (can also be generalized to other architectures). The method relies on a simple weight normalization scheme during training that ensures the Lipschitz constant of every layer is below an upper limit specified by the analyst.
- The algorithm was used to train a powerful, robust, and interpretable discriminator for heavy-flavor decays in the LHCb realtime data-processing system.
- LHCb has adopted this for the major selection algorithms, and looking at it for PID, fake-track killers.

[1] O. Kitouni, N. Nolte, M. Williams "Robust and Provably Monotonic Networks", arXiv:2112.00038