Experimental Opportunities using AI/ML at the EIC

Cristiano Fanelli

POETIC XI, Feb 24-28, 2025

WILLIAM & MARY
CHARTERED 1693

Disclaimer

- *This is not an ePIC talk, though I will frequently reference ePIC.*
- *I will not cover all ongoing AI/ML efforts in the EIC, particularly in ePIC, where most AI/ML activities are still in early stages.*
- *Some examples are drawn from other experiments or inspired by AI4EIC discussions (https://eic.ai) and Allaire, C., et al., Computing and Software for Big Science 8.1 (2024): 5).*
- *The talk primarily explores experimental / data analysis perspectives, showcasing immediately available ML-based approaches that could be widely adopted for the EIC science.*
- *I will not specifically delve into AI/ML for streaming readout, as it warrants a separate talk. However, I will highlight relevant connections and implicitly address aspects of near real-time analysis.*

AI/ML is Ubiquitous

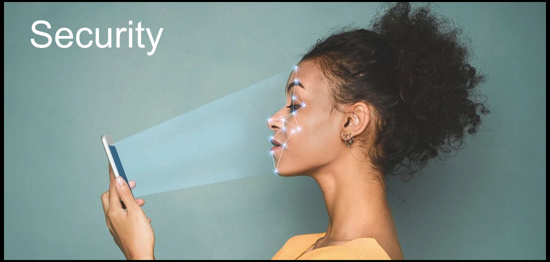
Assistance

AI-Powered POS

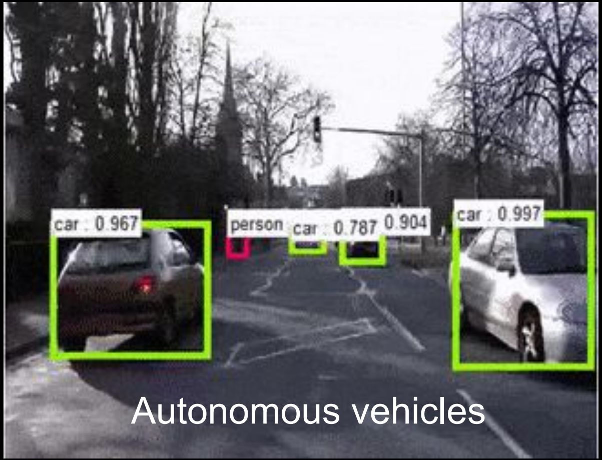
Entertainment

Health

Home automation



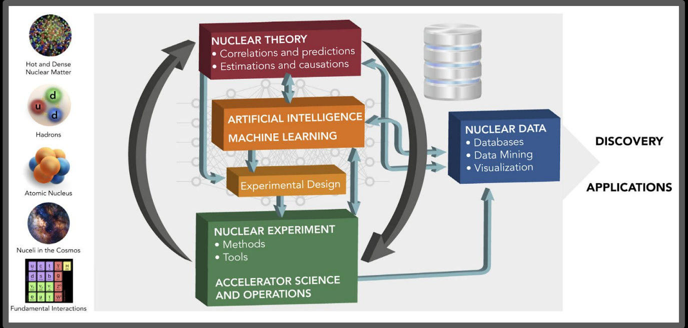
Security



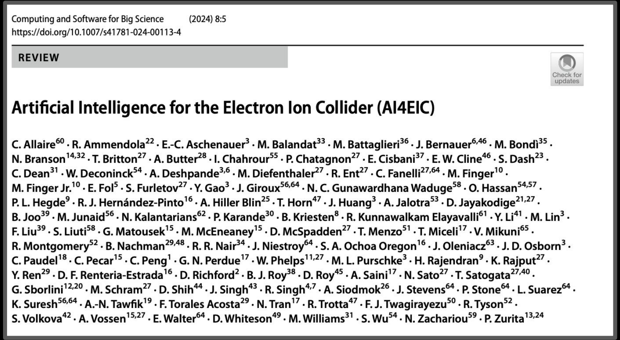
Autonomous vehicles

Experimental Opportunities for EIC

- AI/ML is ubiquitous in HEP and is increasingly shaping NP too.



A. Boehnlein, M. Diefenthaler, CF et al., Machine learning in nuclear physics, Rev. Mod. Phys. **94**, 031003 (2022) and references therein



Computing and Software for Big Science (2024) 8:5
https://doi.org/10.1007/s41781-024-00113-4

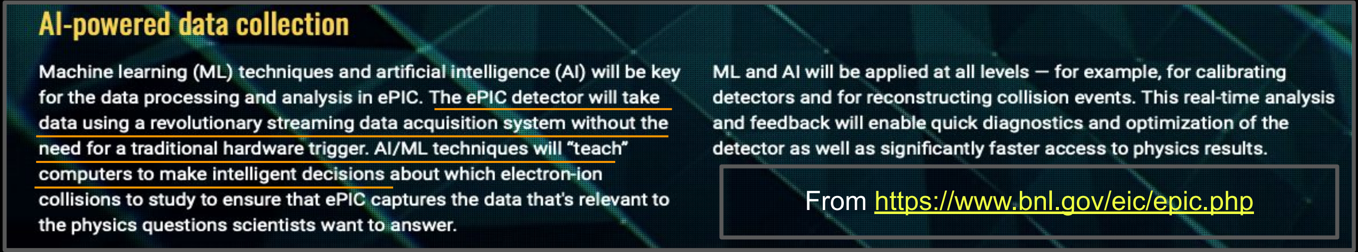
REVIEW

Artificial Intelligence for the Electron Ion Collider (AI4EIC)

C. Allaire[60] · R. Ammendola[22] · E.-C. Aschenauer[3] · M. Balandat[33] · M. Battaglieri[36] · J. Bernauer[6,46] · M. Bondì[35] · N. Branson[14,32] · T. Britton[27] · A. Butter[28] · I. Chahrour[55] · P. Chatagnon[27] · E. Cisbani[37] · E. W. Cline[46] · S. Dash[23] · C. Dean[31] · W. Deconinck[54] · A. Deshpande[3,6] · M. Diefenthaler[27] · R. Ent[27] · C. Fanelli[27,64] · M. Finger[10] · M. Finger Jr.[10] · E. Fol[5] · S. Furletov[27] · Y. Gao[3] · J. Giroux[56,64] · N. C. Gunawardhana Waduge[58] · O. Hassan[54,57] · P. L. Hegde[9] · R. J. Hernández-Pinto[16] · A. Hiller Blin[25] · T. Horn[47] · J. Huang[3] · A. Jalotra[53] · D. Jayakodige[21,27] · B. Joo[39] · M. Junaid[56] · N. Kalantarians[62] · P. Karande[30] · B. Kriesten[8] · R. Kunnawalkam Elayavalli[61] · Y. Li[41] · M. Lin[3] · F. Liu[39] · S. Liuti[58] · G. Matousek[15] · M. McEneaney[15] · D. McSpadden[27] · T. Menzo[51] · T. Miceli[17] · V. Mikuni[65] · R. Montgomery[52] · B. Nachman[29,48] · R. R. Nair[34] · J. Niestroy[64] · S. A. Ochoa Oregon[16] · J. Oleniacz[63] · J. D. Osborn[3] · C. Paudel[18] · C. Pecar[15] · C. Peng[1] · G. N. Perdue[17] · W. Phelps[11,27] · M. L. Purschke[3] · H. Rajendran[9] · K. Rajput[27] · Y. Ren[29] · D. F. Renteria-Estrada[16] · D. Richford[2] · B. J. Roy[38] · D. Roy[45] · A. Saini[17] · N. Sato[27] · T. Satogata[27,40] · G. Sborlini[12,20] · M. Schram[27] · D. Shih[44] · J. Singh[43] · R. Singh[4,7] · A. Siodmok[26] · J. Stevens[64] · P. Stone[64] · L. Suarez[64] · K. Suresh[56,64] · A.-N. Tawfik[19] · F. Torales Acosta[29] · N. Tran[17] · R. Trotta[47] · F. J. Twagirayezu[50] · R. Tyson[52] · S. Volkova[42] · A. Vossen[15,27] · E. Walter[64] · D. Whiteson[49] · M. Williams[31] · S. Wu[54] · N. Zachariou[59] · P. Zurita[13,24]

Allaire, C., et al. "Artificial Intelligence for the Electron Ion Collider (AI4EIC)." Computing and Software for Big Science 8.1 (2024): 5.

- The EIC is being realized during the AI revolution (and the genAI wave), and will be operating in the next decade. The EIC science aims to integrate AI/ML across all phases of the experiment, from design to operations.



AI-powered data collection

Machine learning (ML) techniques and artificial intelligence (AI) will be key for the data processing and analysis in ePIC. The ePIC detector will take data using a revolutionary streaming data acquisition system without the need for a traditional hardware trigger. AI/ML techniques will "teach" computers to make intelligent decisions about which electron-ion collisions to study to ensure that ePIC captures the data that's relevant to the physics questions scientists want to answer.

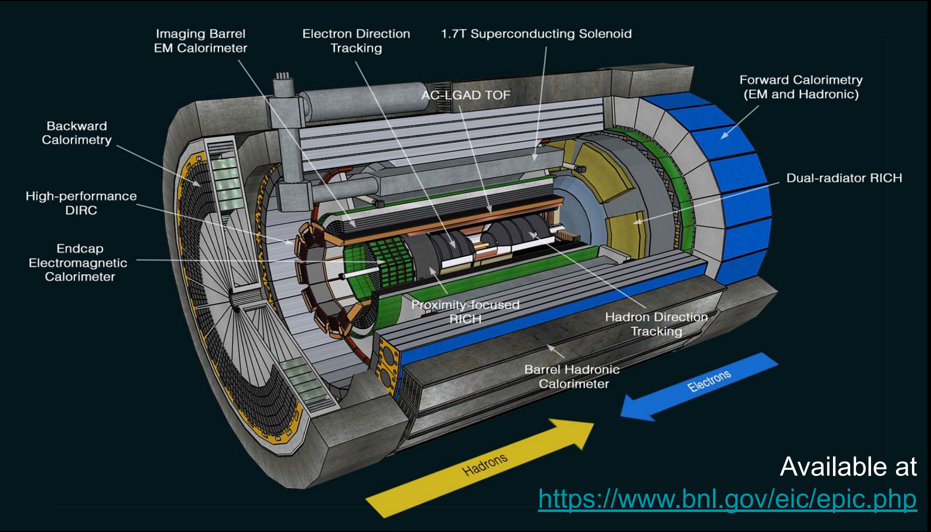
ML and AI will be applied at all levels — for example, for calibrating detectors and for reconstructing collision events. This real-time analysis and feedback will enable quick diagnostics and optimization of the detector as well as significantly faster access to physics results.

From https://www.bnl.gov/eic/epic.php

Experimental Opportunities

1. **Fast, High-Fidelity Simulations** – Emphasis on accurate and efficiently accelerated detector response modeling
2. **Enhanced PID**: Near real-time reconstruction of complex patterns across full kinematic ranges
3. (Bonus) **Data-Driven Learning from Real Data** – Directly leveraging high-purity real data (when available)
4. **Event-Level** – Holistic learning from (all) available event-level information
5. **Uncertainty Quantification** – Understanding and managing uncertainties in ML-based analysis
6. **Event-Level Uncertainty Quantification** – Combining (1) and (2) for more robust event-level insights
7. **Distributed Multi-Objective Optimization for Detector Systems** – Leveraging distributed MOO for detector design, alignment, and calibration while learning tradeoff solutions in a constrained multi-parameter space
8. **Agentic Scientific Workflows** – Multimodal AI assistance with specialized agents for tasks such as experimental shift-taking, data analysis, and EIC science interpretation

AI/ML for Cherenkov Detectors


Imaging Barrel EM Calorimeter
Electron Direction Tracking
1.7T Superconducting Solenoid
AC-LGAD TOF
Forward Calorimetry (EM and Hadronic)
Backward Calorimetry
Dual-radiator RICH
High-performance DIRC
Endcap Electromagnetic Calorimeter
Proximity-focused RICH
Hadron Direction Tracking
Barrel Hadronic Calorimeter
Electrons
Hadrons

Available at https://www.bnl.gov/eic/epic.php

- Cherenkov detectors constitute the backbone of PID (DIRC, dRICH, pfRICH)
- They represent a major simulation bottleneck in that optical photons involve multiple photons that need to be tracked through complex surfaces (S. Joosten's talk at AI4EIC) → need for fast simulations
- All Cherenkov detectors rely on pattern recognition of ring images in the reconstruction, which may become particularly complex like in the case of the DIRC → need to enhance reconstruction

Desiderata:

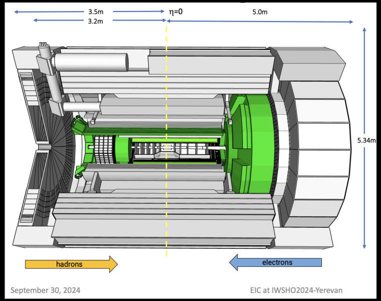
- Reconstruction at the “event-level” rather than “track-level” (*e.g.*, two tracks with overlapping patterns in the same optical box) — N.b. over 10% of SIDIS events involve at least two charged tracks with momenta above 1 GeV/c detected simultaneously in one sector of the hpDIRC
- Possibility of learning directly from real data the detector response.
- Faster algorithms to cope with near real-time analysis

Cherenkov Detectors: DIRC example

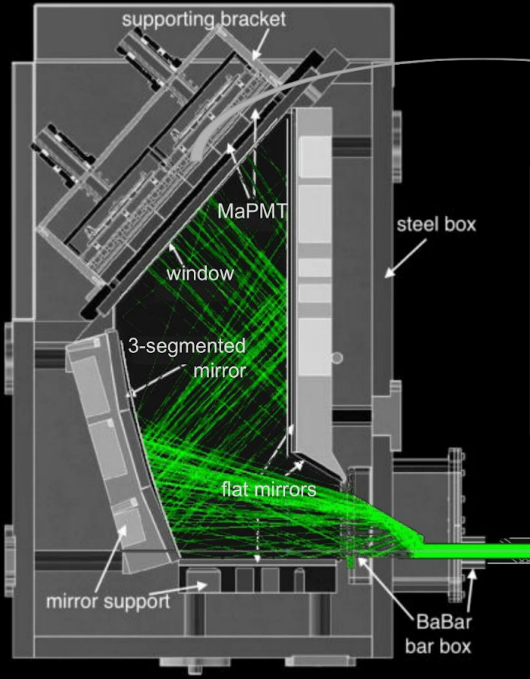
48 fused silica bars segmented into 4 bar boxes

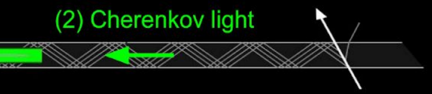
Two optical boxes, containing distilled water and highly reflective focusing mirrors

6 x 18 PMT (8 x 8 pixels) array for photon detection. Provides location and timing information for photons



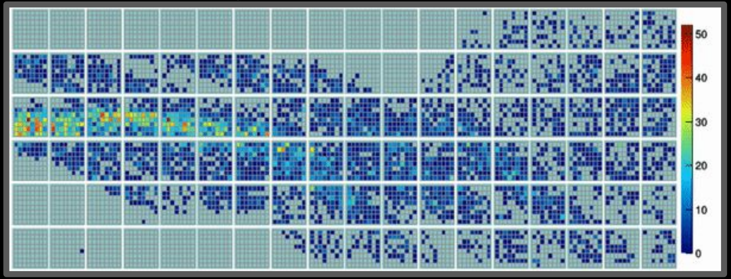
Schematic of GlueX DIRC detector

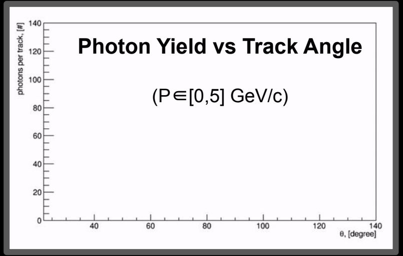




K^{+-} / π^{+-}

Patterns (collecting multiple particles) vs kinematics





Photon yield per particle (each particle produces a sparse hit pattern)

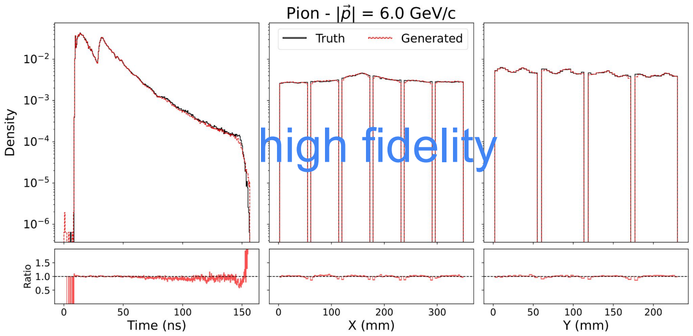
Goal: Characterize hit patterns from K^{+-} / π^{+-} as a function of $< |p| , \theta , \phi >$ (track)

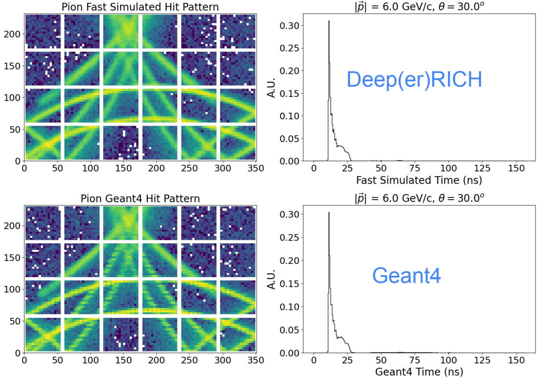
Deep(er)RICH: Fast Sim with NF

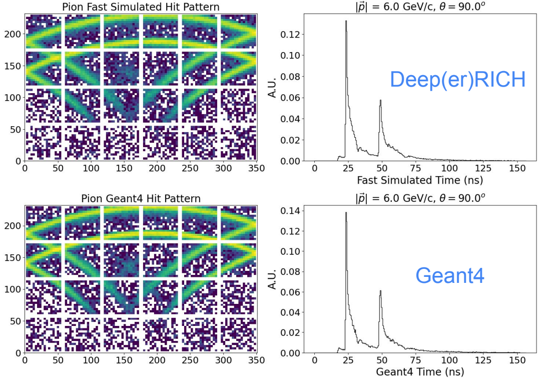
CF, J. Giroux, J. Stevens. “Deep(er)RICH"
Machine Learning: Science and Technology 6.1 (2025): 015028.

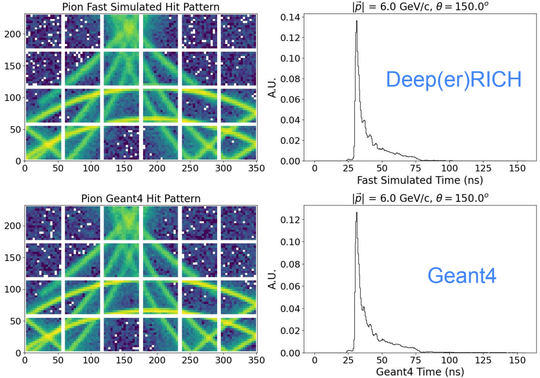
Architecture: Normalizing Flow (NF)

- **Density Transformation** – Define a bijective function and apply a change of variables, conditioning on kinematics parameters to maximize likelihood of expected hit pattern under a base distribution
- **Hit-Level Learning** – Model conditioned on kinematic parameters (|p|,θ,ϕ)
- **Agnostic to Photon Yield** – Ensure model independence from photon yield
- **Abstract away Fixed Input Size** – Address NF limitations with discrete distributions; data preprocessing transform DIRC readout (row, col) to (x,y) in mm and uniformly smear over PMT pixels







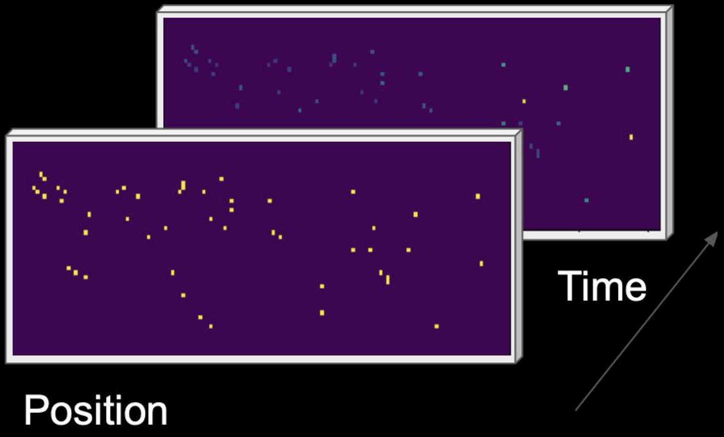


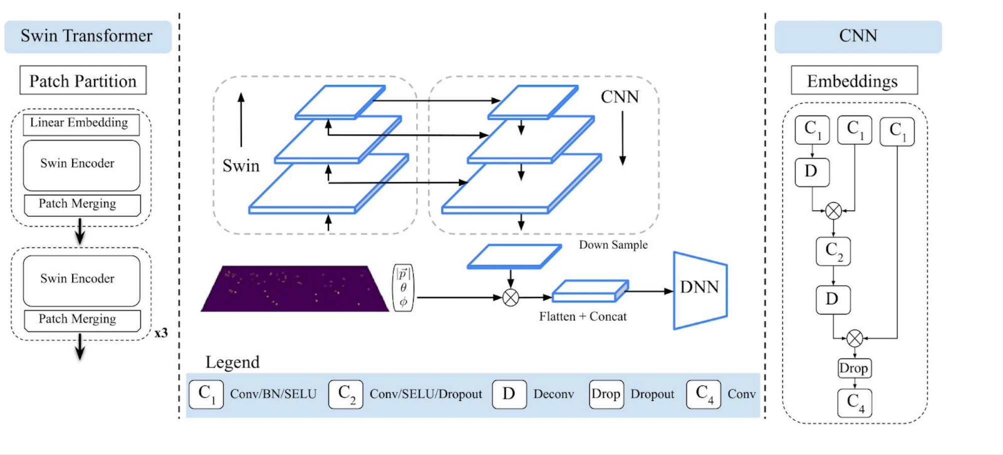
(hpDIRC standalone sim)

Simulation is fast - O(0.5)μs per hit (effective)

Deep(er)RICH: PID

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



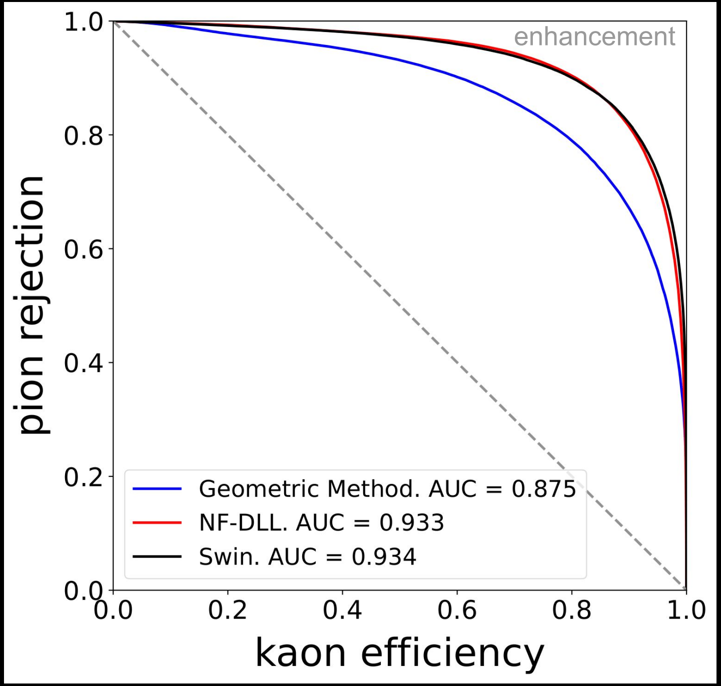


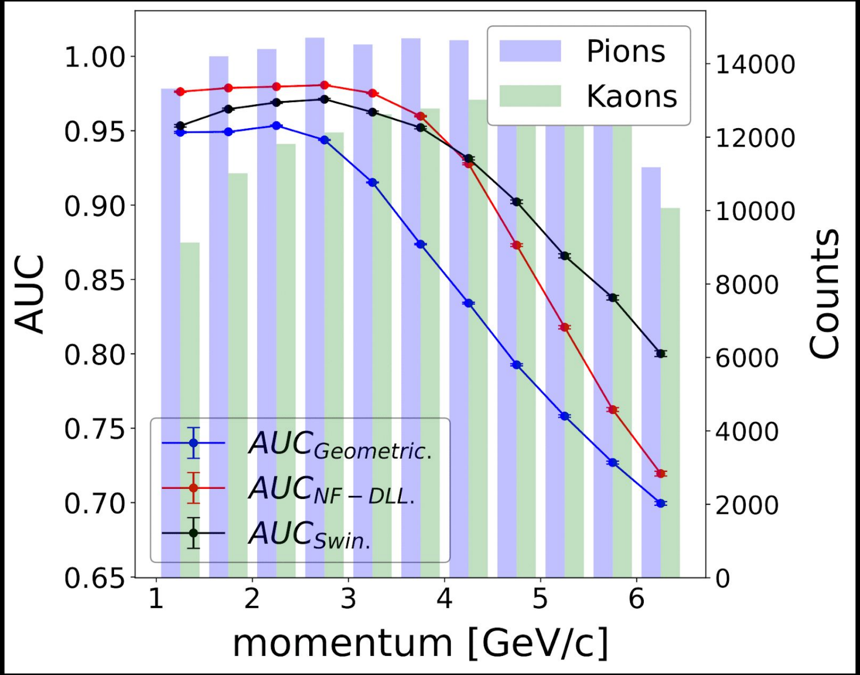
- Individual tracks do form “images” in optical boxes
  - Sparse point representations
- Possibility of overlapping hits
  - Same *x*,*y* - different times
  - Construct these as images as FIFO
  - Tends to be low percentage of overlap

- 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

Deep(er)RICH: PID

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





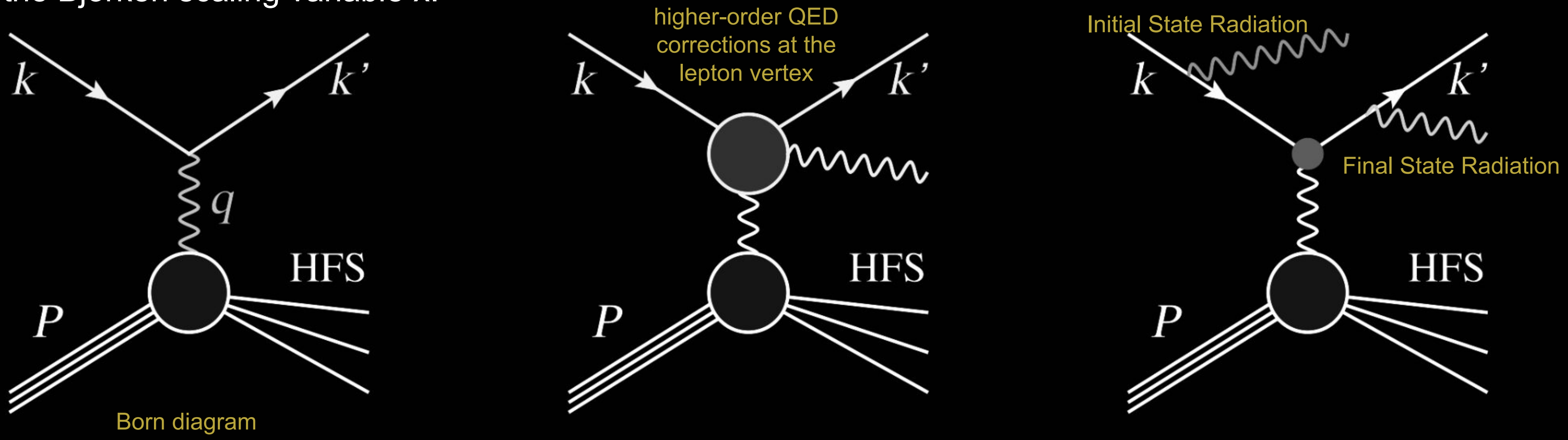
PID is fast - O(9)μs per track with transformer (effective)

Bonus: NF for PID. This method is slightly slower given additional computation needed

(GlueX DIRC sim)

Event-level Reco: DIS example

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 $Q^2 = s \cdot x\, y$, where s is the square of the center-of-mass energy.

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

DIS Kinematics

The expected collision rate for ep DIS 18X275GeV is 83 kHz. More info can be found at this link.

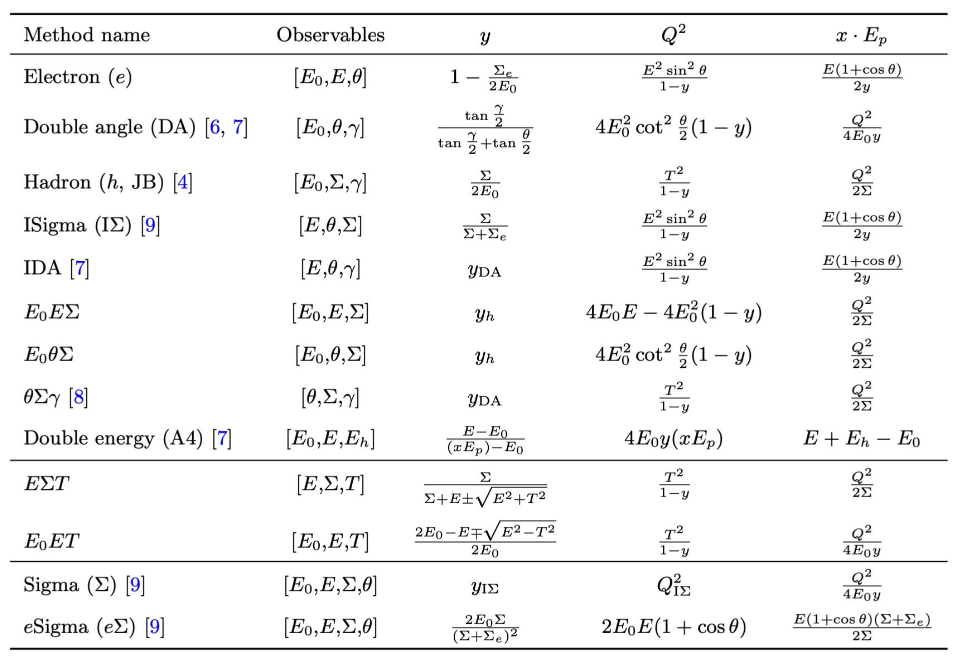
DIS Kinematics:Traditional Methods

- Conservation of momentum and energy over constrain the DIS kinematics and leads to a freedom to calculate x, Q^2, y from measured quantities

- Each method has advantages and disadvantages, and no single approach is optimal over the entire phase space. Each method exhibits different sensitivity to QED radiative effects

- Once (real) higher-order QED effects are considered, various methods yield different results and the calculated quantities for Q^2, y and x are not representative for the γ/Z + p scattering process at the hadronic vertex.

Summary of basic reconstruction methods



Method name	Observables	y	Q^2	$x \cdot E_p$
Electron (e)	$[E_0,E,\theta]$	$1-\frac{\Sigma_e}{2E_0}$	$\frac{E^2\sin^2\theta}{1-y}$	$\frac{E(1+\cos\theta)}{2y}$
Double angle (DA) [6, 7]	$[E_0,\theta,\gamma]$	$\frac{\tan\frac{\gamma}{2}}{\tan\frac{\gamma}{2}+\tan\frac{\theta}{2}}$	$4E_0^2\cot^2\frac{\theta}{2}(1-y)$	$\frac{Q^2}{4E_0y}$
Hadron (h, JB) [4]	$[E_0,\Sigma,\gamma]$	$\frac{\Sigma}{2E_0}$	$\frac{T^2}{1-y}$	$\frac{Q^2}{2\Sigma}$
ISigma (IΣ) [9]	$[E,\theta,\Sigma]$	$\frac{\Sigma}{\Sigma+\Sigma_e}$	$\frac{E^2\sin^2\theta}{1-y}$	$\frac{E(1+\cos\theta)}{2y}$
IDA [7]	$[E,\theta,\gamma]$	y_{DA}	$\frac{E^2\sin^2\theta}{1-y}$	$\frac{E(1+\cos\theta)}{2y}$
$E_0E\Sigma$	$[E_0,E,\Sigma]$	y_h	$4E_0E-4E_0^2(1-y)$	$\frac{Q^2}{2\Sigma}$
$E_0\theta\Sigma$	$[E_0,\theta,\Sigma]$	y_h	$4E_0^2\cot^2\frac{\theta}{2}(1-y)$	$\frac{Q^2}{2\Sigma}$
$\theta\Sigma\gamma$ [8]	$[\theta,\Sigma,\gamma]$	y_{DA}	$\frac{T^2}{1-y}$	$\frac{Q^2}{2\Sigma}$
Double energy (A4) [7]	$[E_0,E,E_h]$	$\frac{E-E_0}{(xE_p)-E_0}$	$4E_0y(xE_p)$	$E+E_h-E_0$
$E\Sigma T$	$[E,\Sigma,T]$	$\frac{\Sigma}{\Sigma+E\pm\sqrt{E^2+T^2}}$	$\frac{T^2}{1-y}$	$\frac{Q^2}{2\Sigma}$
E_0ET	$[E_0,E,T]$	$\frac{2E_0-E\mp\sqrt{E^2-T^2}}{2E_0}$	$\frac{T^2}{1-y}$	$\frac{Q^2}{4E_0y}$
Sigma (Σ) [9]	$[E_0,E,\Sigma,\theta]$	$y_{\mathrm{I}\Sigma}$	$Q^2_{\mathrm{I}\Sigma}$	$\frac{Q^2}{4E_0y}$
eSigma ($e\Sigma$) [9]	$[E_0,E,\Sigma,\theta]$	$\frac{2E_0\Sigma}{(\Sigma+\Sigma_e)^2}$	$2E_0E(1+\cos\theta)$	$\frac{E(1+\cos\theta)(\Sigma+\Sigma_e)}{2\Sigma}$

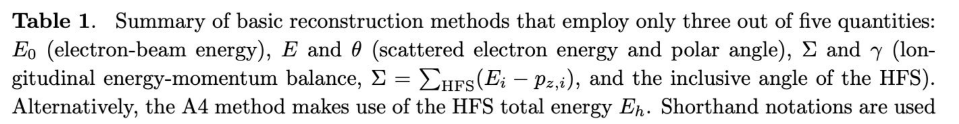
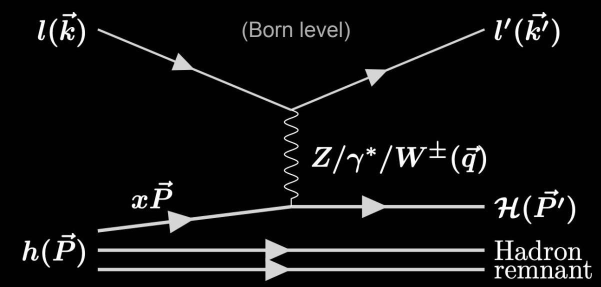


Table 1. Summary of basic reconstruction methods that employ only three out of five quantities: E_0 (electron-beam energy), E and θ (scattered electron energy and polar angle), Σ and γ (longitudinal energy-momentum balance, $\Sigma=\sum_{\mathrm{HFS}}(E_i-p_{z,i})$, and the inclusive angle of the HFS). Alternatively, the A4 method makes use of the HFS total energy E_h. Shorthand notations are used

Table taken from Arratia et al., NIM-A 1025 (2022): 166164

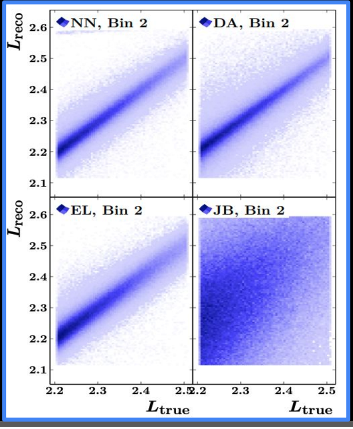
Deeply Learning DIS

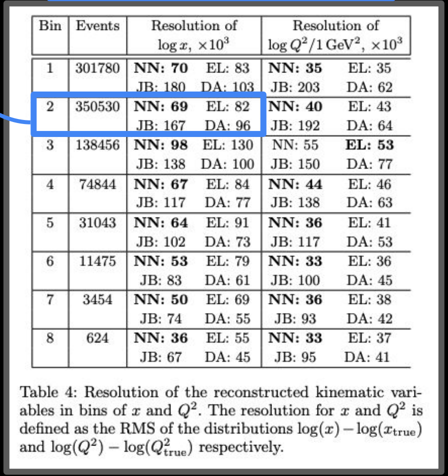
DIS fundamental
process @EIC



DIS beyond the Born approximation has a complicated structure which involve QCD and QED corrections

- Use of DNN to reconstruct the kinematic observable x, Q^2, y in the study of neutral current DIS events at ZEUS and H1 experiments at HERA.
- The performance compared to electron, Jacquet-Blondel and the double-angle methods using data-sets independent of training
- Compared to the classical reconstruction methods, the DNN-based approach enables significant improvements in the resolution of Q^2 and x





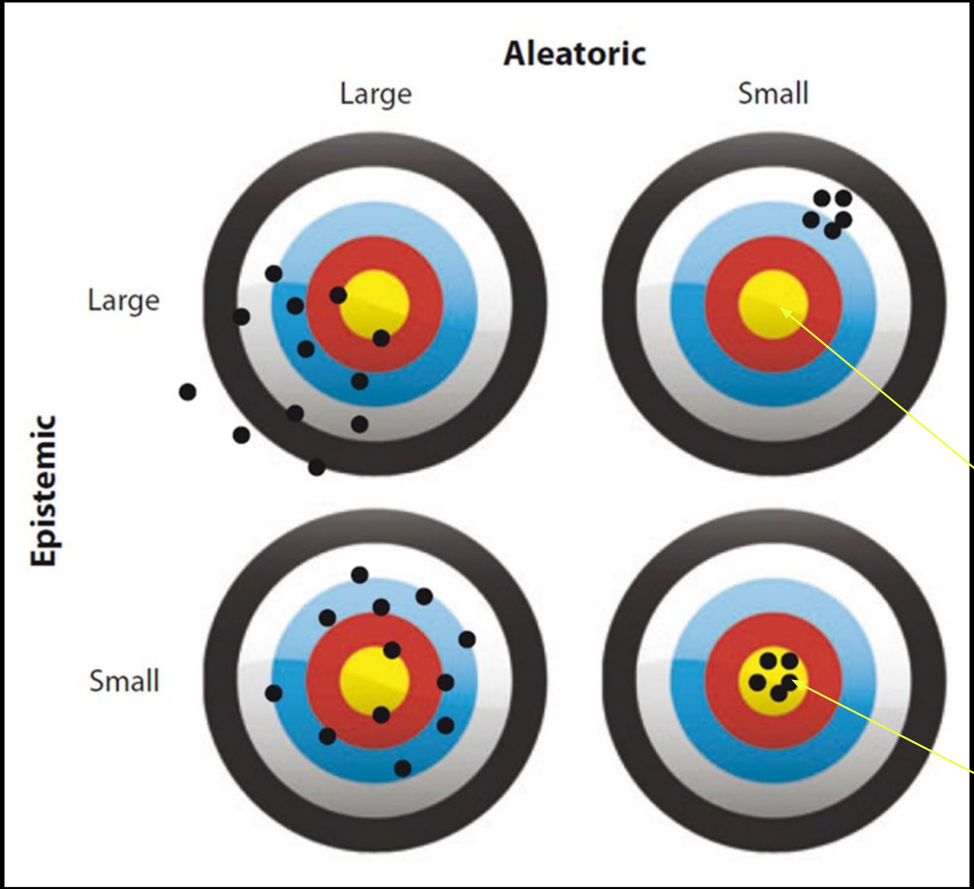
Bin	Events	Resolution of $\log x$, $\times 10^3$		Resolution of $\log Q^2/1\,\mathrm{GeV}^2$, $\times 10^3$	
1	301780	**NN: 70** JB: 180	EL: 83 DA: 103	**NN: 35** JB: 203	EL: 35 DA: 62
2	350530	**NN: 69** JB: 167	EL: 82 DA: 96	**NN: 40** JB: 192	EL: 43 DA: 64
3	138456	**NN: 98** JB: 138	EL: 130 DA: 100	NN: 55 JB: 150	**EL: 53** DA: 77
4	74844	**NN: 67** JB: 117	EL: 84 DA: 77	**NN: 44** JB: 138	EL: 46 DA: 63
5	31043	**NN: 64** JB: 102	EL: 91 DA: 73	**NN: 36** JB: 117	EL: 41 DA: 53
6	11475	**NN: 53** JB: 83	EL: 79 DA: 61	**NN: 33** JB: 100	EL: 36 DA: 45
7	3454	**NN: 50** JB: 74	EL: 69 DA: 55	**NN: 36** JB: 93	EL: 38 DA: 42
8	624	**NN: 36** JB: 67	EL: 55 DA: 45	**NN: 33** JB: 95	EL: 37 DA: 41

Table 4: Resolution of the reconstructed kinematic variables in bins of x and Q^2. The resolution for x and Q^2 is defined as the RMS of the distributions $\log(x) - \log(x_{\mathrm{true}})$ and $\log(Q^2) - \log(Q^2_{\mathrm{true}})$ respectively.

First application of DL for regression of DIS kinematics:

M. Diefenthaler, A. Farhat, A. Verbytskyi, Y Xu. "Deeply learning deep inelastic scattering kinematics." EPJ C 82.11 (2022): 1064.

Can we do regression with UQ?



Ground Truth (Nature)

Ideal case

Input Features

(used H1 MC sim dataset of NIM-A 1025: 166164*)

- Define variables to characterize the strength of QED radiation

$$p_T^{\mathrm{bal}} = 1 - \frac{p_{T,e}}{T} = 1 - \frac{\Sigma_e \tan\frac{\gamma}{2}}{\Sigma \tan\frac{\theta}{2}} \quad \text{and} \quad p_z^{\mathrm{bal}} = 1 - \frac{\Sigma_e + \Sigma}{2\,E_0}.$$

7 features to help indicate QED radiation in the event

- The values of p_T^{bal} and p_z^{bal}.
- The energy, η, and $\Delta\phi$ of the reconstructed photon in the event that is closest to the electron-beam direction, where $\Delta\phi$ is with respect to the scattered electron.
- The sum ECAL energy within a cone of $\Delta R < 0.4$ around the scattered electron divided by the scattered-electron track momentum.
- The number of ECAL clusters within a cone of $\Delta R < 0.4$ around the scattered electron.

+ additional 8 features

- Scattered-electron quantities $p_{T,e}$, $p_{z,e}$ and E.
- HFS four-vector quantities T, $p_{z,h}$ and E_h.
- $\Delta\phi(e,h)$ between the scattered electron and the HFS momentum vector.
- The difference $\Sigma_e - \Sigma$.

Tot. 15 input features

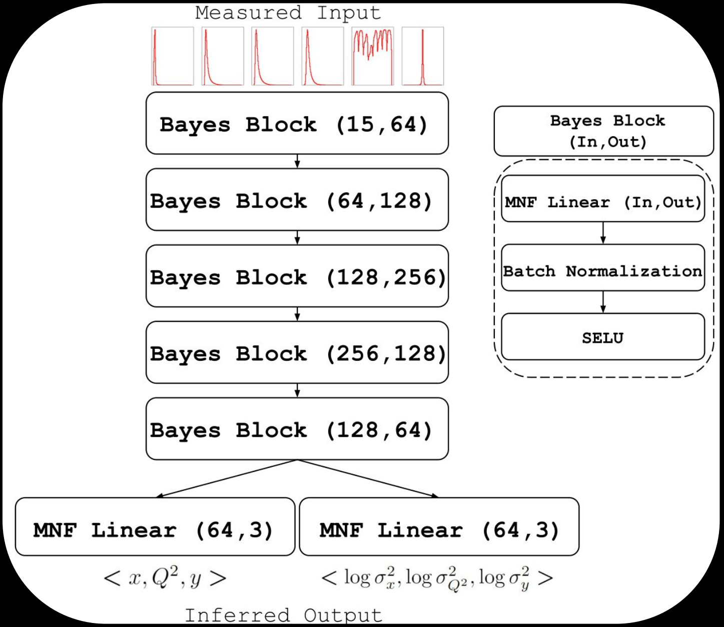
Dataset	Training Events	Validation Events	Testing Events	Size on Disk
H1	8.7×10^6	1.9×10^6	1.9×10^6	8 GB

*M. Arratia, D. Britzger, O. Long, B. Nachman, et al.,
“Reconstructing the kinematics of deep inelastic scattering with deep learning", NIM-A 1025 (2022): 166164

ELUQuant: Event-level UQ

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

Event-Level Uncertainty Quantification



Measured Input

Bayes Block (15,64)
Bayes Block (64,128)
Bayes Block (128,256)
Bayes Block (256,128)
Bayes Block (128,64)
MNF Linear (64,3)
MNF Linear (64,3)

$< x, Q^2, y >$

$< \log\sigma_x^2, \log\sigma_{Q^2}^2, \log\sigma_y^2 >$

Inferred Output

Bayes Block (In,Out)
MNF Linear (In,Out)
Batch Normalization
SELU

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

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

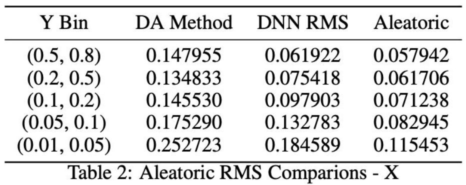
Access aleatoric as a function of regressed output [2]

Constrain the physics

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

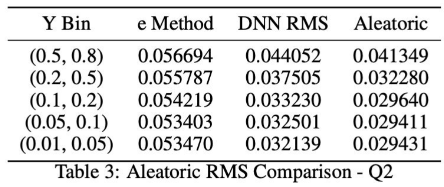
[1] C Louizos, M Welling International Conference on Machine Learning; arXiv:1703.01961 Multiplicative Normalizing Flows for Variational Bayesian Neural Networks
[2] A. Kendall and Y. Gal. "What uncertainties do we need in Bayesian deep learning for computer vision?." Adv. Neural Inf. Process. 30 (2017).

Aleatoric (ELUQ) vs RMS (other)



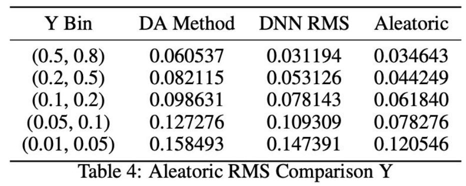
Y Bin	DA Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.147955	0.061922	0.057942
(0.2, 0.5)	0.134833	0.075418	0.061706
(0.1, 0.2)	0.145530	0.097903	0.071238
(0.05, 0.1)	0.175290	0.132783	0.082945
(0.01, 0.05)	0.252723	0.184589	0.115453

Table 2: Aleatoric RMS Comparions - X



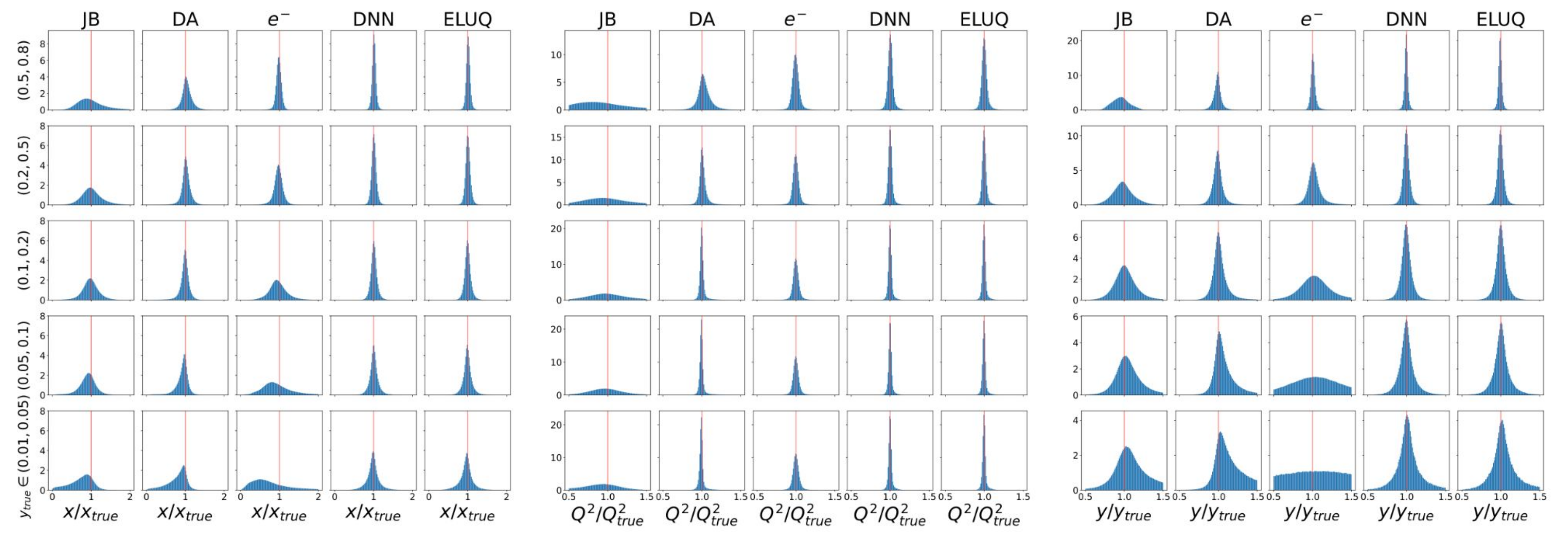
Y Bin	e Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.056694	0.044052	0.041349
(0.2, 0.5)	0.055787	0.037505	0.032280
(0.1, 0.2)	0.054219	0.033230	0.029640
(0.05, 0.1)	0.053403	0.032501	0.029411
(0.01, 0.05)	0.053470	0.032139	0.029431

Table 3: Aleatoric RMS Comparison - Q2

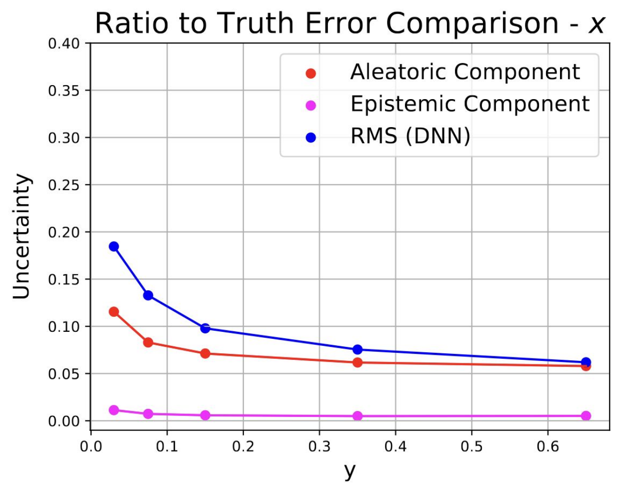


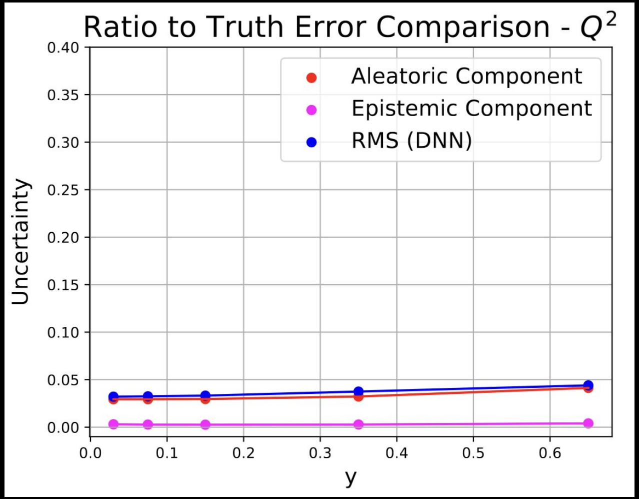
Y Bin	DA Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.060537	0.031194	0.034643
(0.2, 0.5)	0.082115	0.053126	0.044249
(0.1, 0.2)	0.098631	0.078143	0.061840
(0.05, 0.1)	0.127276	0.109309	0.078276
(0.01, 0.05)	0.158493	0.147391	0.120546

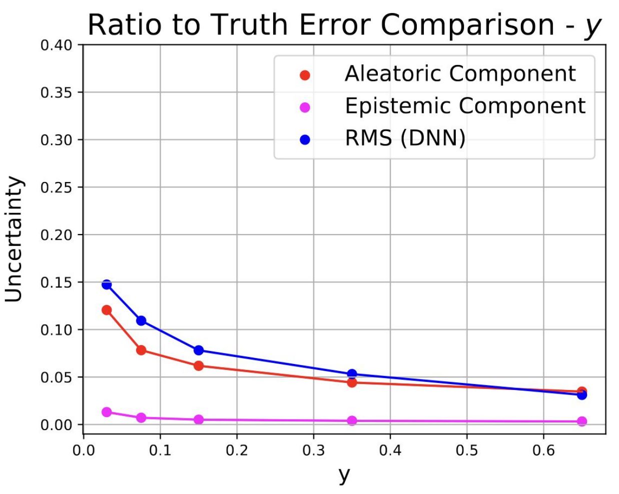
Table 4: Aleatoric RMS Comparison Y



ELUQ vs DNN

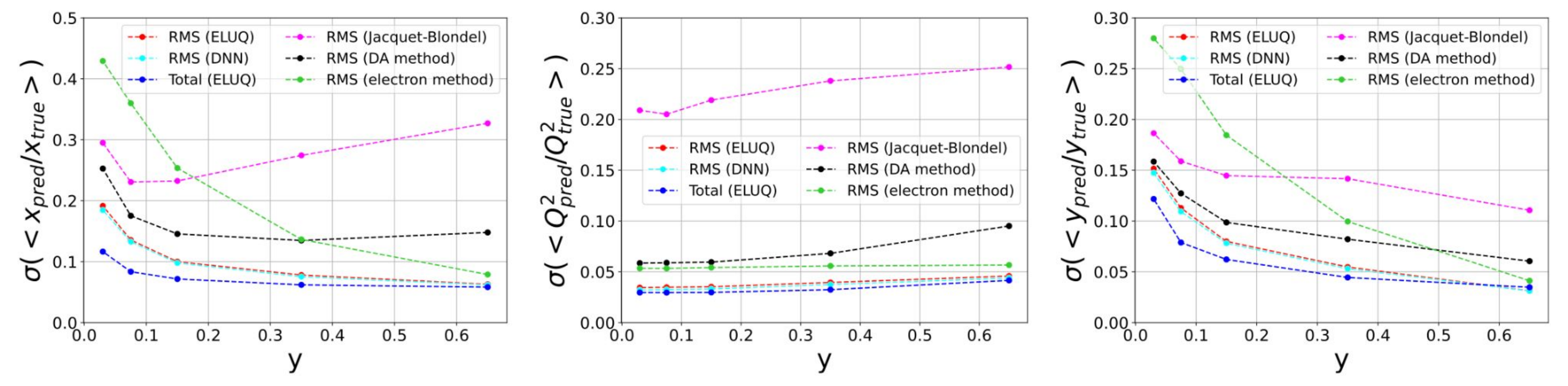

Ratio to Truth Error Comparison - x


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

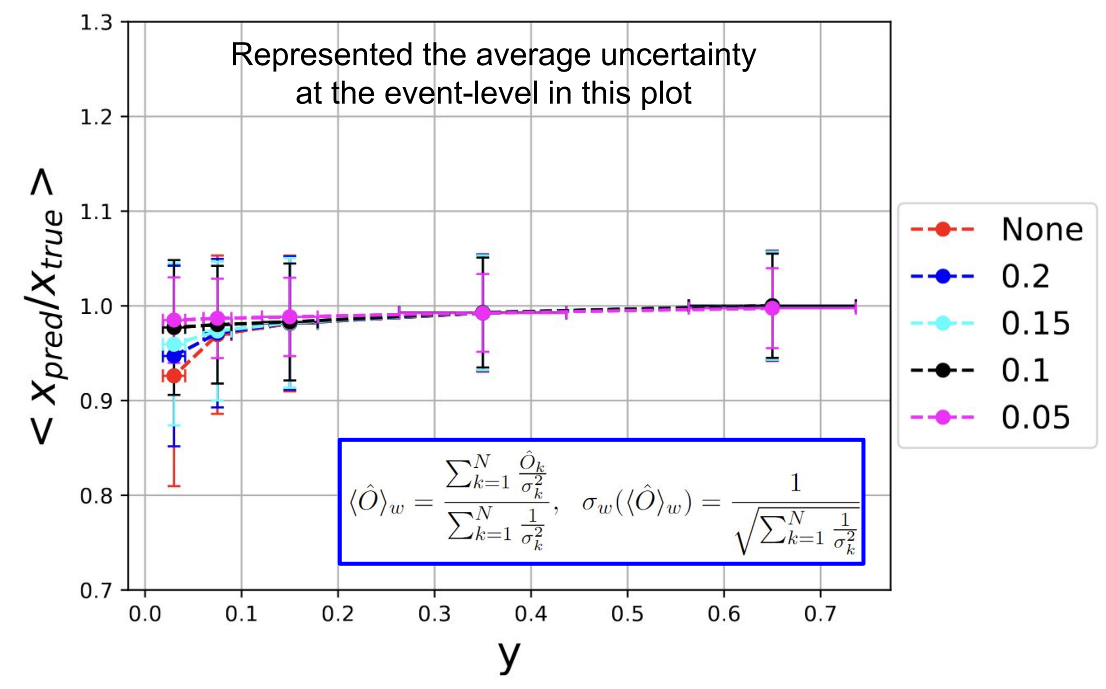
All Methods Compared



- 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*”

— Reporting uncertainty at the level of the event for ELUQuant; RMS from other methods and also for ELUQuant for comparison with DNN
—

Leveraging Event-Level Information



Represented the average uncertainty at the event-level in this plot

$$\langle \hat{O} \rangle_w = \frac{\sum_{k=1}^{N} \frac{\hat{O}_k}{\sigma_k^2}}{\sum_{k=1}^{N} \frac{1}{\sigma_k^2}}, \quad \sigma_w(\langle \hat{O} \rangle_w) = \frac{1}{\sqrt{\sum_{k=1}^{N} \frac{1}{\sigma_k^2}}}$$

- 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 —

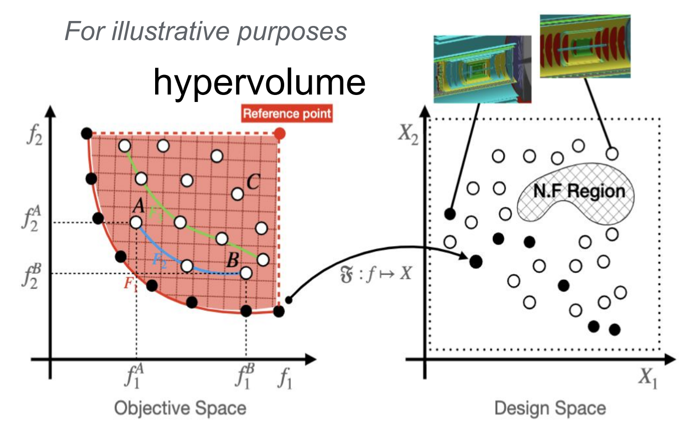
Multi-Objective Optimization

Hot take: every optimization problem is fundamentally a multi-objective optimization problem.

Maximizing Clarity

vs

Minimizing Talk Duration

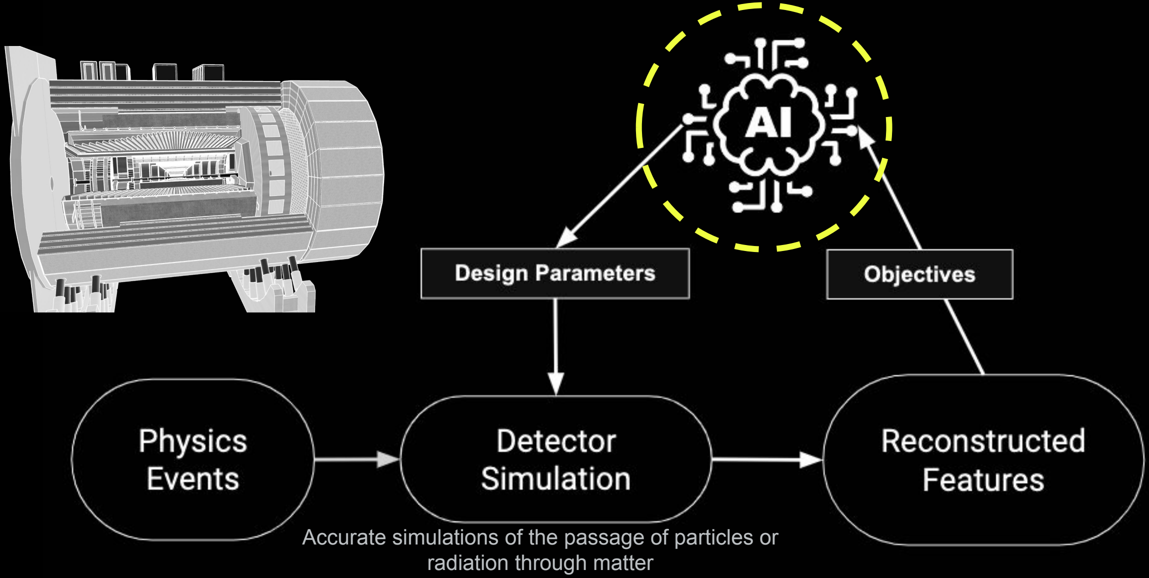


For illustrative purposes

hypervolume

AI-assisted Detector Design

Compute-intensive simulation pipelines



AI

Design Parameters

Objectives

Physics Events

Detector Simulation

Reconstructed Features

Accurate simulations of the passage of particles or radiation through matter

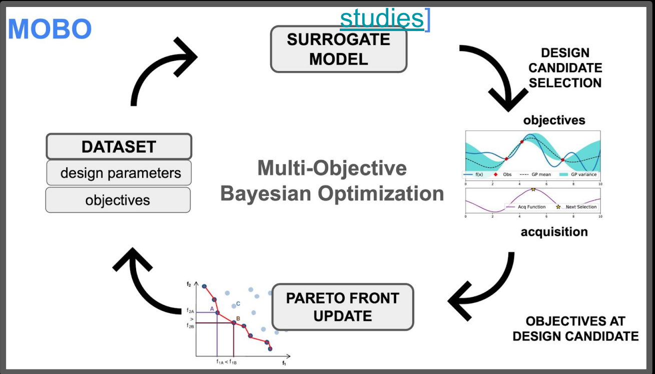
- Benefits from rapid turnaround time from simulations to analysis of high-level reconstructed observables
- The EIC SW stack offers multiple features that facilitate AI-assisted design (e.g., modularity of simulation, reconstruction, analysis, easy access to design parameters, automated checks, etc.)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

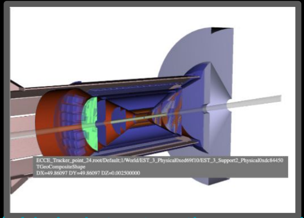
AI-assisted Detector Design @EIC

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

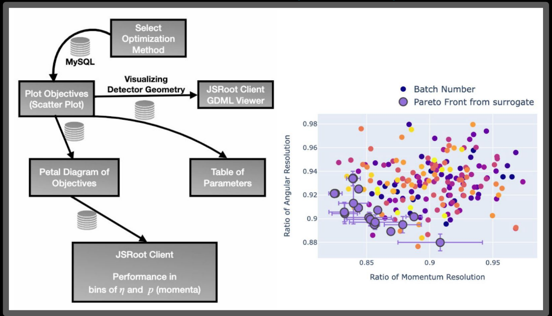
[Link to complexity studies]


MOBO
SURROGATE MODEL
DESIGN CANDIDATE SELECTION
objectives
acquisition
DATASET
design parameters
objectives
Multi-Objective Bayesian Optimization
PARETO FRONT UPDATE
OBJECTIVES AT DESIGN CANDIDATE

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

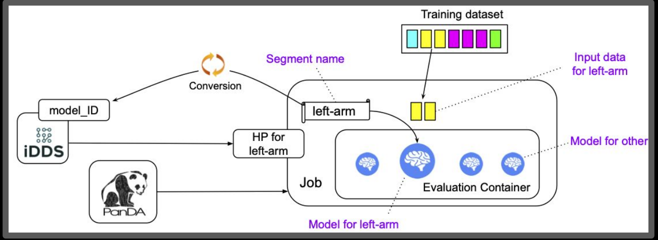


https://ai4eicdetopt.pythonanywhere.com/


Select Optimization Method
MySQL
Visualizing Detector Geometry
Plot Objectives (Scatter Plot)
JSRoot Client GDML Viewer
Petal Diagram of Objectives
Table of Parameters
JSRoot Client
Performance in bins of η and p (momenta)
Batch Number
Pareto Front from surrogate
Ratio of Angular Resolution
Ratio of Momentum Resolution
0.98
0.96
0.94
0.92
0.9
0.88
0.85
0.9
0.95

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

Distributed, scalable!



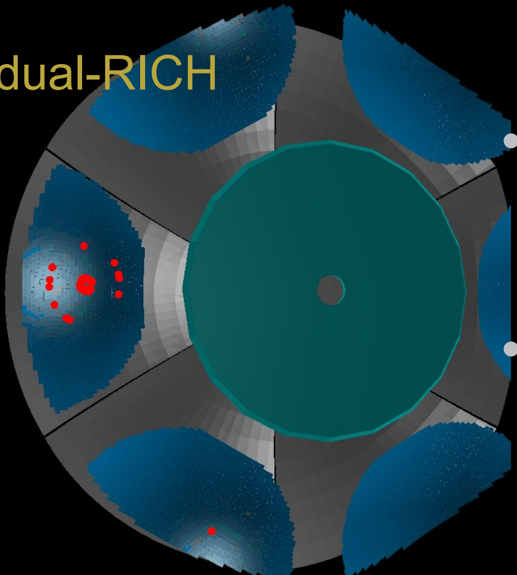
(iii) Leverage cutting-edge workload management systems capable of operating at massive data and handle complex workflows

- 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

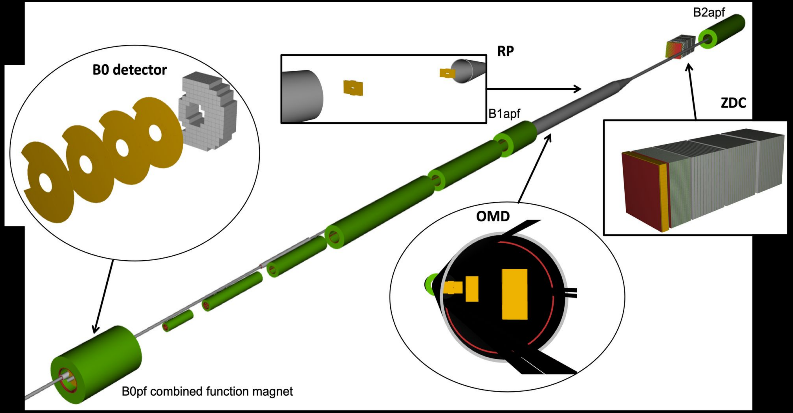


- *Mirror, sensor placement, gas, mirror material (lower costs material)...*
- *PID performance, costs, …*

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

Far-Forward

- *B0 magnetic field map, distance between layers, central location of tracker*
- *Momentum resolution, acceptance*

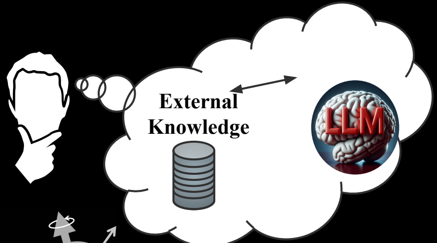


- AID2E framework can <u>support the design</u> of a possible detector-2; it is also being used for other compute-intensive tasks, such as <u>alignment</u> and <u>calibration</u>

AI Agents for EIC

https://rags4eic-ai4eic.streamlit.app/RAG-ChatBot

K. Suresh, N. Kackar, L Schleck, CF
"Towards a RAG-based summarization for the Electron Ion Collider." JINST 19.07 (2024): C07006.

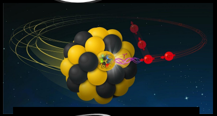


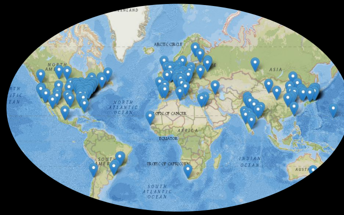
What is Retrieval Augmented Generation (RAG)?

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

Why need RAG for Large Scale Physics Experiments?

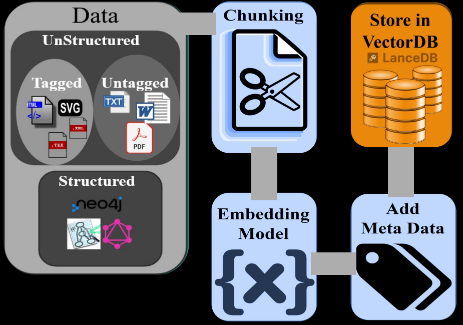
- EIC large scale experiment (e.g., EICUG ~1,400 users, ePIC 170+ institutions)
- Regular updates to documents, Run Wiki
- Newbies may take months to get to know the full experimental details.
- Tot document size approximately proportional to scale of experiment





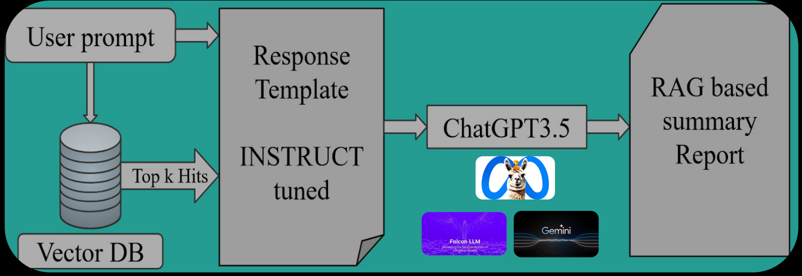
“Ingestion” of data

- Creation of the 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



AI Agents for EIC

Methods to evaluate RAG’s performance

- RAGAS score
- LLM as Judge[1]

1. Faithfulness –

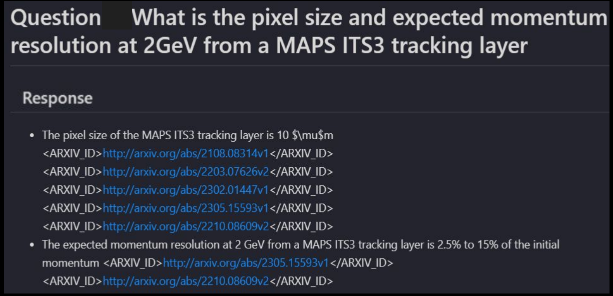
$$\text{Faithfulness score} = \frac{|\text{Number of claims that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

2. Answer relevance –

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

3. Context relevance –

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

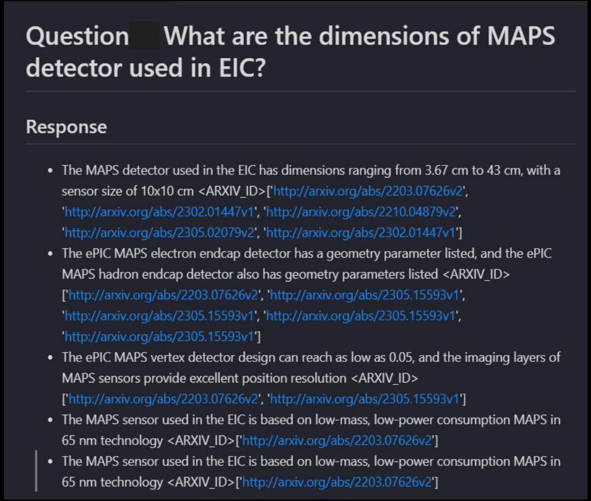


Question What is the pixel size and expected momentum resolution at 2GeV from a MAPS ITS3 tracking layer

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

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…

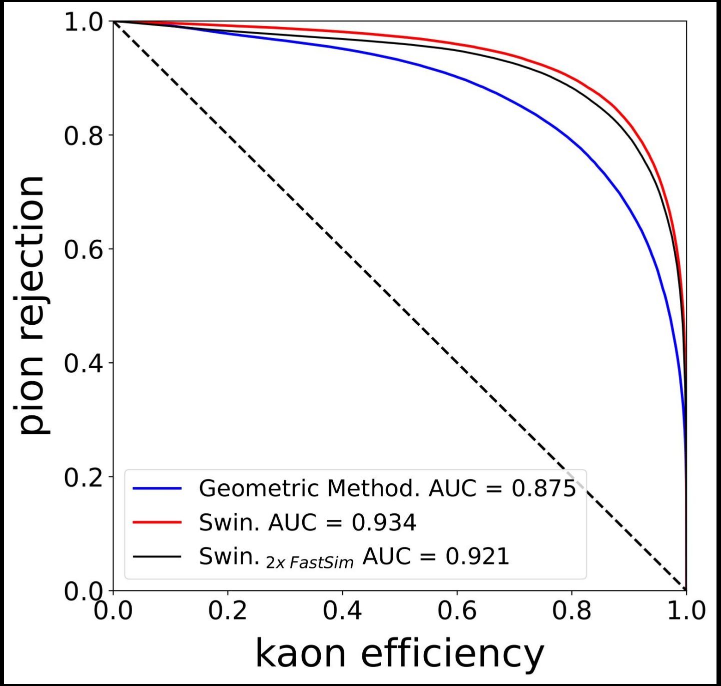
Conclusions

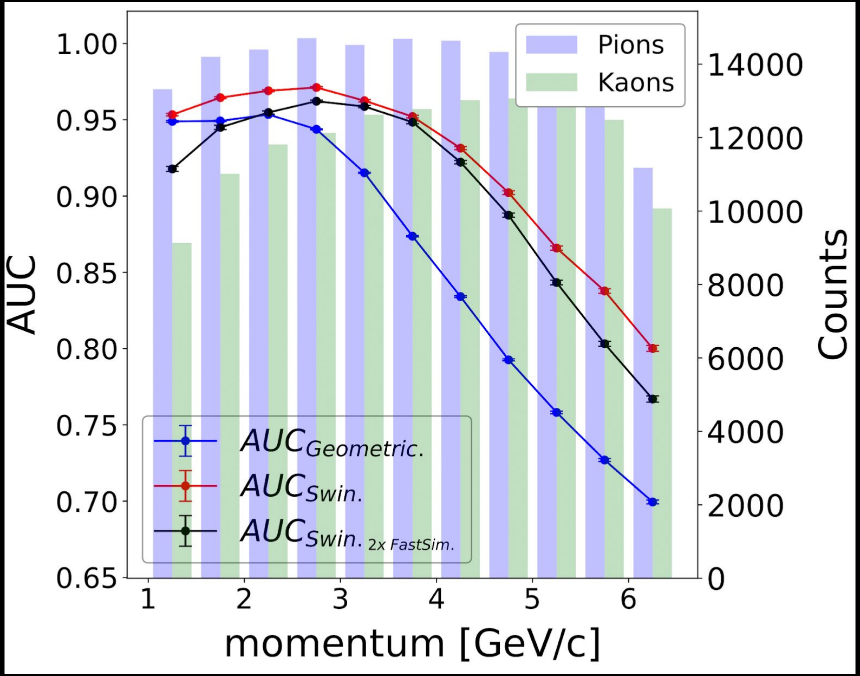
- AI/ML Integration: AI/ML will be integrated across all phases of EIC's science.
- Streaming Computing Model: The EIC community is developing a streaming computing model to enable near real-time AI/ML analysis, promising substantial advancements in live data processing.
- AI/ML will be key for data processing and analysis at EIC:
    - E.g., Cherenkov detectors are extremely important at EIC and present unique computational challenges; Newly developed methods enable generalization across continuous phase spaces and direct learning from real data.
        - Transformers: Delivers fast inference, approximately 9 μs/track, facilitating complex event-level topology learning. This is particularly valuable for critical physics channels like SIDIS
        - Normalizing Flows: Delivers high-fidelity, hit-level simulations conditioned on charged track kinematics, effectively overcoming major computational bottlenecks in simulations.
- Uncertainty Quantification at the event-level: Essential for precision experiments at EIC; leveraging advanced simulations including radiative corrections, we can use deep learning to improve both kinematic reconstruction and UQ at the event level
- Large-scale experiments leveraging AI-assisted Design and Optimization: Experiments at EIC aim to be among the first to be realized with the assistance of AI (this was already done during the proto-collaboration phase). The AID2E project is realizing a distributed framework that can optimize holistically large-scale detectors. The Detector-2 can be an ideal candidate. AID2E will be an ideal tool to optimize design changes with objectives (e.g., reduce cost). Can be also used for alignment and calibration tasks.
- Towards Multimodal Agents: Imagine you're an EIC owl shift-taker doing some complex task. To resolve it efficiently, you need to retrieve relevant information while analyzing available plots. A multimodal AI agent can assist in interpreting visual data, cross-referencing documentation (logbooks, wiki, etc), and providing actionable insights in near real-time…

Backup

The road to EIC, as seen from South Florida...

Fast Sim: Closure Tests





<u>Here **trained on fast simulation**</u>
2x size of original dataset (full sim)

Tested on full simulations

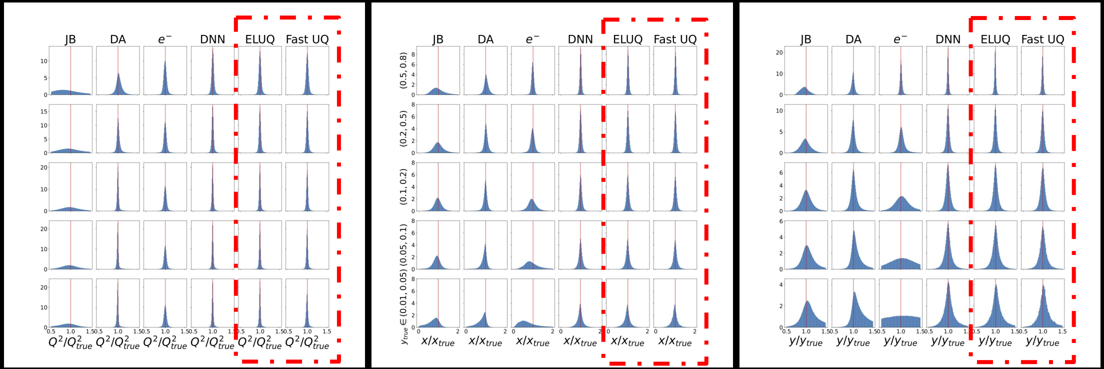
ELUQuant Time Performance

(2)

Inference Parameter	value
Number of Samples (N)	10k
Batch Size	100
Inference GPU Memory	$\sim$ 24GB
Inference Time per Event	$\sim 20ms$

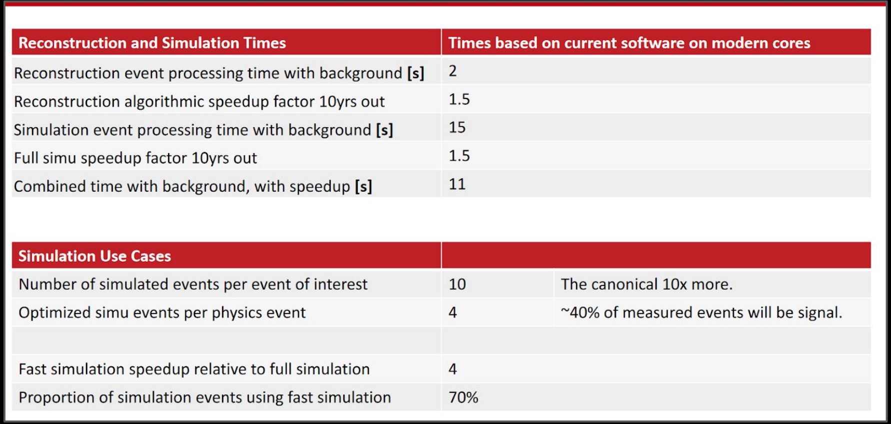
Training Parameter	value
Max Epochs	100
Batch Size	1024
Decay Steps	50
Decay Factor (γ)	0.1
Physics Loss Scale (α)	1.0
KL Scale (β)	0.01
Training GPU Memory	$\sim$ 1GB
Network memory on local storage	$\sim$ 7MB
Trainable parameters	611,247
Wall Time	$\sim$ 1 Day

- In computational terms, ELUQuant at inference showed an impressive rate of 10,000 samples/event within a 20 milliseconds on an RTX 3090.
- Can we do faster than this?
  - Several ways. A quick and dirty approach is distilling this knowledge in a simpler but faster network (we explored a DNN with 450k parameters) called in the following "Fast UQ", obtaining an effective inference time of 7-8us/event using batch ~0.5M events



Simulations Campaign (1)

- Large simulation campaigns needed since proto-collaboration phase (the selected “ECCE detector” concept utilized AI-assistance), where we adopted solutions with containerized software with distribution over the OSG — NIM-A: 1047 (2023):167859 (ECCE Computing model)
- This typically entails a large volume of events which are simulated for any given design of the detector (“design point”); i.e., in principle, 1 design point => 1 simulation campaign
- Therefore, and in general, more simulations needed to explore multiple design points
- Current simulation campaigns produce up to 15-20 TB / month (T. Britton, Oct 2024)
- Towards a quantitative computing model (M. Diefenthaler, Sep 2024)



Reconstruction and Simulation Times	Times based on current software on modern cores
Reconstruction event processing time with background **[s]**	2
Reconstruction algorithmic speedup factor 10yrs out	1.5
Simulation event processing time with background **[s]**	15
Full simu speedup factor 10yrs out	1.5
Combined time with background, with speedup **[s]**	11

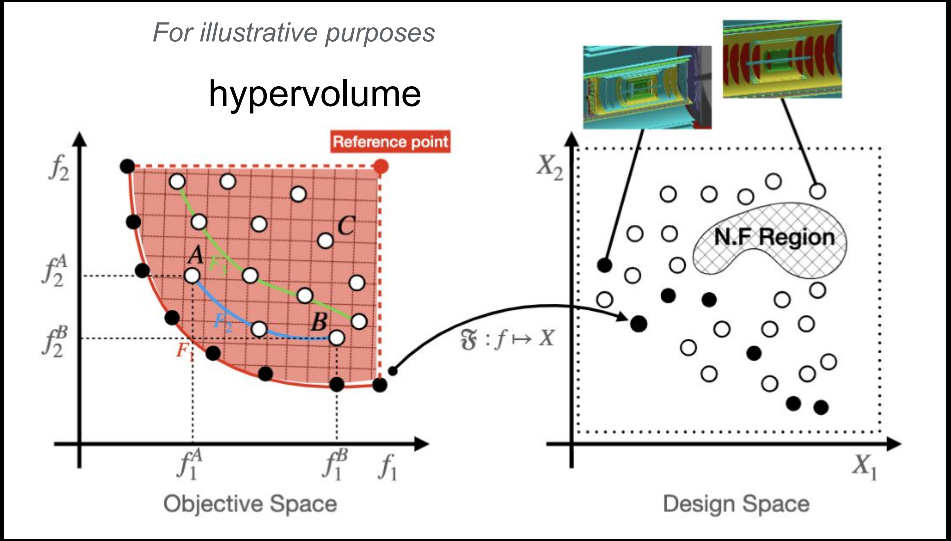
Simulation Use Cases		
Number of simulated events per event of interest	10	The canonical 10x more.
Optimized simu events per physics event	4	~40% of measured events will be signal.
Fast simulation speedup relative to full simulation	4	
Proportion of simulation events using fast simulation	70%	

500k tracks for the tracker and PID would require O(1.5k) cpu core hours including the entire simulation pipeline

Multi-Objective Optimization (1)

MOO is needed to optimize a system of sub-detectors

- 3 Types of Objectives
    - **Intrinsic detector performance** (resolutions, efficiencies) for each sub-detector — Tracking, calorimetry, PID — noisy
    - **Physics-performance** — Multiple physics channels, equally important in the EIC physics program
    - **Costs** (e.g., material costs, provided a reliable parametrization)
- Objectives can be competing with each other
    - E.g. Better detector response come with higher costs; better resolutions may imply lower efficiencies; etc.



For illustrative purposes

MOBO

n: number of design points
d: design dimensionality (each point)
M: objectives

Gaussian Process $O(n^3)$

- Surrogate model.
- SAAS[1] priors have been proven to be successful up to 388 design dimensions
- Assumes several design variables has increased importance compared to others
- Computational expensive as iteration increases
- Benefit from GPU hardware acceleration

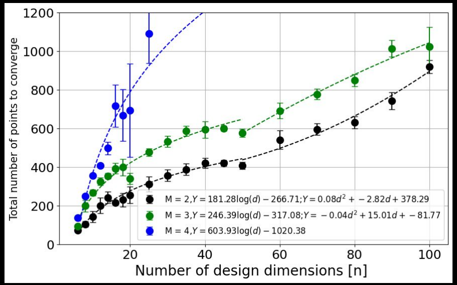
Bayesian Sampling from posteriors NUTS – $O(Md^{5/4})$[NUTS]

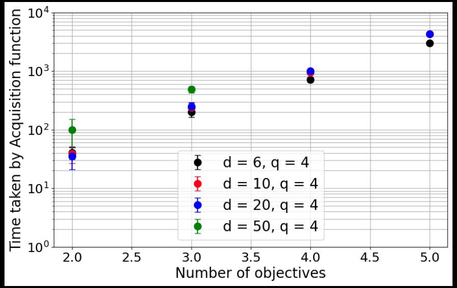
- Sample L points from the posterior distribution
- HMC is a popular algorithm, NUTS is a variant
- Mainly depends on the number of objectives and design space dimensions
- Has minimal dependence on iteration.
- GPU acceleration through JAX backend.

Acquisition function qNEHVI – $O(Md(n+i)^M)$[2]

- Captures HV improvement
- A "cheaper" function to evaluate as a proxy for the black box function
- Scales nonlinearly with iteration, total points explored, design space and objective space.
- Partially benefitted by GPU acceleration.

Complexity Studies



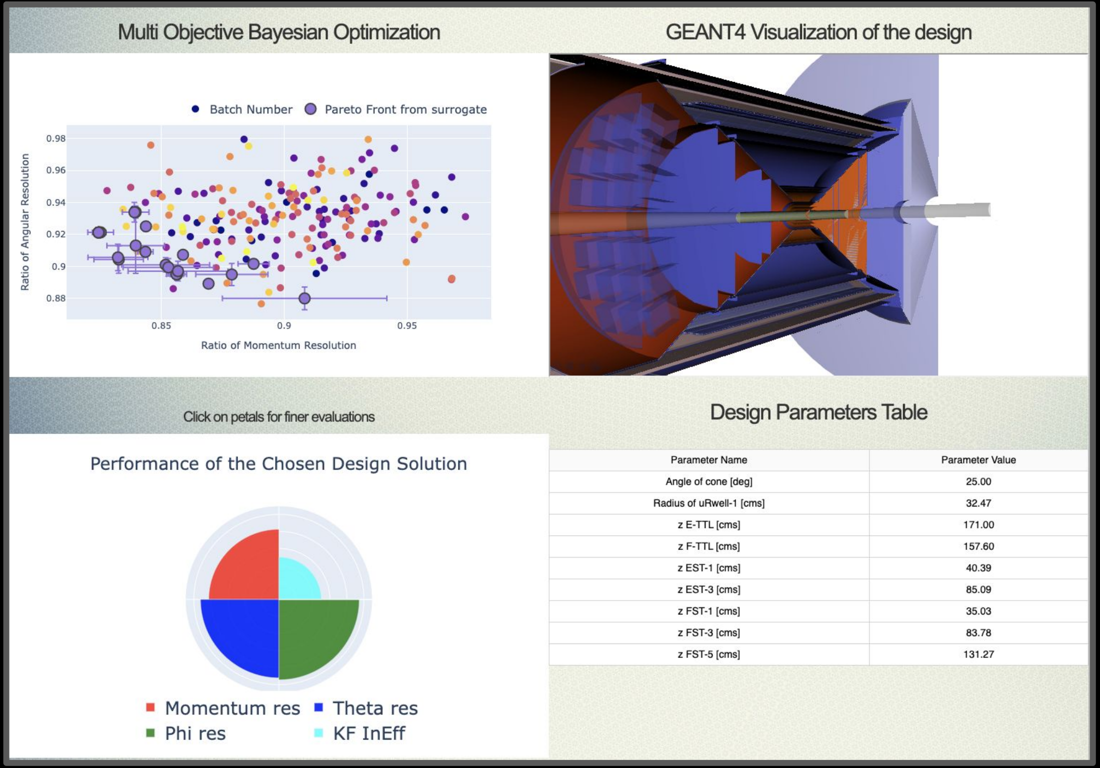


q: batch size

- Benefitting from GPU acceleration
- With sufficient parallelization, if possible, the time associated to the MOBO part at some point becomes dominant (bottom plot shown at 15th iteration with number of points between ~70-160)

Interactively Navigate Pareto (1)

C.Fanelli et al, NIM A, 2023, 167748



Multi Objective Bayesian Optimization

GEANT4 Visualization of the design

Click on petals for finer evaluations

Performance of the Chosen Design Solution

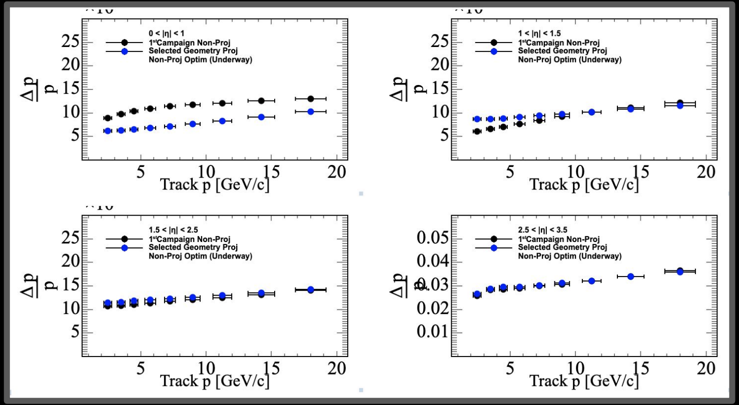
Design Parameters Table

Parameter Name	Parameter Value
Angle of cone [deg]	25.00
Radius of uRwell-1 [cms]	32.47
z E-TTL [cms]	171.00
z F-TTL [cms]	157.60
z EST-1 [cms]	40.39
z EST-3 [cms]	85.09
z FST-1 [cms]	35.03
z FST-3 [cms]	83.78
z FST-5 [cms]	131.27

● Batch Number
● Pareto Front from surrogate
Ratio of Angular Resolution
0.98
0.96
0.94
0.92
0.9
0.88
0.85
0.9
0.95
Ratio of Momentum Resolution
Momentum res
Theta res
Phi res
KF InEff

The whole idea of the AI-assisted design is that of determining trade-off optimal solutions in a multidimensional design space driven by multiple objectives

For an interactive visualization: https://ai4eicdetopt.pythonanywhere.com



0 < |η| < 1
1st Campaign Non-Proj
Selected Geometry Proj
Non-Proj Optim (Underway)
1 < |η| < 1.5
1.5 < |η| < 2.5
2.5 < |η| < 3.5
Δp/p
Track p [GeV/c]
25
20
15
10
5
0.05
0.04
0.03
0.02
0.01