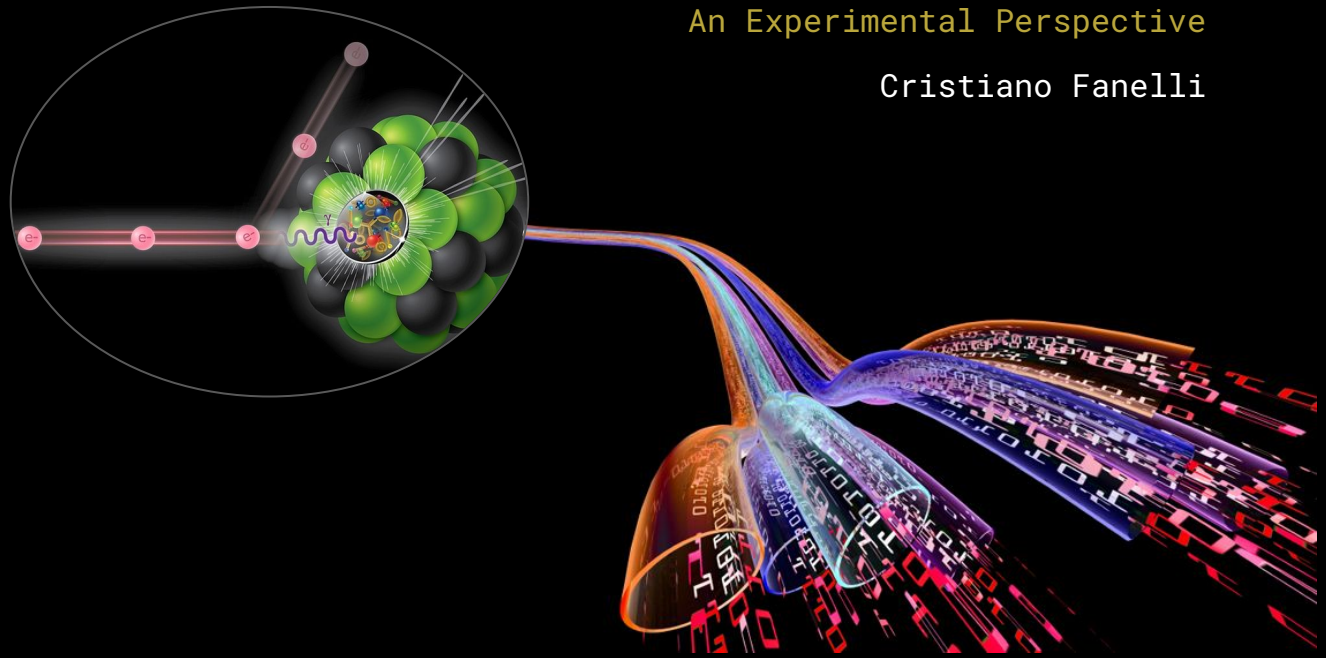
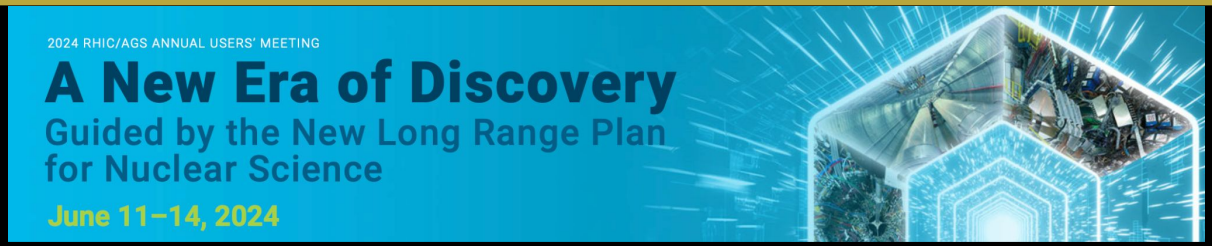
AI/ML Applications for EIC

An Experimental Perspective

Cristiano Fanelli



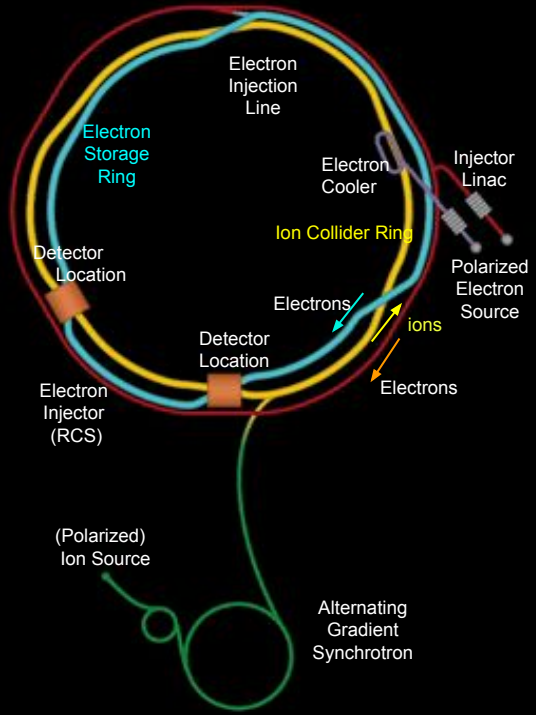


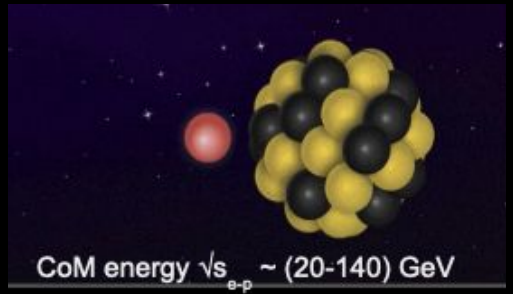
WILLIAM & MARY
CHARTERED 1693

Electron Ion Collider

A US-led and international effort to build the ultimate precision machine to study the “glue” that binds us all

polarized electron - polarized protons/ions





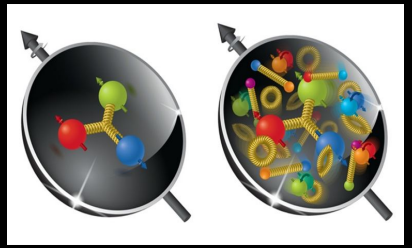
World-wide interest, thousands of users and hundreds of institutions already involved

Total estimated cost ~ $1.6-2.6B

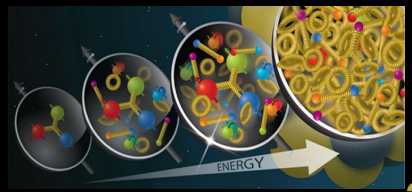
3 fundamental questions

How does the mass of the nucleon arise?

How does the spin of the nucleon arise?

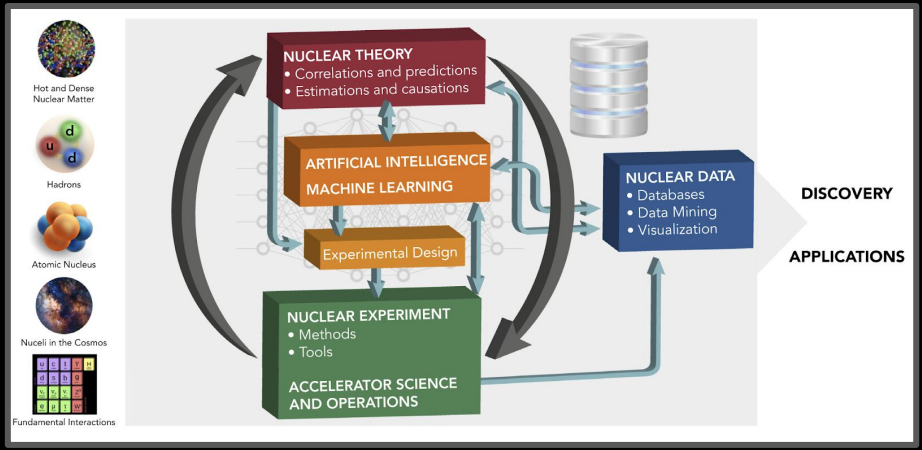


What are the emergent properties of dense systems of gluons?



AI/ML at the Electron Ion Collider

- AI/ML is becoming ubiquitous in NP



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

- Next generation QCD experiment like EIC have the opportunity to integrate AI/ML from the initial phases of their development

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

Community perspective:

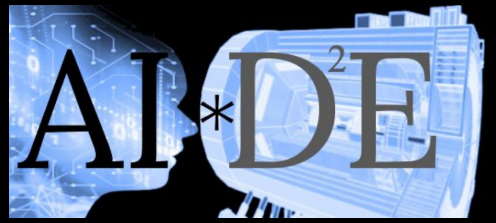
Some recommendations from the NSAC Long-Range Plan Town Hall Meeting on Hot and Cold QCD (MIT, 2022):

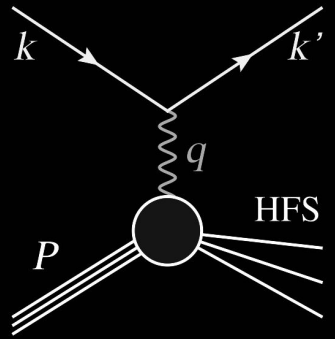
- The completion of the EIC is recommended as the highest priority for facility construction
- Increased investments in computational nuclear physics, AI/ML, HPC, HTC, data systems, and interdisciplinary workforce development, are essential for advancing nuclear physics.

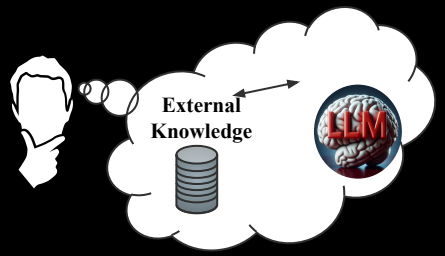
https://indico.mit.edu/event/538/
The Present and Future of QCD, Nucl.Phys.A 1047 (2024): 122874.

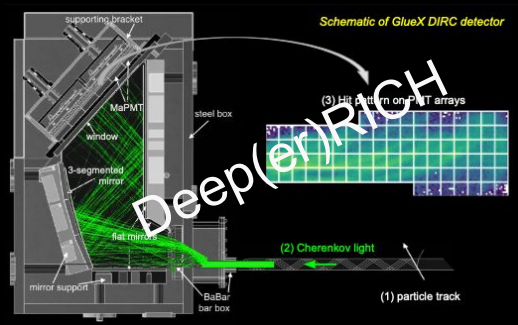
Outline

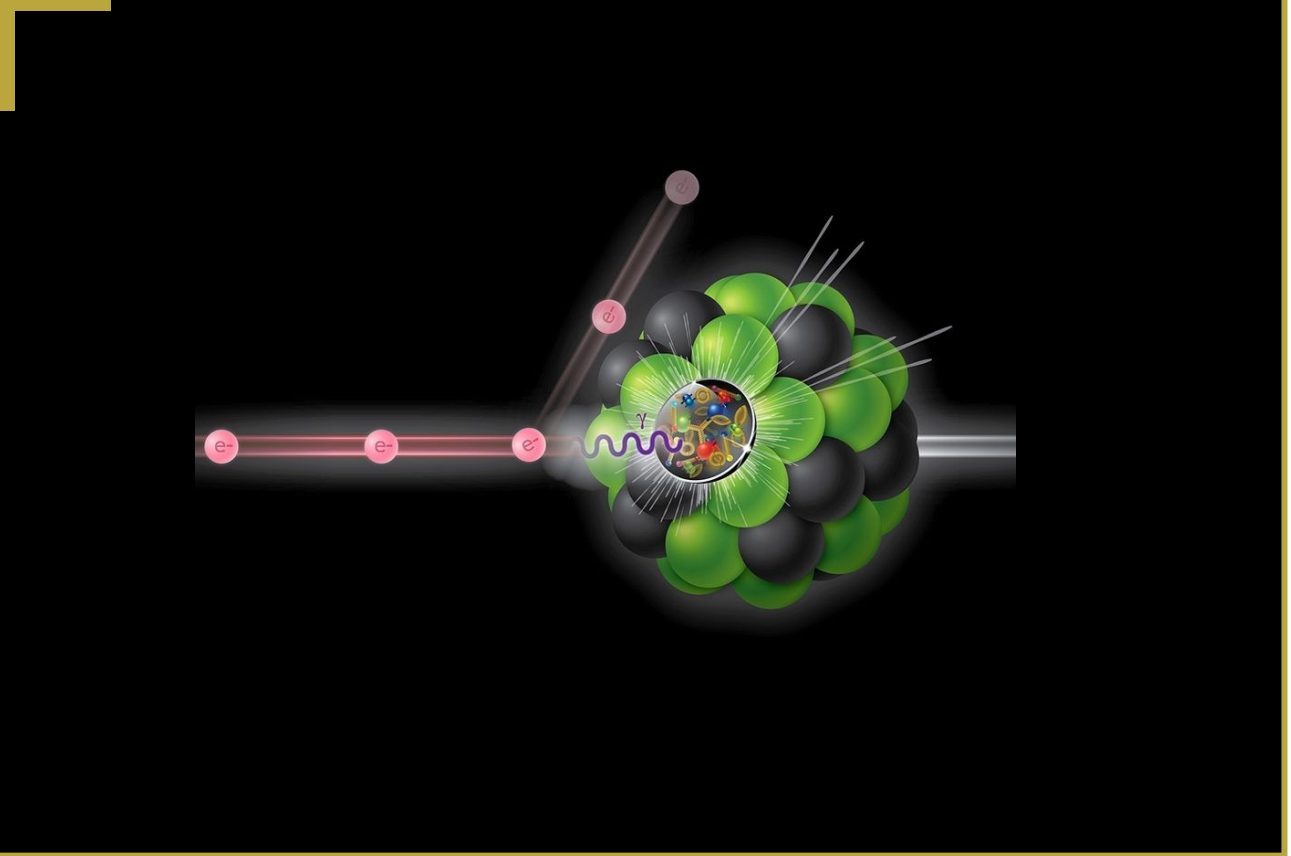
- Multiple ongoing AI/ML activities for the EIC
  - I will focus on the following examples:
    - AI-assisted design for EIC —> AID2E
    - Reconstruction of Deep Inelastic Scattering —> ELUQuant
    - Particle Identification —> Deep(er)RICH
    - AI-assistants for the EIC —> RAG-based

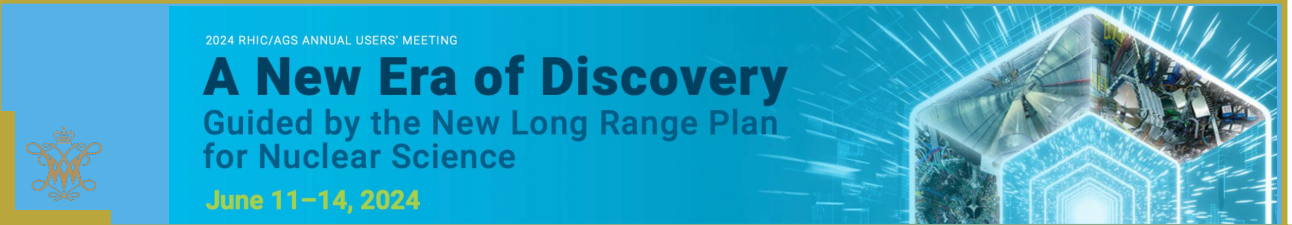










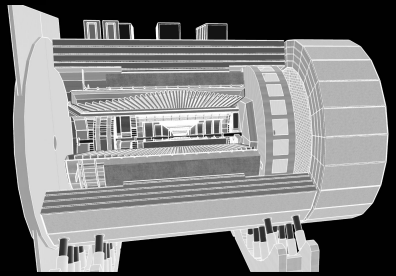


AID2E

AI-assisted Detector Design at EIC

AI-Assisted Detector Design

Experimental design embraces all the the main steps of the simulation pipeline…



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

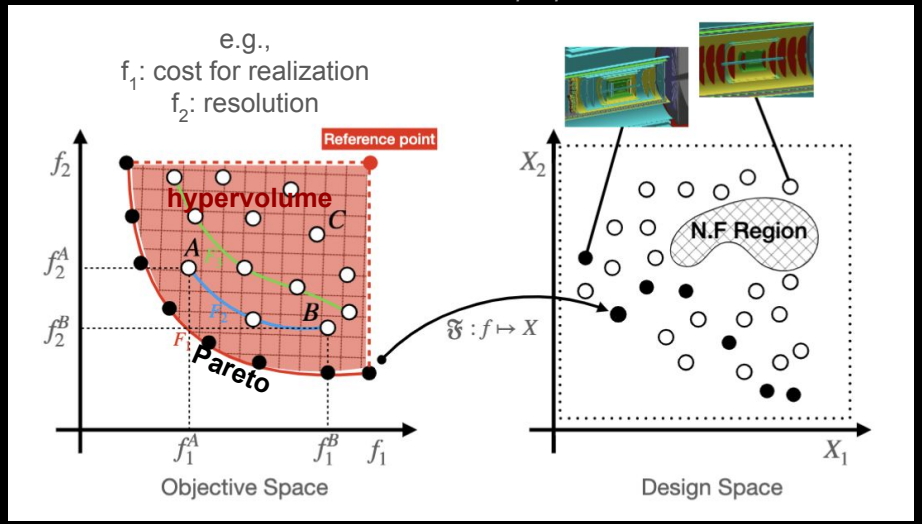
Those at EIC can be the first large-scale experiments ever realized with the assistance of AI

AI-Assisted Detector Design

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

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

For illustrative purposes



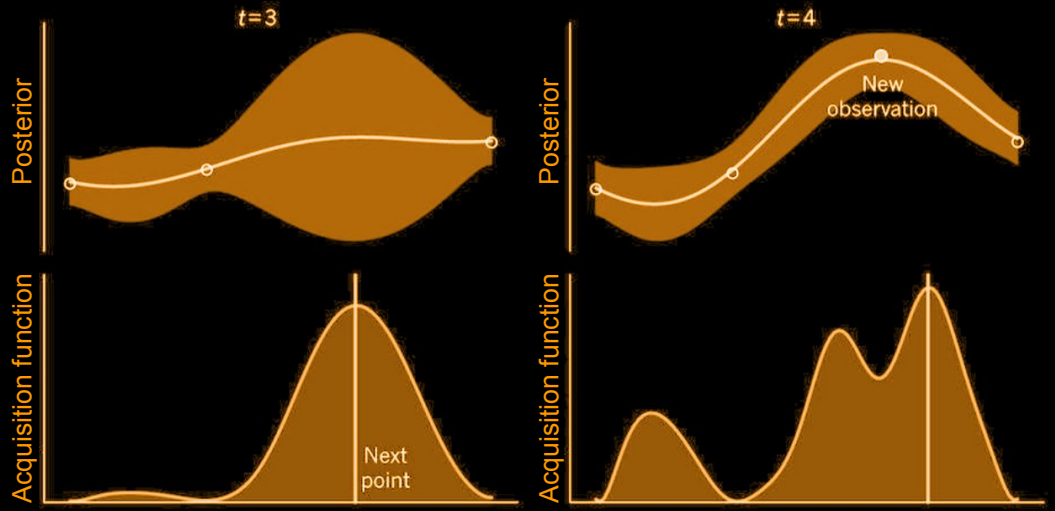
Goal: Determine Pareto front (trade off solutions)

- Examining solutions on the Pareto front of EIC detectors at different values of the budget can have great cost benefits.
- A fractional improvement in objectives leads to more efficient beam time use, significantly reducing the lifetime costs of the EIC.

Bayesian Optimization

- BO is a sequential strategy developed for global optimization.
- After gathering evaluations we builds a posterior distribution used to construct an **acquisition function**.
- This cheap function determines what is **next query point**.

With a single objective, for illustration



1. Select a Sample by Optimizing the Acquisition Function.
2. Evaluate the Sample With the Objective Function.
3. Update the Data and, in turn, the Surrogate Function.
4. Go To 1.

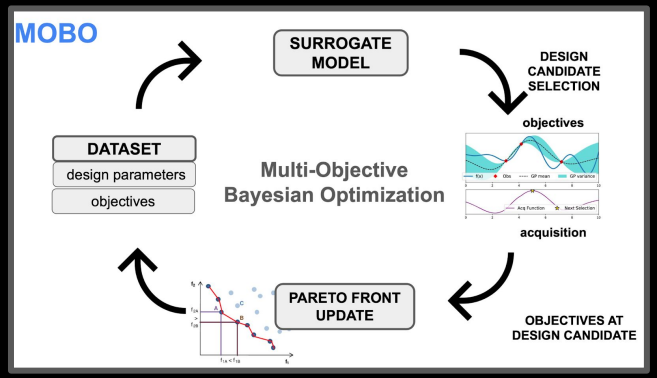
This strategy can be generalized to multi-objective optimization to find the Pareto front

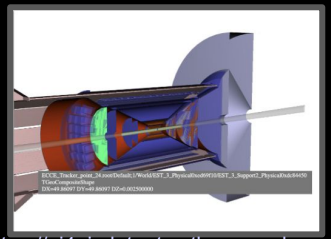
AID2E

AI-assisted Detector Design for EIC

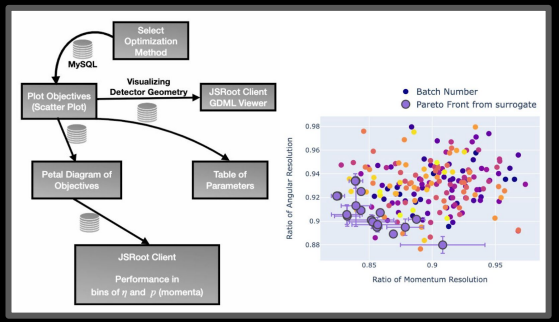
AID(2)E Coll. , arXiv:2405.16279 (2024)

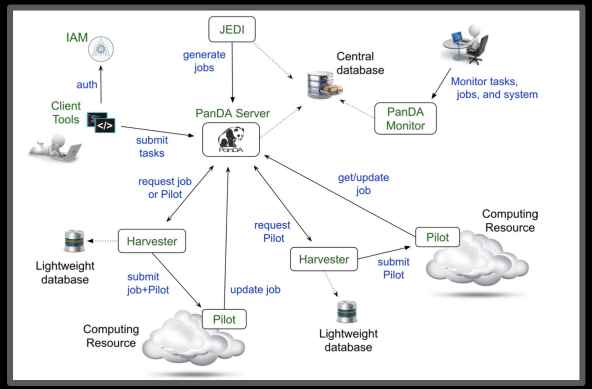
BNL, CUA, Duke, JLab, W&M





https://ai4eicdetopt.pythonanywhere.com/





(i) Will contribute to advance state of the art MOBO complexity to accommodate a large number of objectives and will explore usage of physics-inspired approaches

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

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

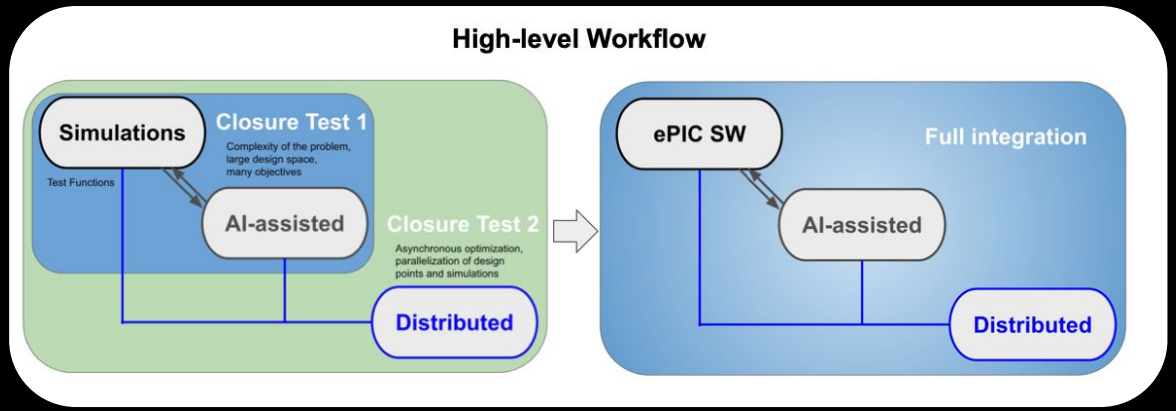
CF, Z. Papandreou, K. Suresh, et al. "AI-assisted optimization of the ECCE tracking system at the Electron Ion Collider." NIMA: 1047 (2023): 167748.
CF "Design of detectors at the electron ion collider with artificial intelligence." JINST 17.04 (2022): C04038.
Maeno, Tadashi, et al. "PanDA: Production and Distributed Analysis System." Computing and Software for Big Science 8.1 (2024): 1-21.

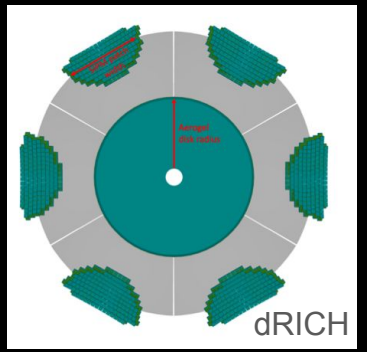
AID2E

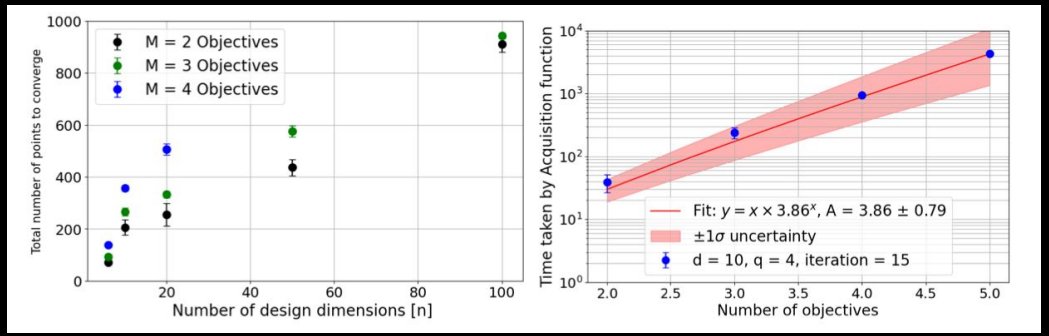
AI-assisted Detector Design for EIC

AID(2)E Coll. , arXiv:2405.16279 (2024)

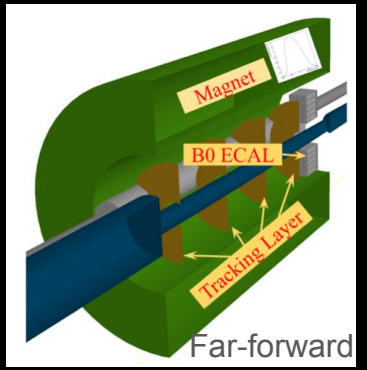
BNL, CUA, Duke, JLab, W&M



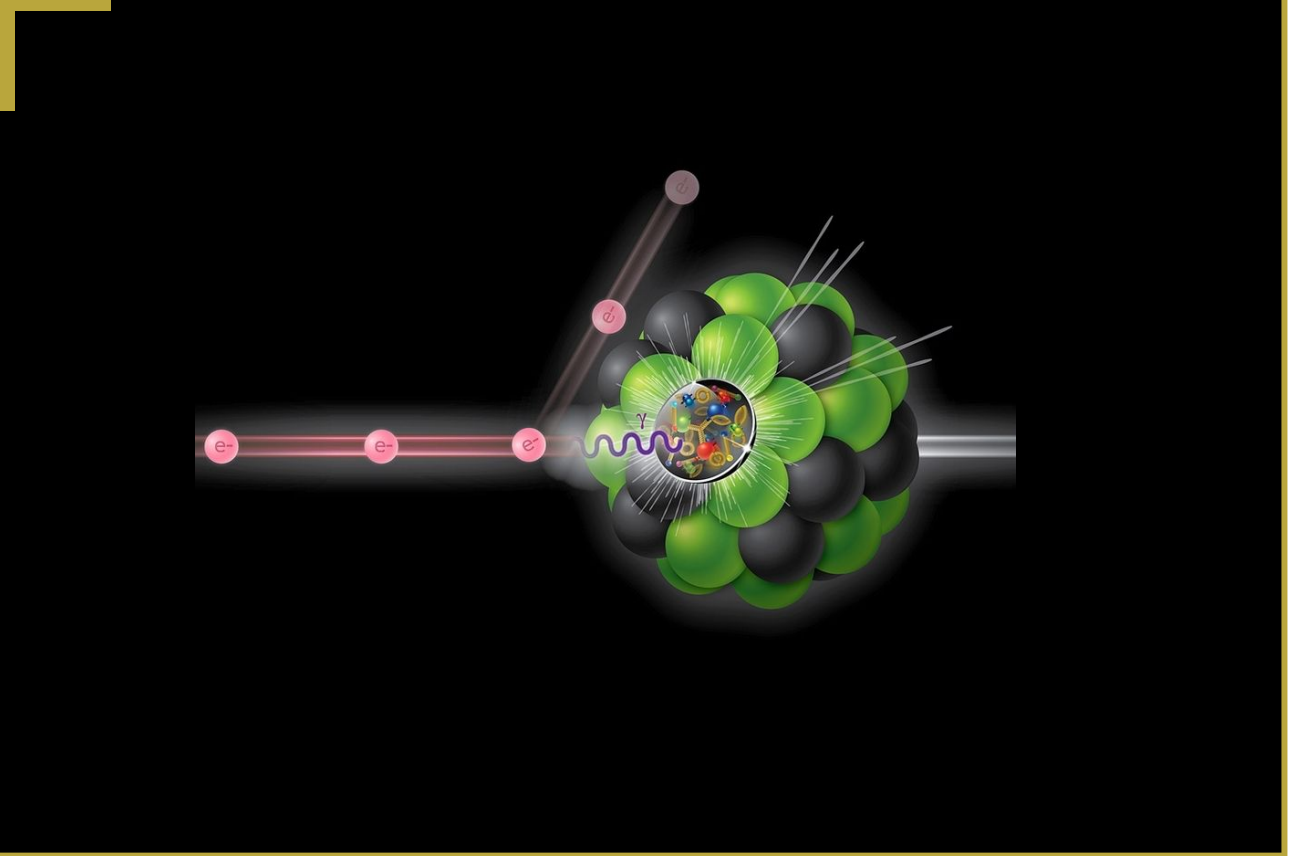


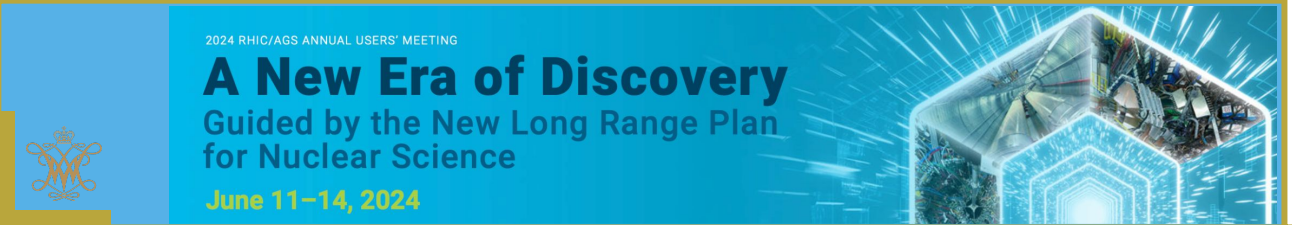


Complexity studies with DTLZ-2 benchmark problem



Applications for ePIC sub-detector systems



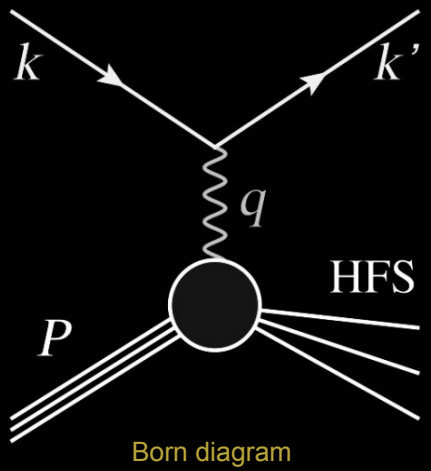


ELUQuant

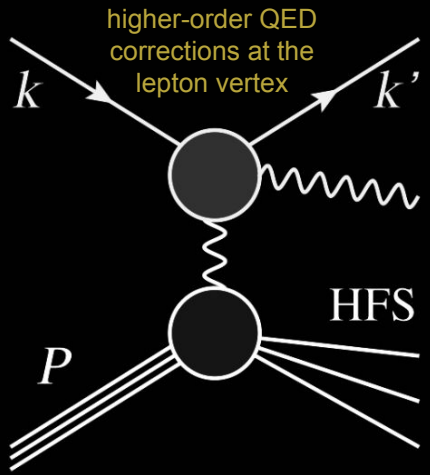
Event-Level Uncertainty Quant.

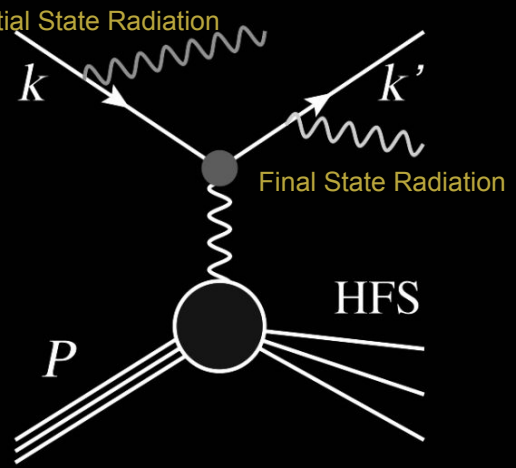
Deep Inelastic Scattering

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



Born diagram





These kinematic variables are related via the relation $Q^2 = sxy$, where s is the square of the center-of-mass energy.

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

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

Reconstruction 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, the 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

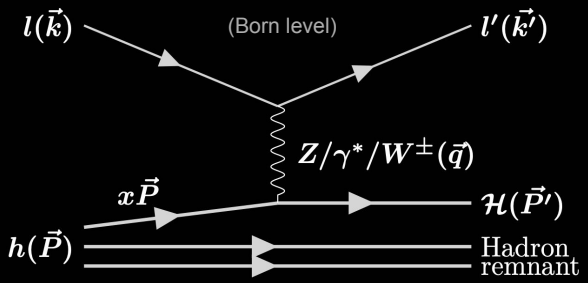
Method name	Observables	y	Q^2	$x \cdot E_p$
Electron (e)	$[E_0,E,\theta]$	$1-\frac{\Sigma_e}{2E_0}$	$\frac{E^2\sin^2\theta}{1-y}$	$\frac{E(1+\cos\theta)}{2y}$
Double angle (DA) [6, 7]	$[E_0,\theta,\gamma]$	$\frac{\tan\frac{\gamma}{2}}{\tan\frac{\gamma}{2}+\tan\frac{\theta}{2}}$	$4E_0^2\cot^2\frac{\theta}{2}(1-y)$	$\frac{Q^2}{4E_0y}$
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

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

Deeply Learning DIS

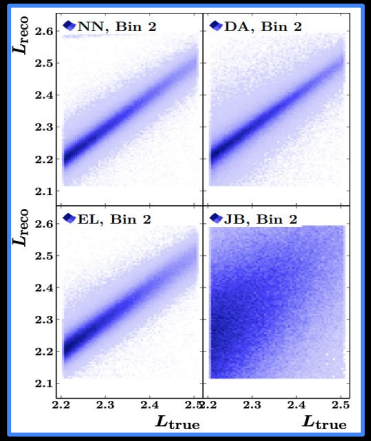
DIS fundamental process @EIC



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

Example in one specific bin



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

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

Input features of ELUQuant

Utilized input features and H1 MC dataset of paper NIM-A 1025 (2022): 16616

- Define variables to characterize the strength of QED radiation

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

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 $\Sigma_e - \Sigma$.

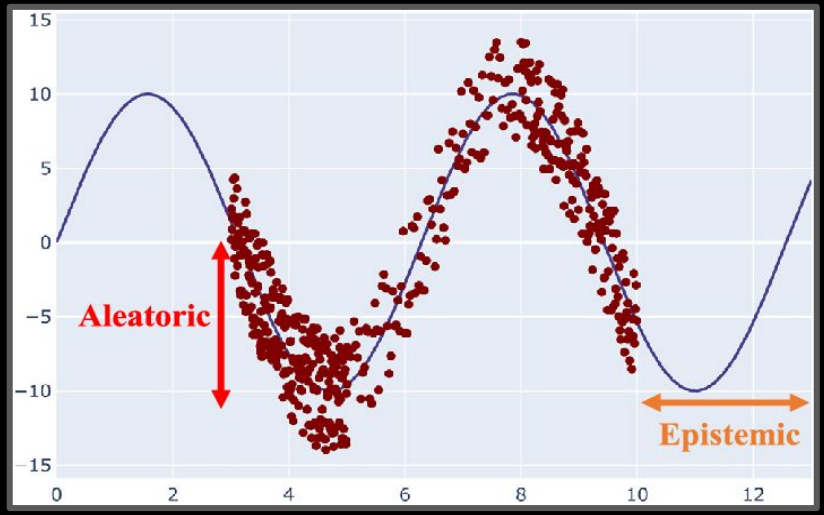
Tot. 15 input features

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

Epistemic vs Aleatoric

- Epistemic Uncertainty
  - Arises from a lack of knowledge which is reflected in the effectiveness of the model in describing the data. Can be reduced as more information or data becomes available, and by improving the model. It can be affected by inaccuracy.
- Aleatoric Uncertainty
  - Due to inherent variability or randomness in a process or system and cannot be reduced by collecting more data.

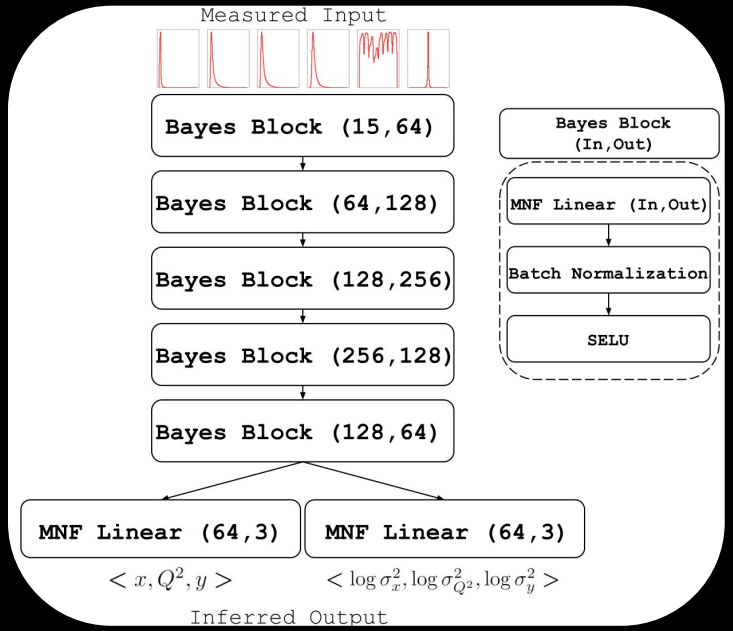


Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." Information fusion 76 (2021): 243-297.

ELUQuant

C. Fanelli, and J. Giroux. Machine Learning: Science and Technology 5.1 (2024): 015017.

Event-Level Uncertainty Quantification



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

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 (systematic) uncertainty through sampling MNF [1] layers

Learn the regression transformation

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

epistemic aleatoric

Access aleatoric (statistical) as a function of regressed output [2]

Constrain the physics

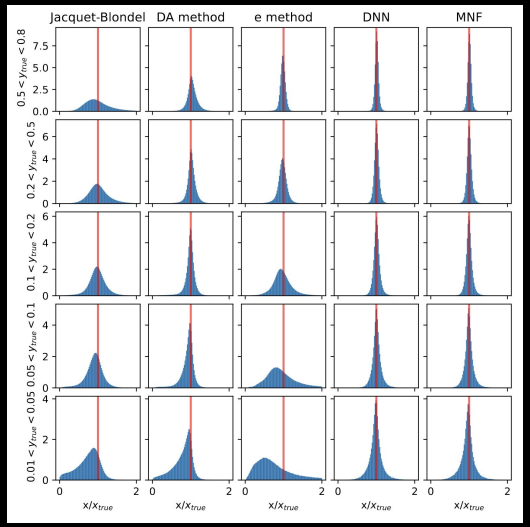
$$\mathcal{L}_{Phys.} = \frac{1}{N}\sum_i \log\hat{Q}_i^2 - (\log s_i + \log\hat{x}_i + \log\hat{y}_i)$$

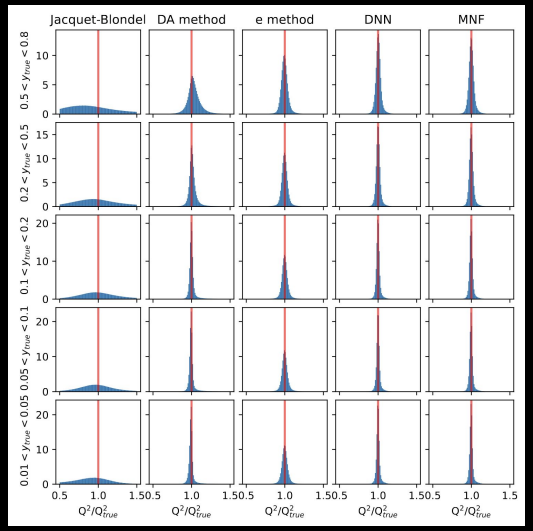
DIS

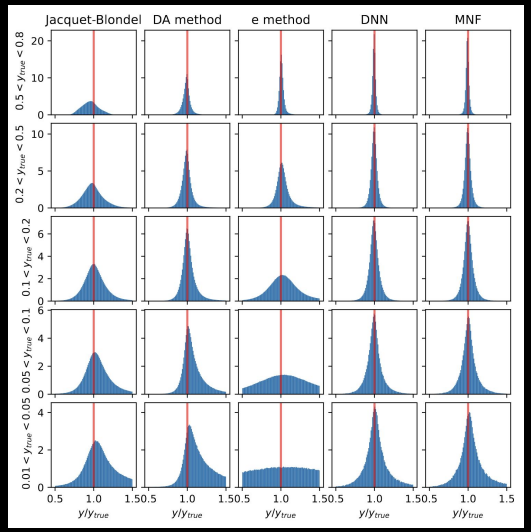
[1] C Louizos, M Welling International Conference on Machine Learning; arXiv:1703.01961 Multiplicative Normalizing Flows for Variational Bayesian Neural Networks
[2] A. Kendall and Y. Gal. "What uncertainties do we need in Bayesian deep learning for computer vision?." Adv. Neural Inf. Process. 30 (2017).

Aleatoric vs RMS

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







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

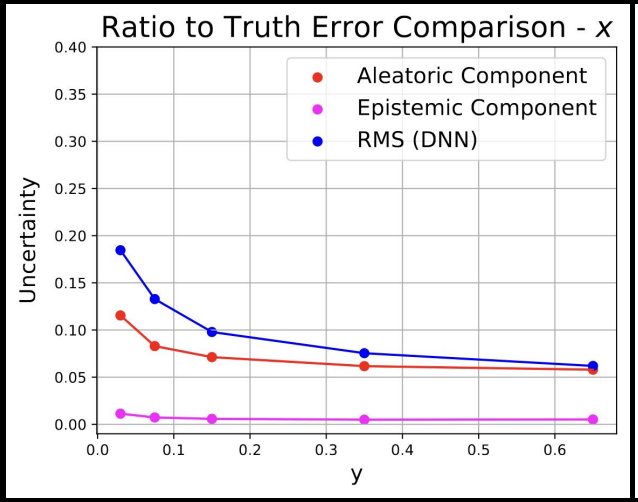
Y Bin	DA Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.060537	0.031194	0.034643
(0.2, 0.5)	0.082115	0.053126	0.044249
(0.1, 0.2)	0.098631	0.078143	0.061840
(0.05, 0.1)	0.127276	0.109309	0.078276
(0.01, 0.05)	0.158493	0.147391	0.120546

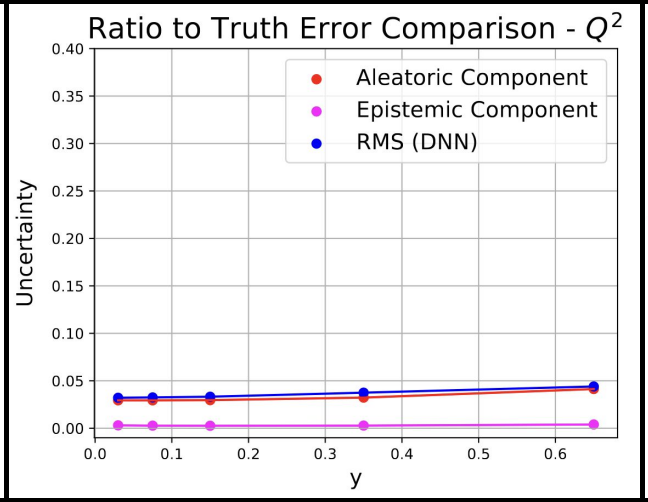
Table 4: Aleatoric RMS Comparison Y

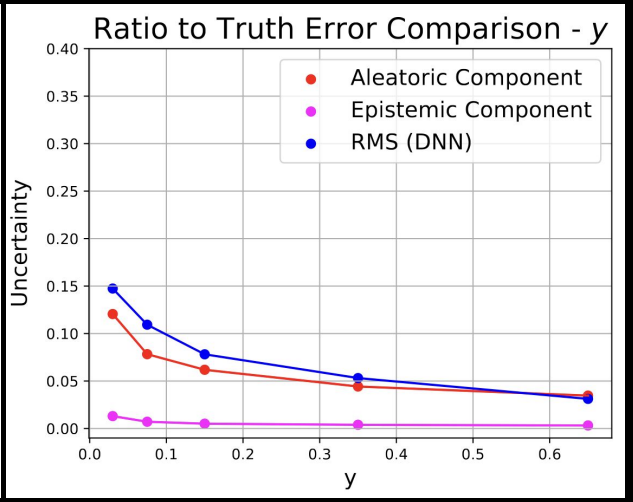
- ELUQuant performance similar to DNN
- Closure test on aleatoric when epistemic is negligible and distribution is gaussian

*M. Arratia, D. Britzger, O. Long, B. Nachman, et al., "NIM-A 1025 (2022): 166164 (DNN)

Comparison between DNN and BNN



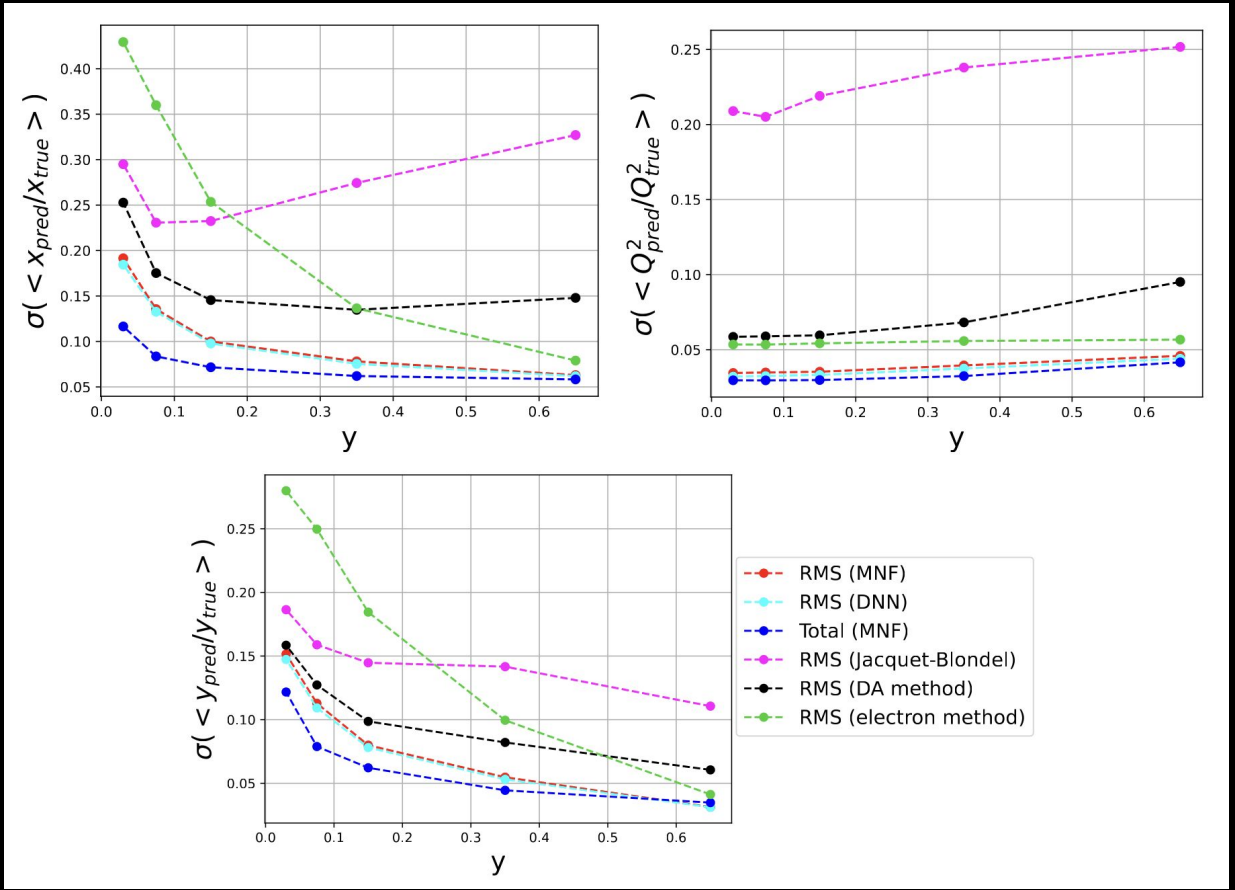




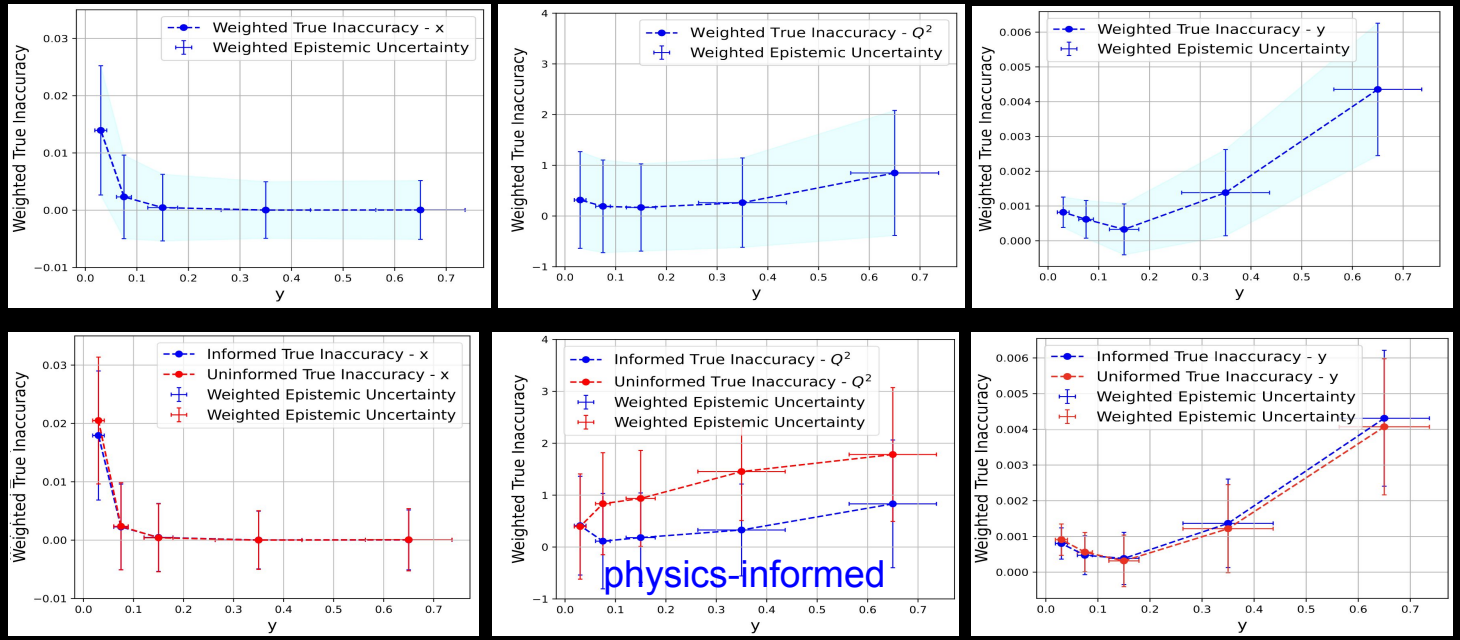
- (from table before) The RMS (ELUQ) roughly coincide with that of DNN as seen previously
- The RMS (DNN) for x and y is larger at low y given the distributions are broader
- The epistemic is systematically smaller than aleatoric component.
- At large y, for x and y the total uncertainty (epistemic+aleatoric)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



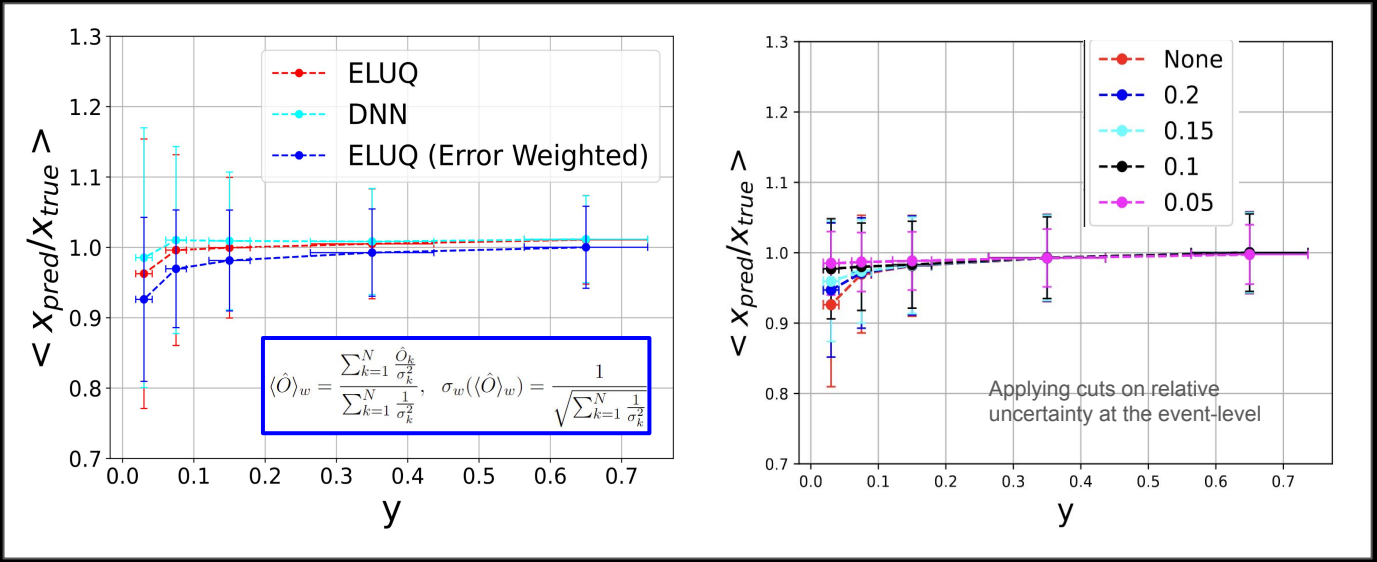
Epistemic vs True Inaccuracy



1. The plots show that the epistemic uncertainty is larger when the true inaccuracy is larger — N.b.: at inference, we are agnostic to the true inaccuracy

2. The physics-informed term (blue) contributes to decrease the true inaccuracy.

Leveraging event-level information



- A “simple” DNN does not have per se uncertainty at the event level. In the plots we use the RMS from final distributions. We also compare to a weighted average.

- 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

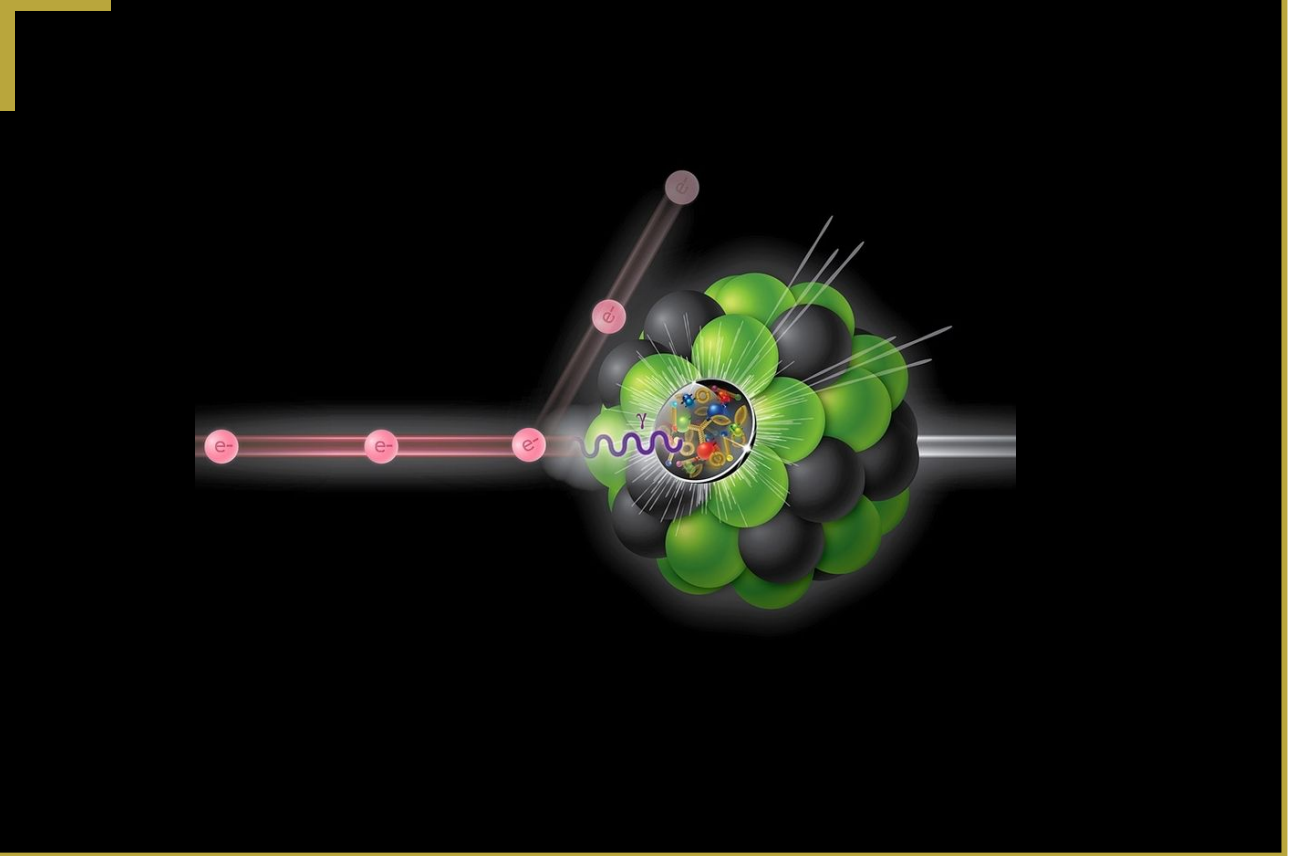
- We know that ELUQuant is sensitive to anomaly detection. Performance studies are underway.

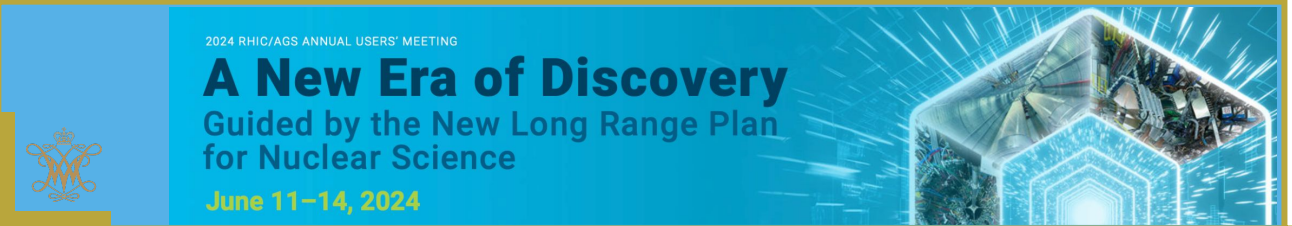
Computing performance

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

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

- In computational terms, ELUQuant at inference showed an impressive rate of 10,000 samples/event within a 20 milliseconds on an RTX 3090.



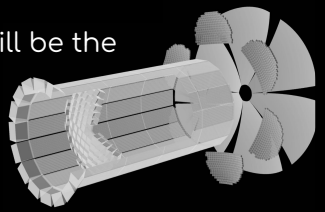


Deep(er)RICH

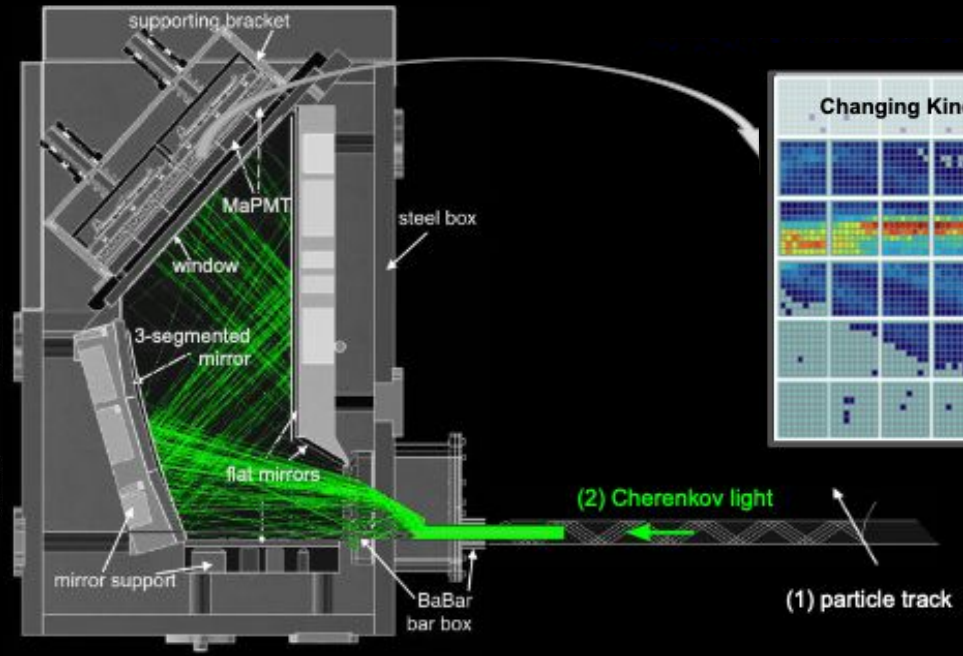
Deep Reconstruction of Imaging CHerenkov

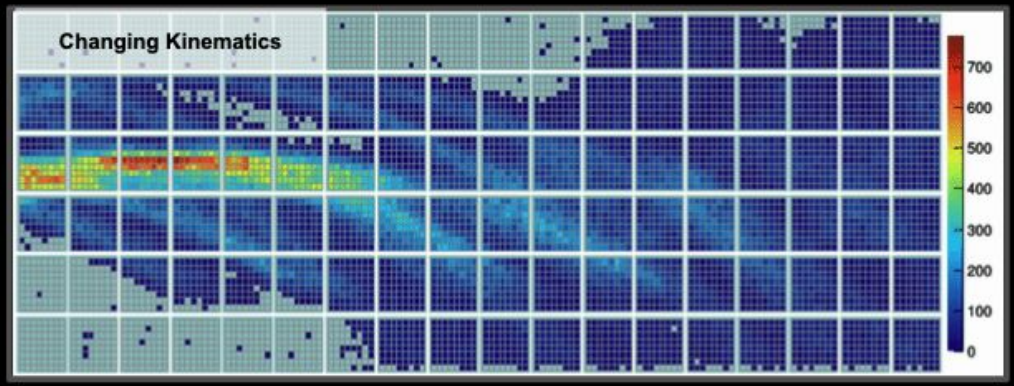
PID with Cherenkov

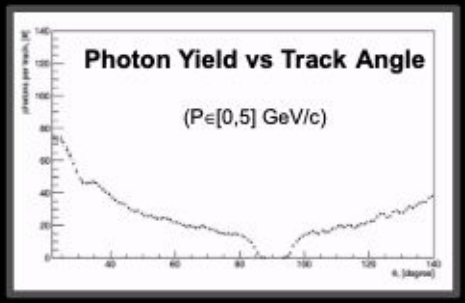
Cherenkov detectors will be the backbone of PID at EIC



DIRC at GlueX is instrumental for PID







[1] C. Fanelli, J. Pomponi, “DeepRICH: learning deeply Cherenkov detectors”, Mach. Learn.: Sci. Technol., 1.1 (2020): 015010
[2] C. Fanelli, "Machine learning for imaging Cherenkov detectors." JINST 15.02 (2020): C02012.

PID with Cherenkov

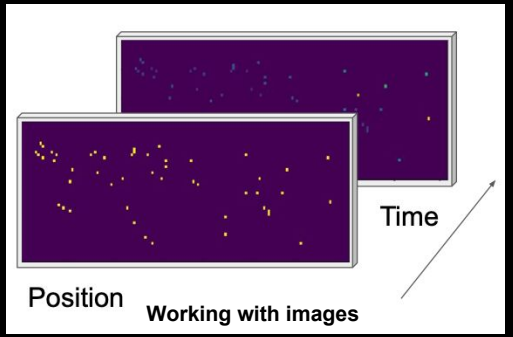
Deep(er)RICH

Challenges:

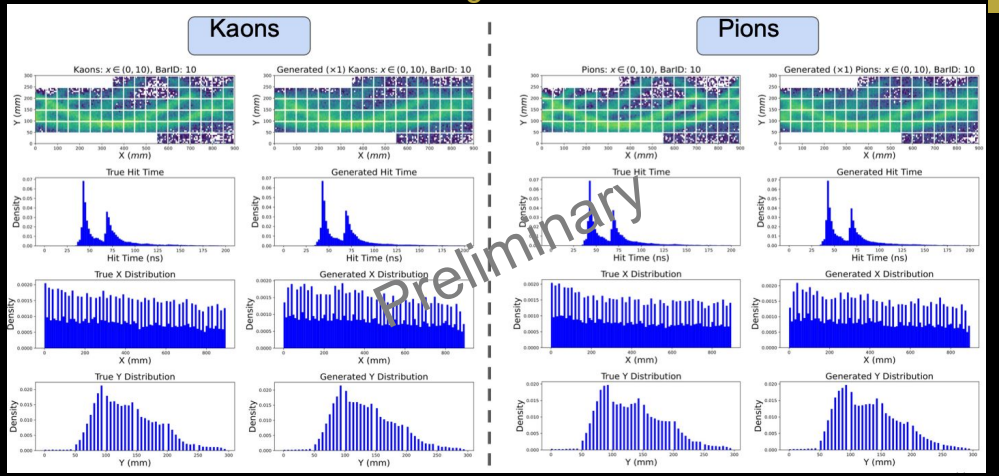
- Complex hit patterns (DIRC is the most complex), sparse data, response vs kinematics, simultaneous tracks
- Expensive simulations for optical processes with many photons tracked through complex surfaces

AI/ML Solutions:

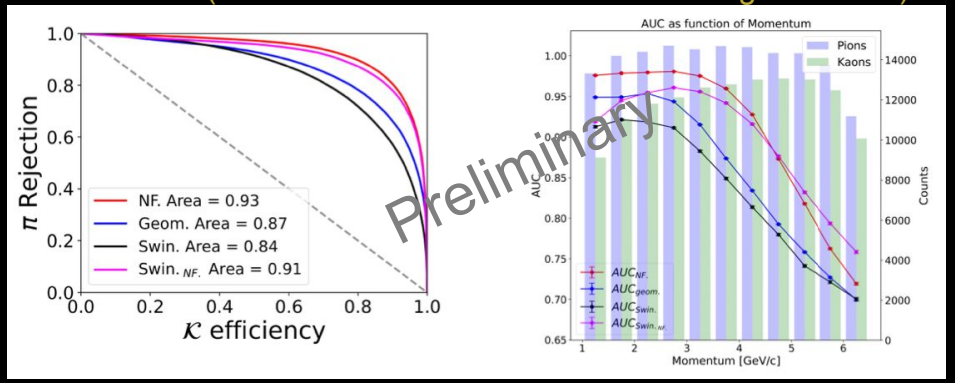
- With DeepRICH (Mach. Learn.: Sci. Technol. 1 015010, 2020) we showed same reconstruction performance of best reconstruction algorithm with ~4 orders of magnitude speed-up in inference time on GPU
- Fast simulation from generative models - Possibility to learn at the event-level the detector response using real data
- Possibility to deal with overlapping hit patterns from simultaneously detected tracks



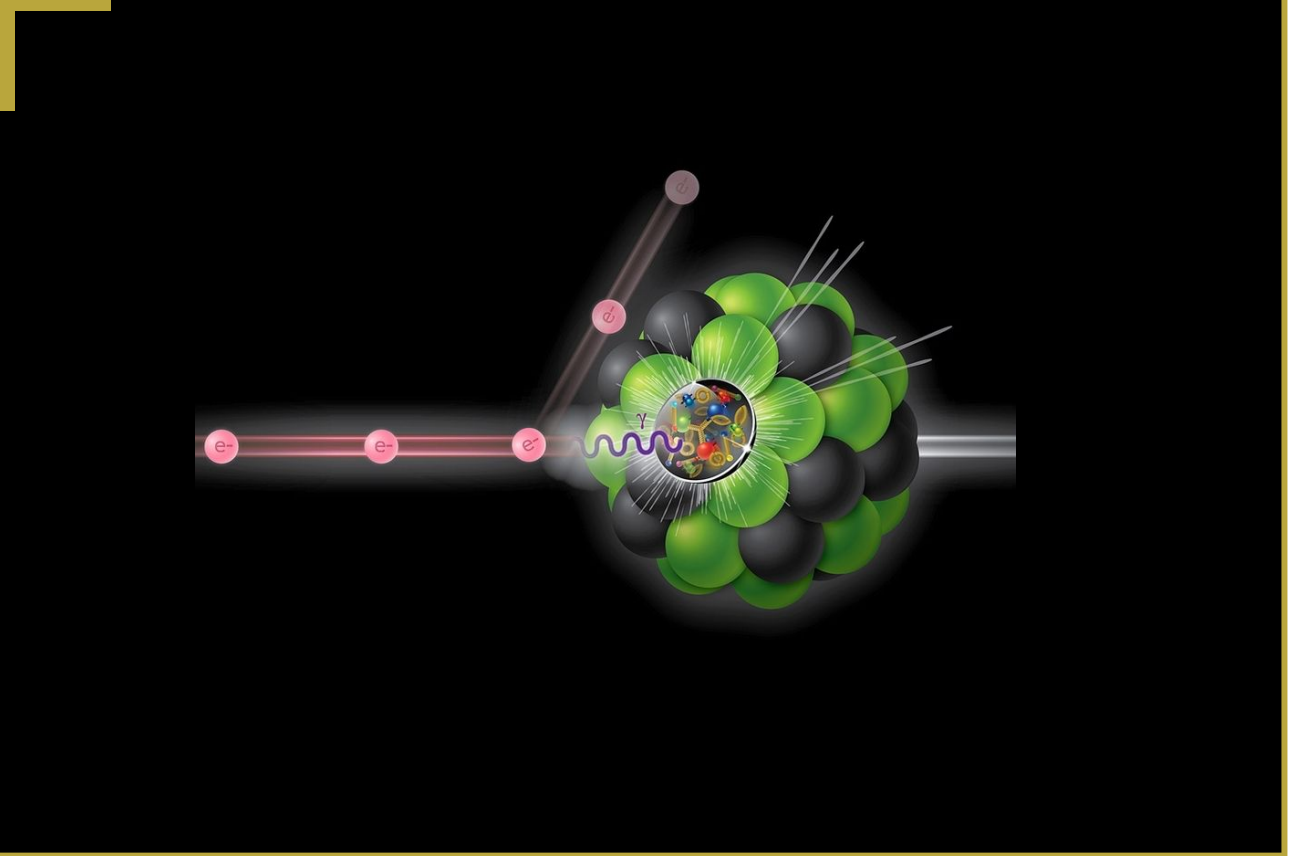
Fast simulations with normalizing flows

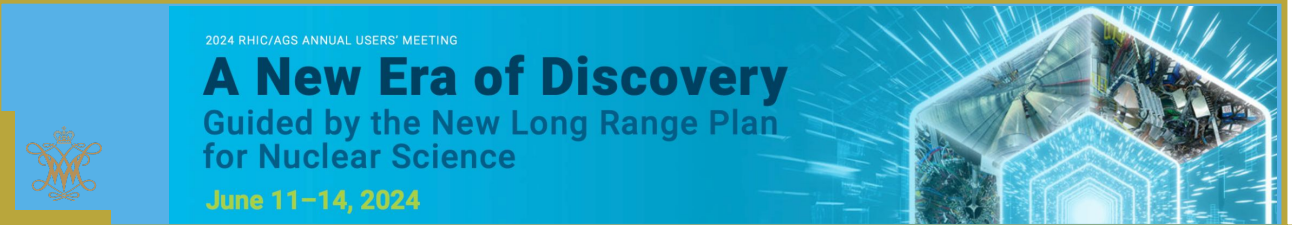


PID methods (Swin Transformer vs NF vs classical geom. reco)



C. Fanelli, J. Giroux., J. Stevens (in progress)

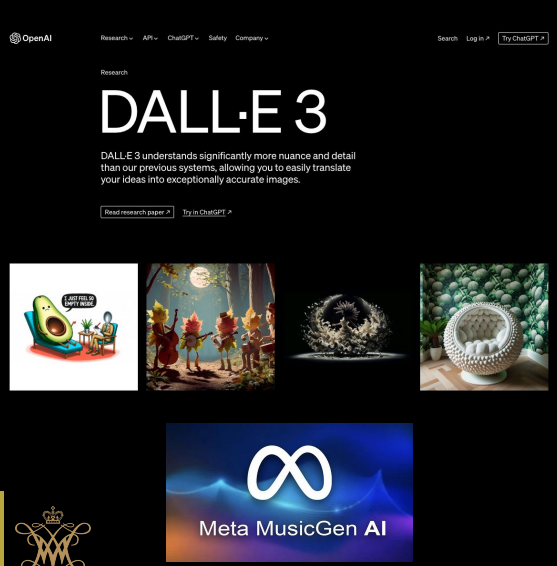


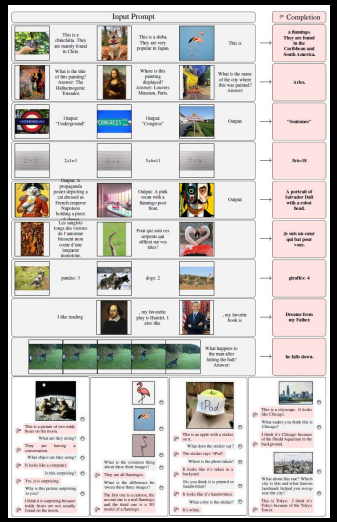


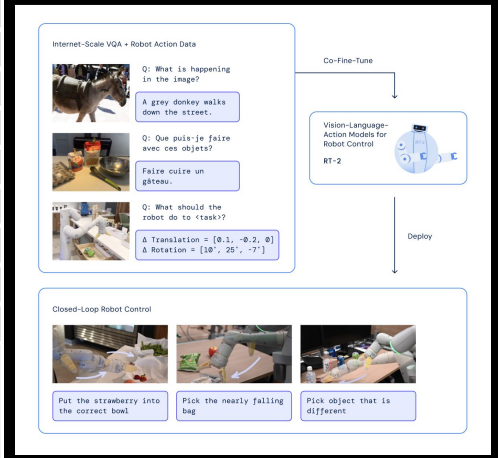
RAG-based Agent

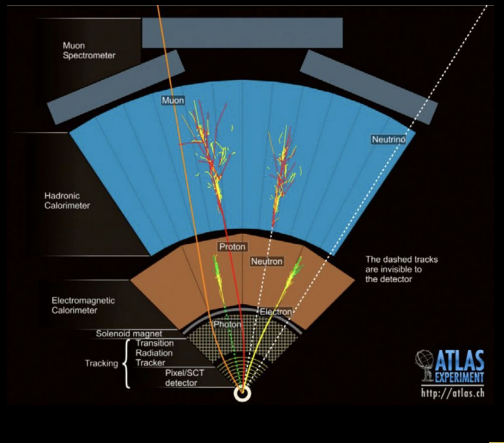
Foundation Models

- Foundation models are AI models trained on broad data sets, allowing for versatility across multiple applications, and have been pivotal in transforming AI, particularly in powering advanced chatbots and generative AI applications.
- The first examples of foundation models were pre-trained language models like Google's BERT and the "GPT-n" series of OpenAI's
- Foundation models have been developed across a range of modalities, see, e.g., DALL-E and Flamingo for images, MusicGen for music, RT-2 for robotic control, etc
- Emerging interesting research activities in particle physics (see, e.g., trackingBERT talk at AI4EIC) inspired by these approaches









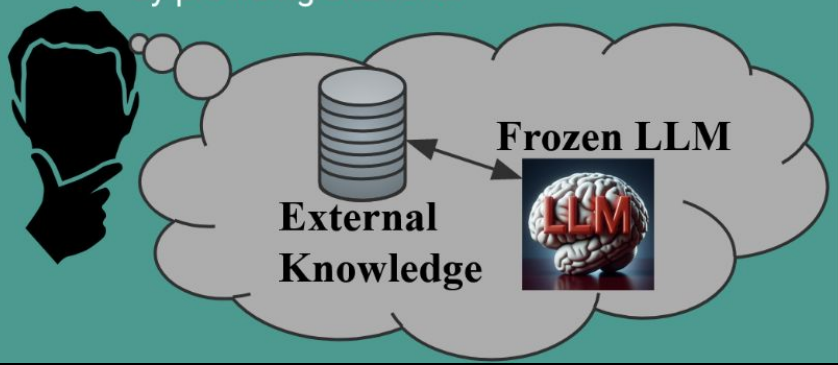
See AI4EIC2023, session "Foundation Models and Trends in Data Science"

RAG-based summarization AI for EIC

K. Suresh, N. Kackar, L. Schleck, CF, arXiv preprint arXiv:2403.15729 (2024)

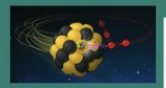
What is RAG ?

- 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?

- Electron Ion Collider (EIC) is a large scale experiment.
- Regular updates to documents, Run Wiki
- Newbies may take~ months to get to know the full experimental details.
- Document size ∝ Scale of experiment





EICUG

1400 physicists, 240 institutions and 38 countries

RAG-based EIC agent: https://rags4eic-ai4eic.streamlit.app/

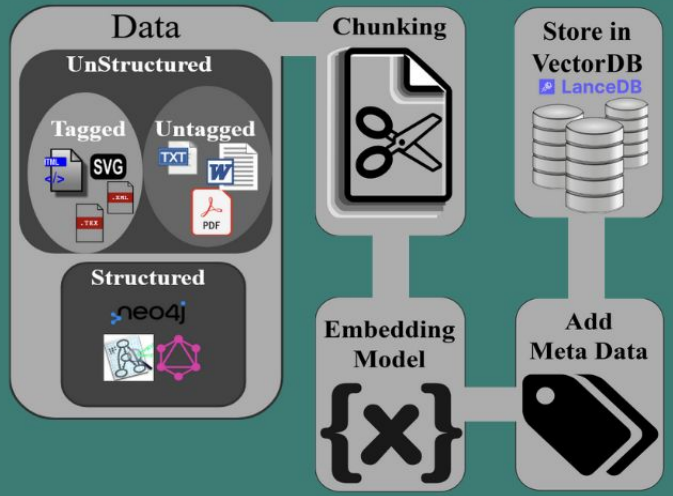
https://github.com/ai4eic/EIC-RAG-Project

RAG-based summarization AI for EIC

K. Suresh, N. Kackar, L. Schleck, CF, arXiv preprint arXiv:2403.15729 (2024)

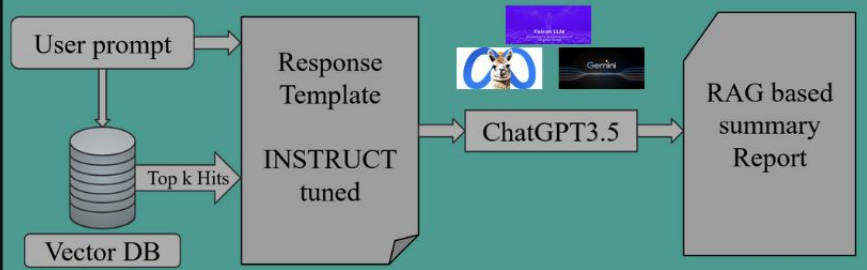
Ingestion

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



The inference*

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



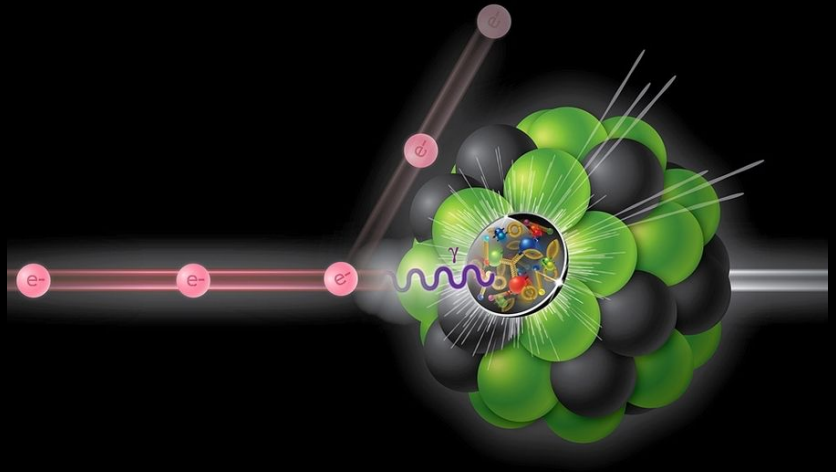
*Naive RAG pipeline

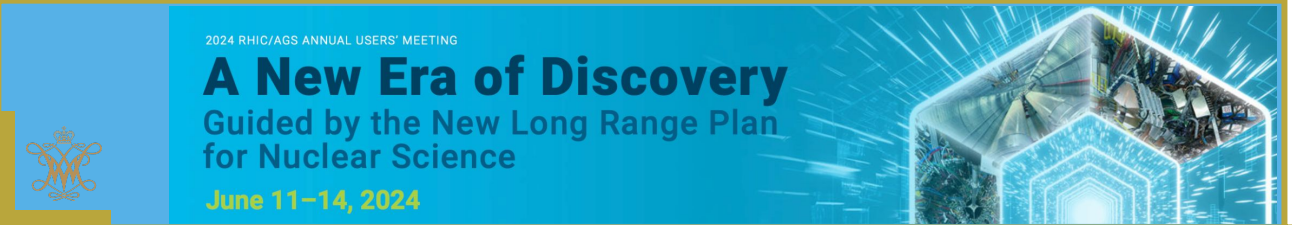
RAG-based EIC agent: https://rags4eic-ai4eic.streamlit.app/

https://github.com/ai4eic/EIC-RAG-Project

Conclusions

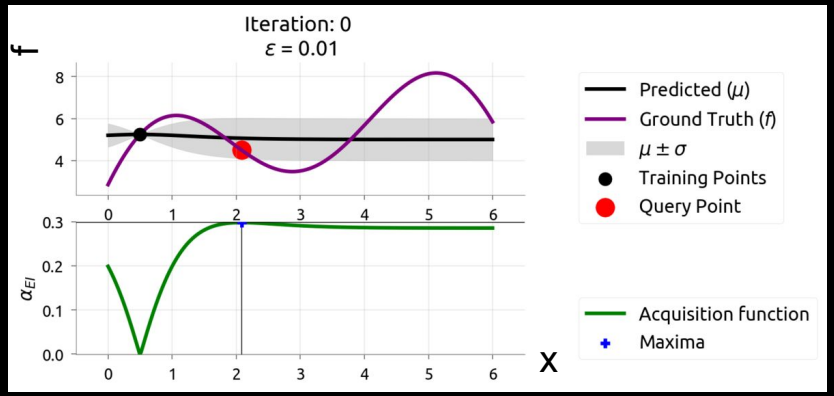
- AI/ML can be integrated into nearly all aspects of data processing pipelines for NP experiments
- Next-generation QCD experiments, such as the EIC, are being conceptualized during the AI revolution, enabling the incorporation of AI/ML from the design and R&D phases. The EIC detectors could be the first large-scale detectors optimized using AI/ML.
- Hadronic physics will see increasing benefits from ML; when studying non-perturbative effects, ML provides a comprehensive approach by utilizing full event information and can be trained on real data.
- Upcoming QCD experiments will harness the power of SRO and AI with heterogeneous computing:
  - Implement near real-time analysis and control, such as intelligent and autonomous detectors. A key focus is the application of AI methods with well-understood uncertainty quantification, both systematic and statistical.
  - Understanding uncertainties and biases in near real-time analysis with SRO could lead to a paradigm shift for next-generation QCD experiments, enabling faster turnaround times for producing scientific results.
- Foundation models have become some of the most powerful tools available today, yet their potential in our field remains largely untapped.


e-
e-
e-
γ


2024 RHIC/AGS ANNUAL USERS' MEETING
A New Era of Discovery
Guided by the New Long Range Plan
for Nuclear Science
June 11–14, 2024

Backup

Acquisition Functions



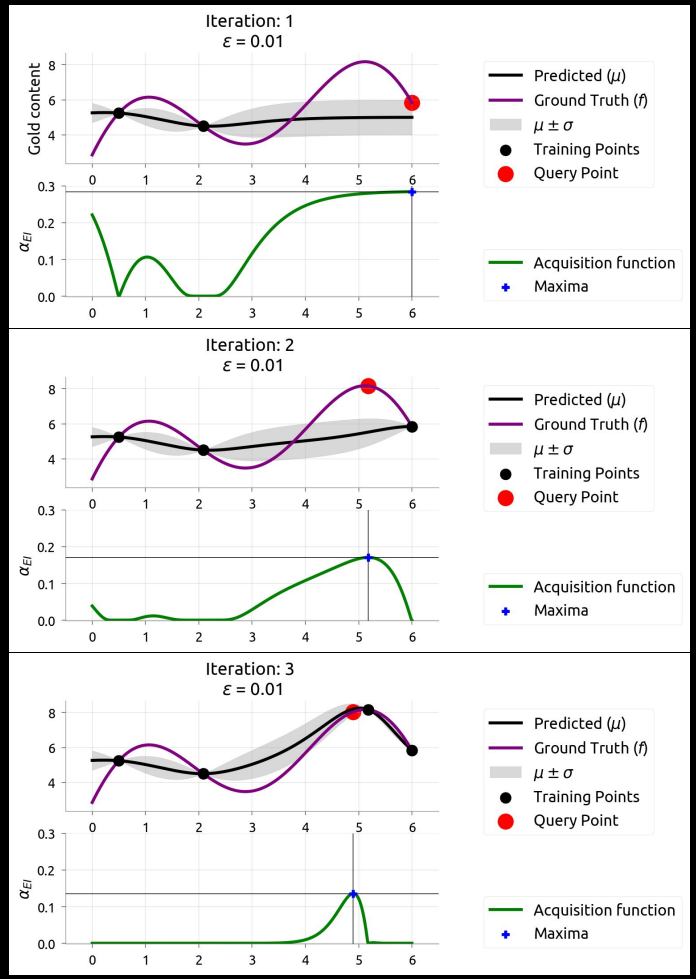
Utilization **Exploration**

$$EI(x) = \begin{cases} (\mu_t(x) - f(x^+) - \epsilon)\Phi(Z) + \sigma_t(x)\phi(Z), & \text{if } \sigma_t(x) > 0 \\ 0, & \text{if } \sigma_t(x) = 0 \end{cases}$$

Best found so far

$$Z = \frac{\mu_t(x) - f(x^+) - \epsilon}{\sigma_t(x)}$$

We are sampling x



- **"Exploitation"**: search where μ is high
- **"Exploration"**: search where σ is high

MOO

- In the following we will refer to the multi-objective optimization based on evolutionary algorithms [1], and in particular pymoo [2], written in Python, which also includes visualization and decision making tools.
- The definition of a generic MOO problem can be formulated as:

$$
\begin{aligned}
\min \quad & f_m(\mathbf{x}) && m = 1, .., M, \\
\text{s.t.} \quad & g_j(\mathbf{x}) \leq 0, && j = 1, .., J, \\
& h_k(\mathbf{x}) = 0, && k = 1, .., K, \\
& x_i^L \leq x_i \leq x_i^U, && i = 1, .., N.
\end{aligned}
$$

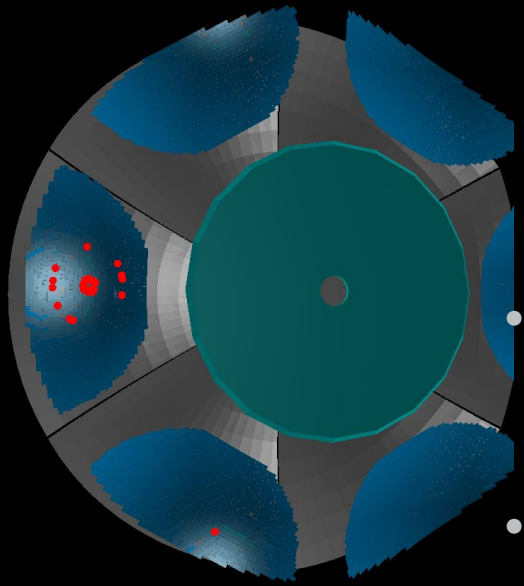
- M objective functions *f(x)* to optimize. By construction, pymoo performs minimization so a function to maximize needs a minus sign.
- There can be J inequalities *g(x)*
- There can be K equality constraints *h(x)*
- There are N variables x_i with lower and upper boundaries.

[1] Deb, Kalyanmoy. *Multi-objective optimization using evolutionary algorithms*. Vol. 16. John Wiley & Sons, 2001.
[2] Blank, Julian, and Kalyanmoy Deb. "pymoo: Multi-objective Optimization in Python." *IEEE Access* 8 (2020): 89497-89509

Candidates for Optimization in ePIC

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

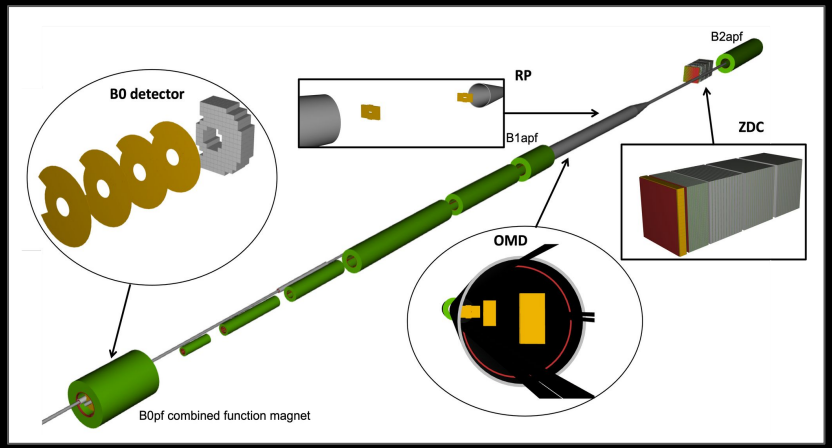
E. Cisbani *et al* 2020 *JINST* **15** P05009



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

dual-RICH

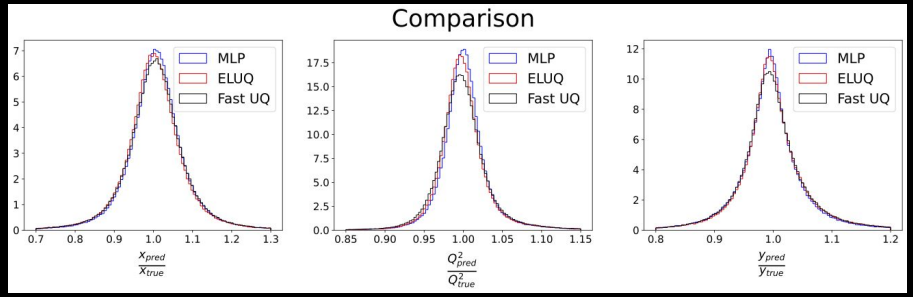
- *B0 magnetic field map, distance between space (always considered even), central location of tracker*
- *Momentum resolution, acceptance*

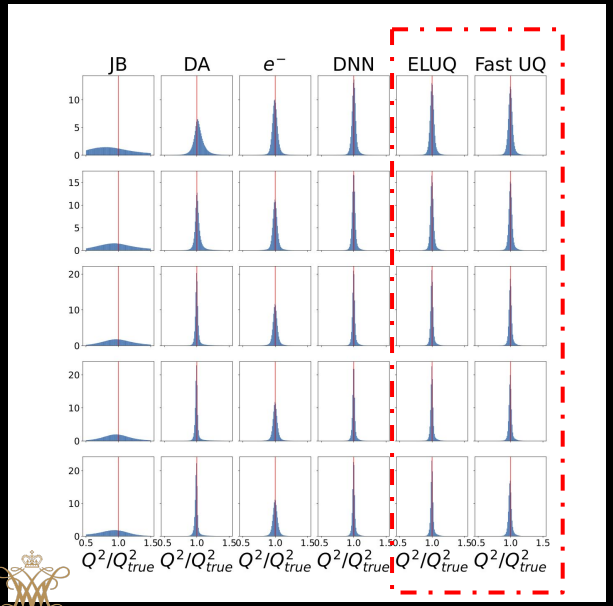


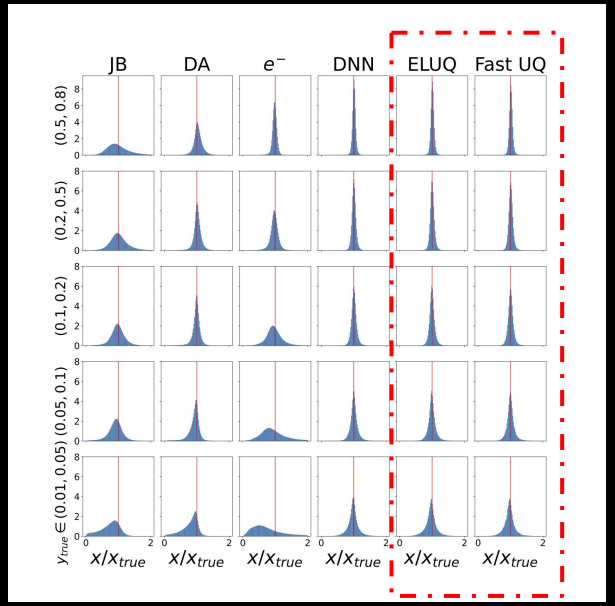
Far-Forward

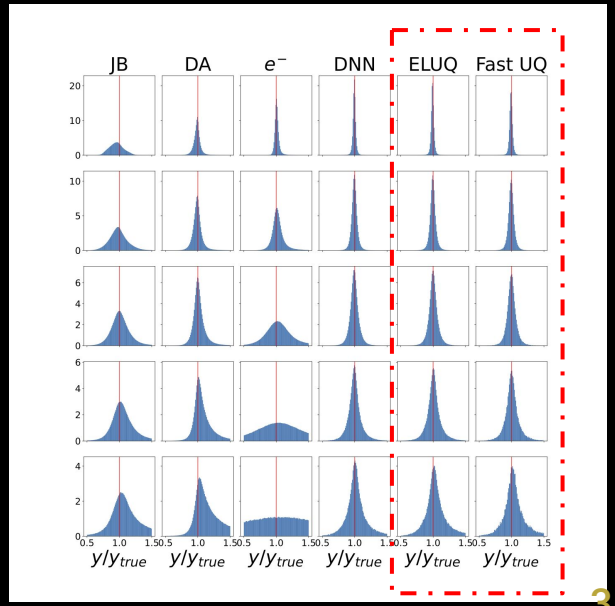
Ongoing discussion with working groups to identify potential

Towards near real-time

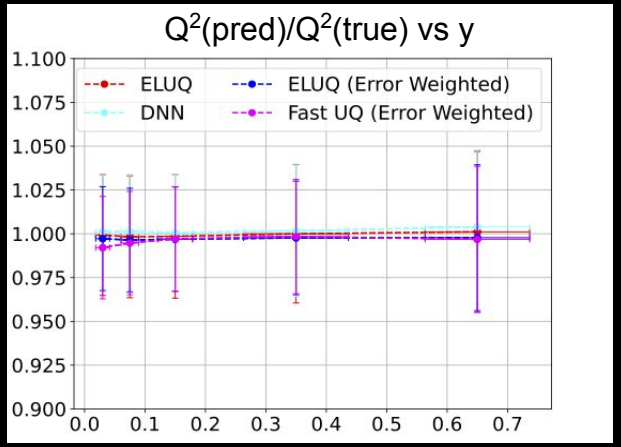


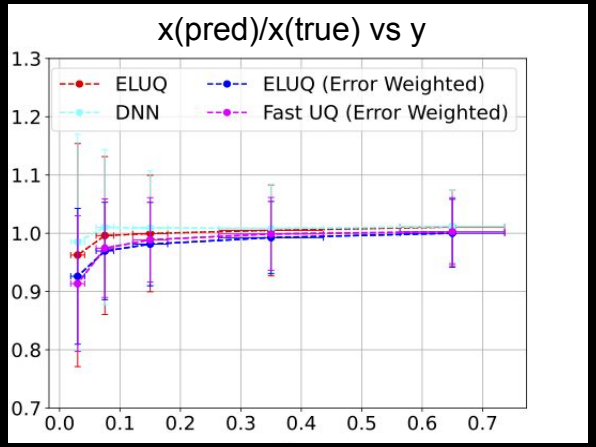


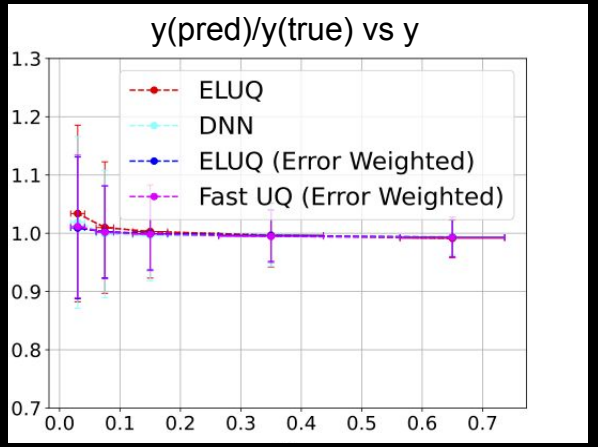


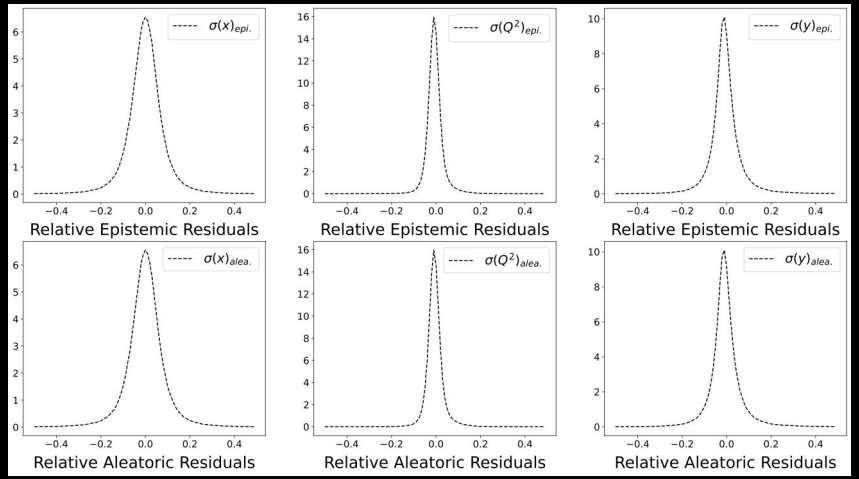


Towards near real-time









ELUQuant/Fast UQ: Very similar performance at the event level, predictions on kinematics and epistemic + aleatoric uncertainties within ~5% on average

RAG-based summarization AI 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 –

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

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.

Question 3: What is the pixel size and expected momentum resolution at 2GeV from a MAPS ITS3 tracking layer

Iteration 1

- 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 3: 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']

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

RAG-based EIC agent: https://rags4eic-ai4eic.streamlit.app/