Towards Uncertainty Quantification in Near Real-Time Analysis

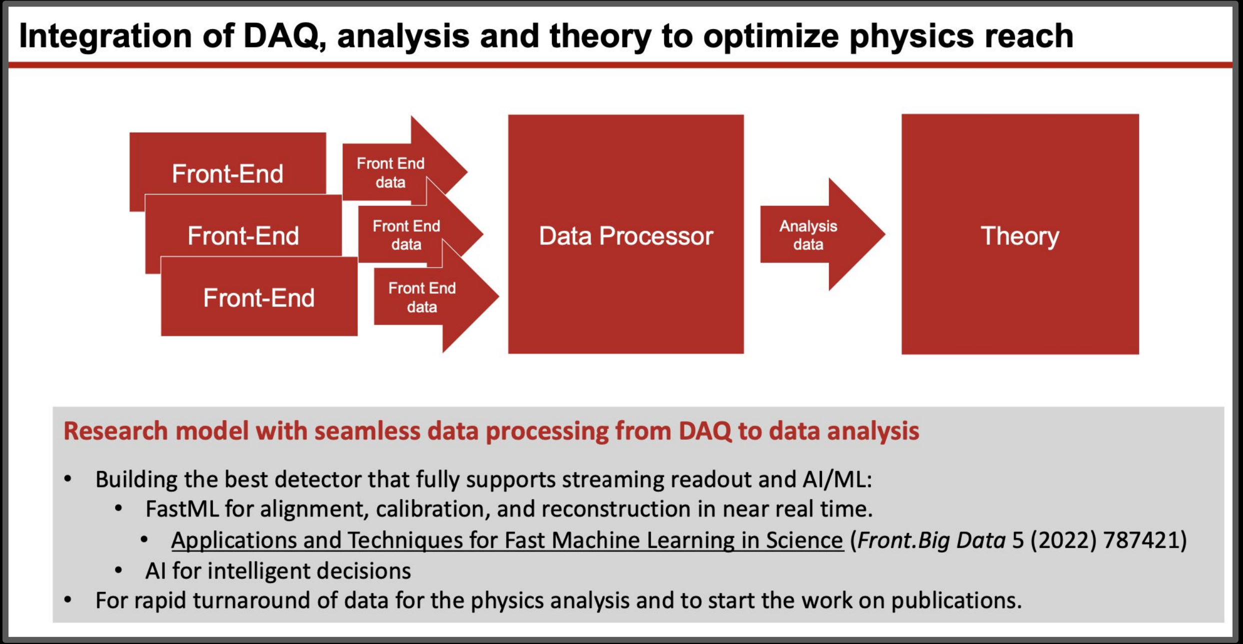
WILLIAM & MARY
CHARTERED 1693

Cristiano Fanelli

Streaming Readout XI, Nov-28 - Dec-2, 2023

Introduction

- Streaming readout supports near real-time analysis, and open up opportunities for integration of AI/ML for rapid data interpretation and decision-making



Integration of DAQ, analysis and theory to optimize physics reach

Front-End

Front-End

Front-End

Front End data

Front End data

Front End data

Data Processor

Analysis data

Theory

Research model with seamless data processing from DAQ to data analysis

- Building the best detector that fully supports streaming readout and AI/ML:
  - FastML for alignment, calibration, and reconstruction in near real time.
    - Applications and Techniques for Fast Machine Learning in Science (*Front.Big Data* 5 (2022) 787421)
  - AI for intelligent decisions
- For rapid turnaround of data for the physics analysis and to start the work on publications.

Introduction

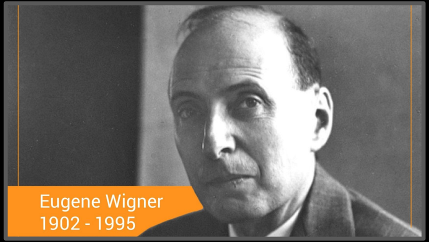
- Streaming readout supports near real-time analysis, and open up opportunities for integration of AI/ML for rapid data interpretation and decision-making
- Addressing uncertainty quantification in data processing and analysis is prominent for machine learning and deep learning applications. Neglecting UQ can have dramatic effects downstream in the near real-time data processing pipeline.

Not much taken into account yet

Introduction

- Streaming readout supports near real-time analysis, and open up opportunities for integration of AI/ML for rapid data interpretation and decision-making
- Addressing uncertainty quantification in data processing and analysis is prominent for machine learning and deep learning applications. Neglecting UQ can have dramatic effects downstream in the near real-time data processing pipeline.

E. Wigner: “The optimist regards the future as uncertain”


Eugene Wigner
1902 - 1995

- While including UQ may look daunting and increases “complexity”, at the same time addressing it, if possible in a streaming environment, would open possibilities, e.g.:
    - Uncertainty-aware models, making decisions also based on uncertainty
    - Multifold applications, spanning from data filtering and data quality monitoring to anomaly detection

Introduction

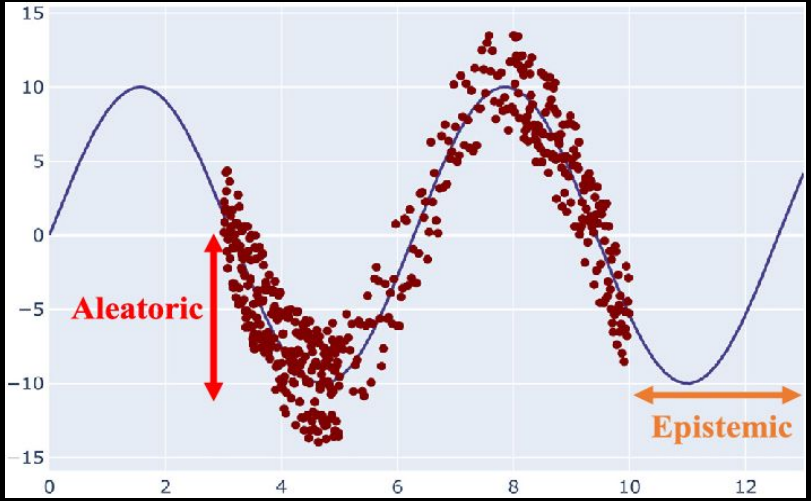
- Streaming readout supports near real-time analysis, and open up opportunities for integration of AI/ML for rapid data interpretation and decision-making
- Addressing uncertainty quantification in data processing and analysis is prominent for machine learning and deep learning applications. Neglecting UQ can have dramatic effects downstream in the near real-time data processing pipeline. This regards also ML/DL applications with streamed data.
- This can extend to fast reconstruction of abundant topologies collected in our detectors, and analyses at the event-level (or particle-level, depending on the application)
- For an Electron Ion Collider, one focus could be DIS events

event(particle)-level, uncertainty quantification (near real-time)
What I am going to show is heavily based on a recent paper accepted on NeurIPS'23 C. Fanelli, J. Giroux, "ELUQuant: Event-Level Uncertainty Quantification" arxiv:2310.02913 [cs.LG] (and references therein)

Disclaimer: Similar arguments can also be applied to other near real-time applications at the event/particle level

Epistemic vs Aleatoric

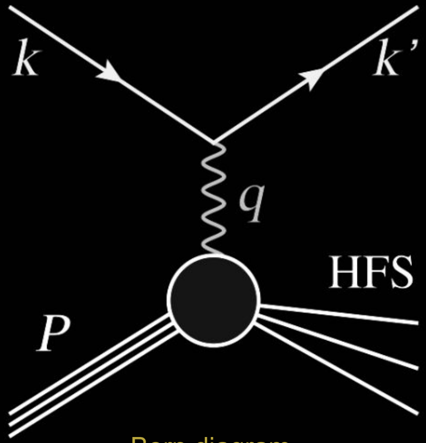
- Epistemic Uncertainty: This type of uncertainty arises from a lack of knowledge which is reflected in the effectiveness of the model in describing the data. It can be reduced as more information or data becomes available, and by improving the model. It can be affected by inaccuracy.
- Aleatoric Uncertainty: This uncertainty is due to inherent variability or randomness in a process or system and cannot be reduced by collecting more data. For example, even if we know the probability of getting heads when flipping a fair coin, the outcome of each individual flip is still uncertain.



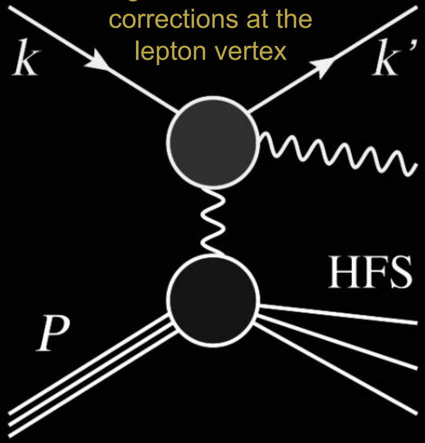
Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." Information fusion 76 (2021): 243-297.

Deep Inelastic Scattering

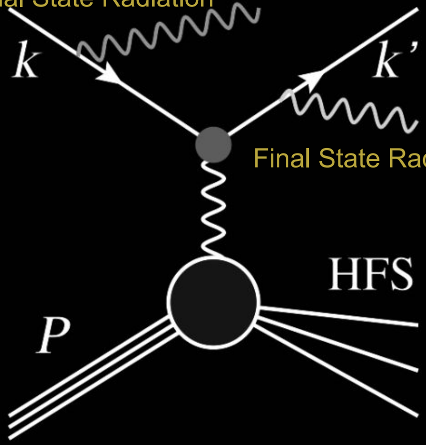
DIS is governed by the four-momentum transfer squared of the exchanged boson Q^2, the inelasticity y, and the Bjorken scaling variable x.



Born diagram



higher-order QED corrections at the lepton vertex



Initial State Radiation

Final State Radiation

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

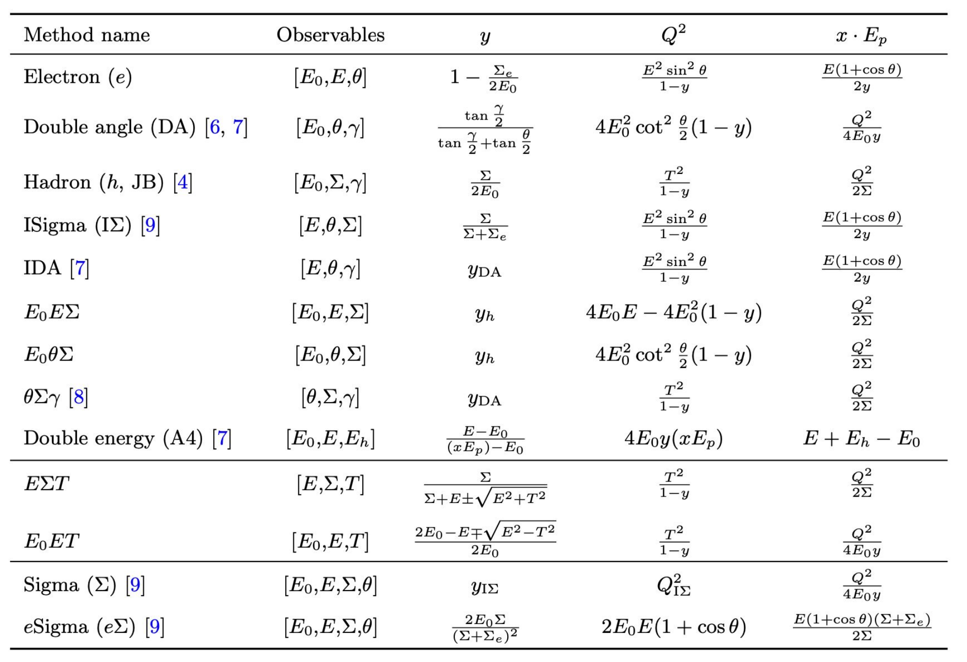
DIS kinematics: Traditional Methods

- 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]	$[E_0,E,\Sigma,\theta]$	$\frac{2E_0\Sigma}{(\Sigma+\Sigma_e)^2}$	$2E_0E(1+\cos\theta)$	$\frac{E(1+\cos\theta)(\Sigma+\Sigma_e)}{2\Sigma}$

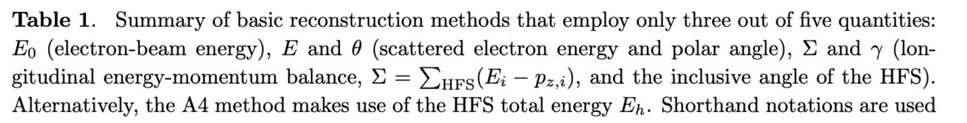
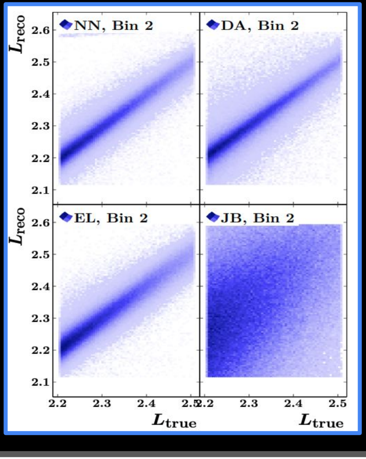


Table 1. Summary of basic reconstruction methods that employ only three out of five quantities: E_0 (electron-beam energy), E and θ (scattered electron energy and polar angle), Σ and γ (longitudinal energy-momentum balance, $\Sigma = \sum_{\mathrm{HFS}}(E_i - p_{z,i})$, and the inclusive angle of the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

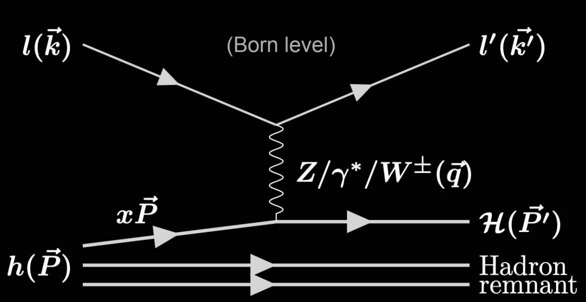
Table taken from Arratia et al., NIM-A 1025 (2022): 166164

Deeply Learning DIS

Example in one specific bin

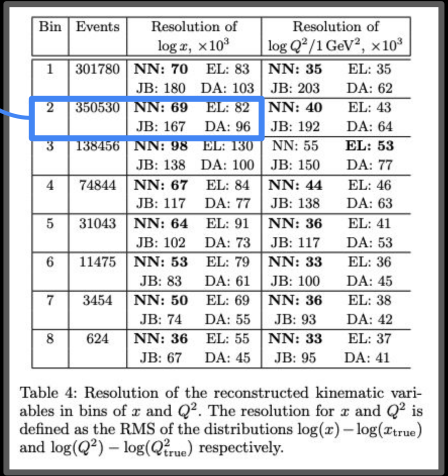


DIS fundamental process @EIC



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**	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 $\log(x) - \log(x_{\mathrm{true}})$ and $\log(Q^2) - \log(Q^2_{\mathrm{true}})$ respectively.

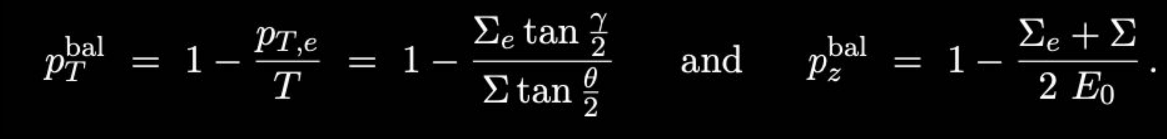
First application of DL for regression of DIS kinematics

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

Input features of ELUQuant

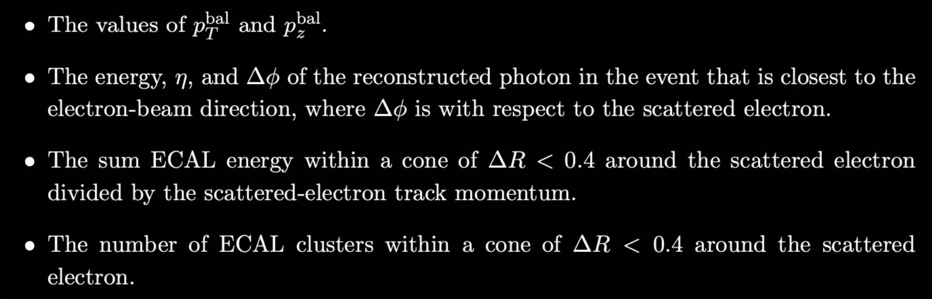
Utilized input features and H1 MC dataset of paper NIM-A 1025 (2022): 1661

- Define variables to characterize the strength of QED radiation



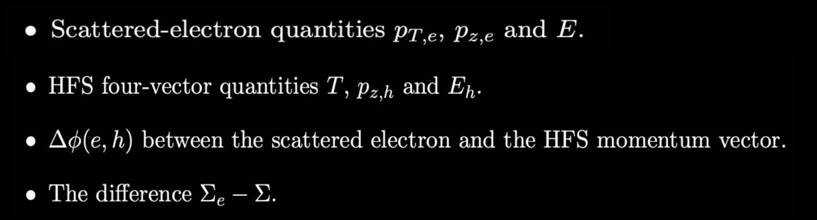
$$p_T^{\mathrm{bal}} = 1 - \frac{p_{T,e}}{T} = 1 - \frac{\Sigma_e \tan\frac{\gamma}{2}}{\Sigma \tan\frac{\theta}{2}} \quad \text{and} \quad p_z^{\mathrm{bal}} = 1 - \frac{\Sigma_e + \Sigma}{2\ E_0}.$$

7 features to help indicate QED radiation in the event



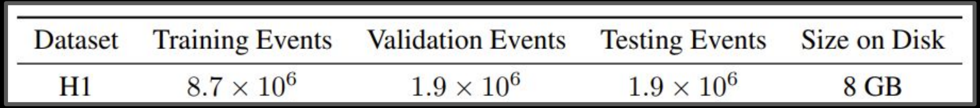
- The values of p_T^{bal} and p_z^{bal}.
- The energy, η, and $\Delta\phi$ of the reconstructed photon in the event that is closest to the electron-beam direction, where $\Delta\phi$ is with respect to the scattered electron.
- The sum ECAL energy within a cone of $\Delta R < 0.4$ around the scattered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

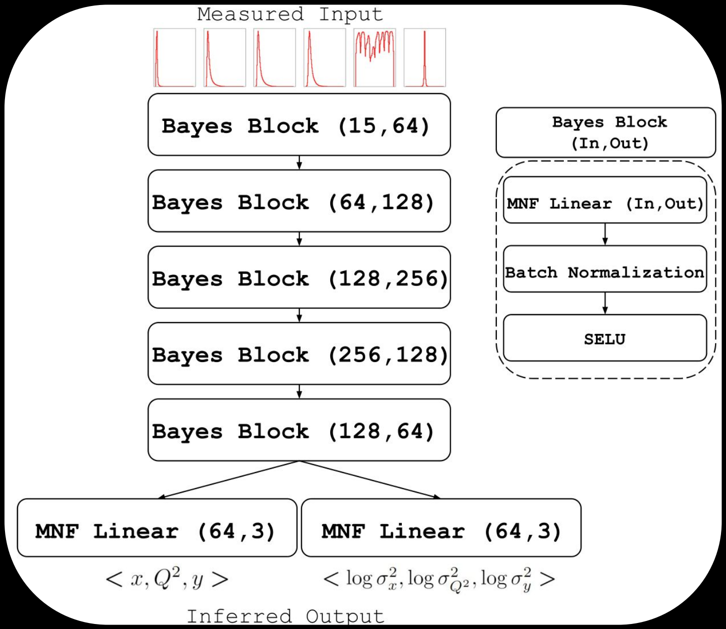


Dataset	Training Events	Validation Events	Testing Events	Size on Disk
H1	8.7×10^6	1.9×10^6	1.9×10^6	8 GB

*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

C. Fanelli, J. Giroux, “ELUQuant: Event-Level Uncertainty Quantification” arxiv:2310.02913 [cs.LG]



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

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

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

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

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

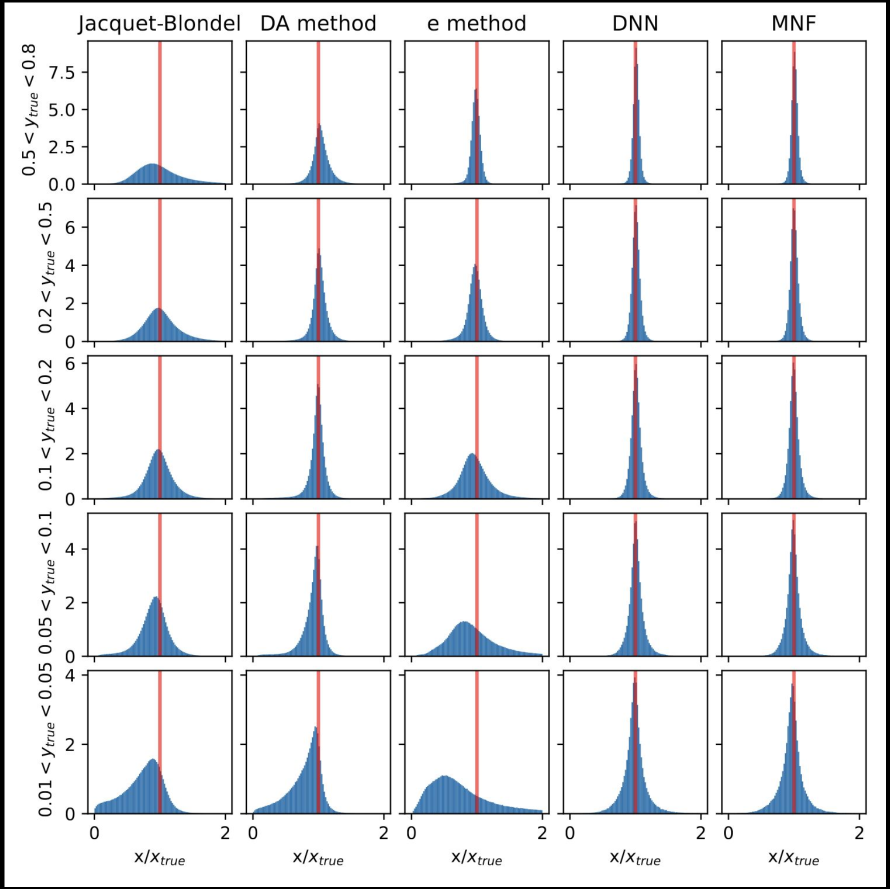
Constrain the physics

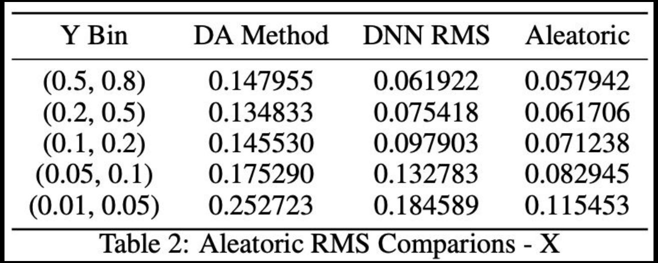
$$\mathcal{L}_{Phys.} = \frac{1}{N}\sum_i \log\hat{Q}_i^2 - (\log s_i + \log\hat{x}_i + \log\hat{y}_i)$$

[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-RMS comparison

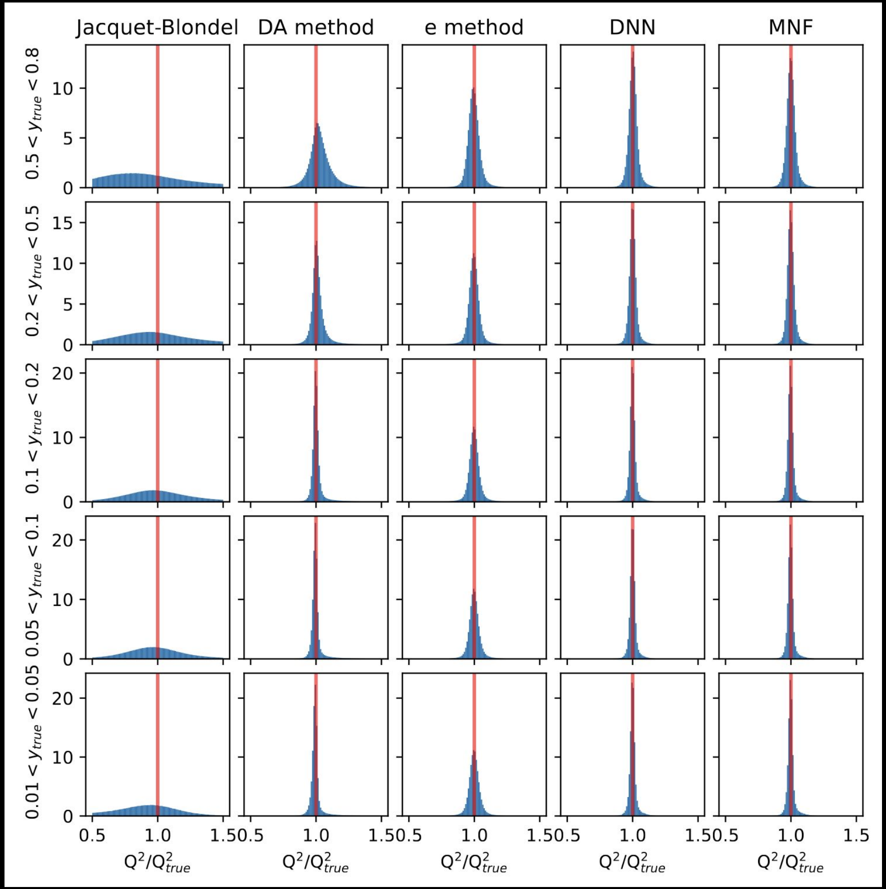


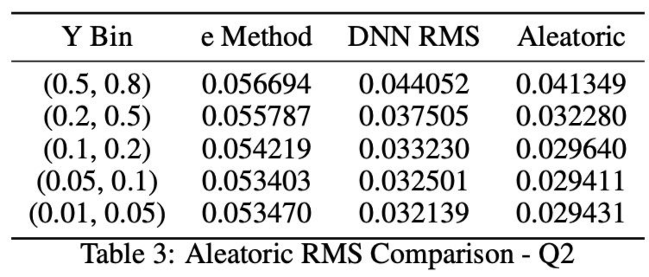


Y Bin	DA Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.147955	0.061922	0.057942
(0.2, 0.5)	0.134833	0.075418	0.061706
(0.1, 0.2)	0.145530	0.097903	0.071238
(0.05, 0.1)	0.175290	0.132783	0.082945
(0.01, 0.05)	0.252723	0.184589	0.115453

Table 2: Aleatoric RMS Comparions - X

Aleatoric-RMS comparison

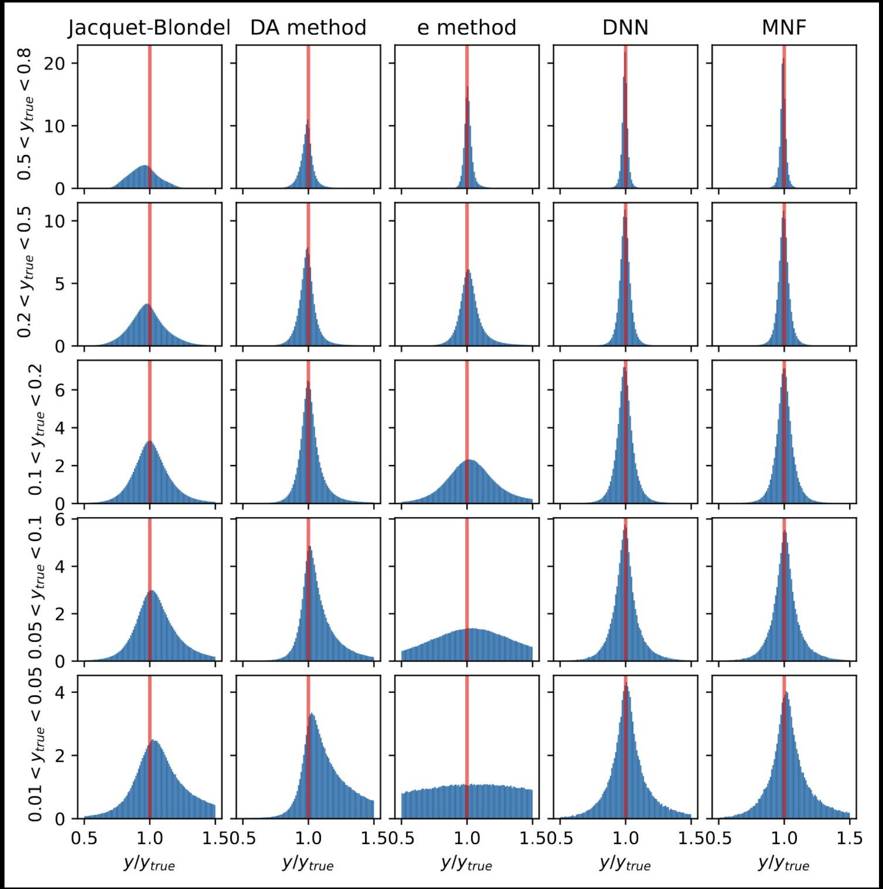


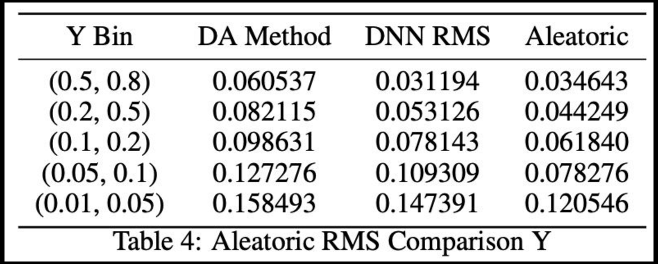


Y Bin	e Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.056694	0.044052	0.041349
(0.2, 0.5)	0.055787	0.037505	0.032280
(0.1, 0.2)	0.054219	0.033230	0.029640
(0.05, 0.1)	0.053403	0.032501	0.029411
(0.01, 0.05)	0.053470	0.032139	0.029431

Table 3: Aleatoric RMS Comparison - Q2

Aleatoric-RMS comparison

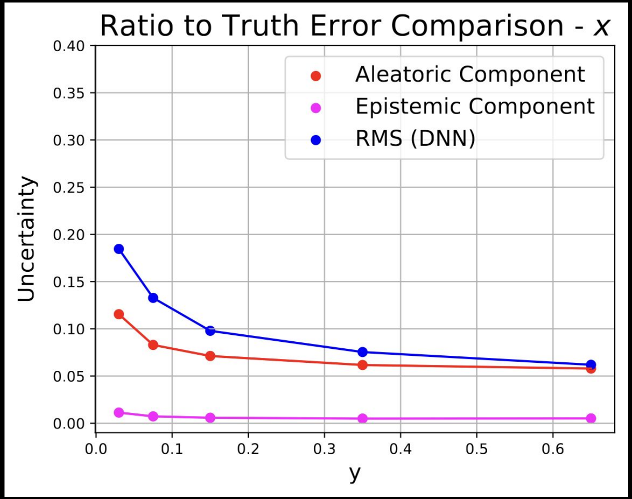


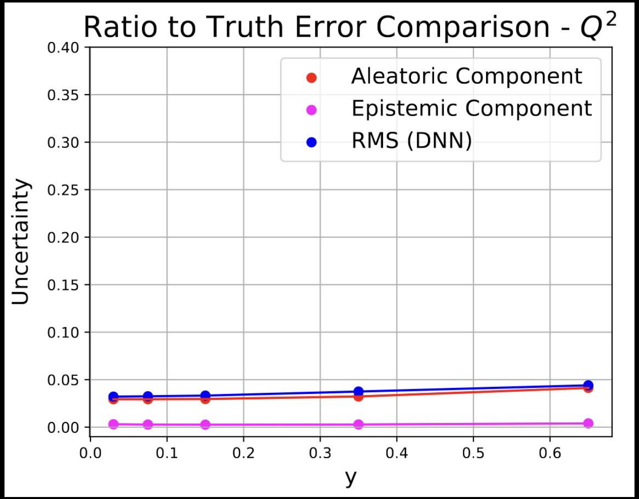


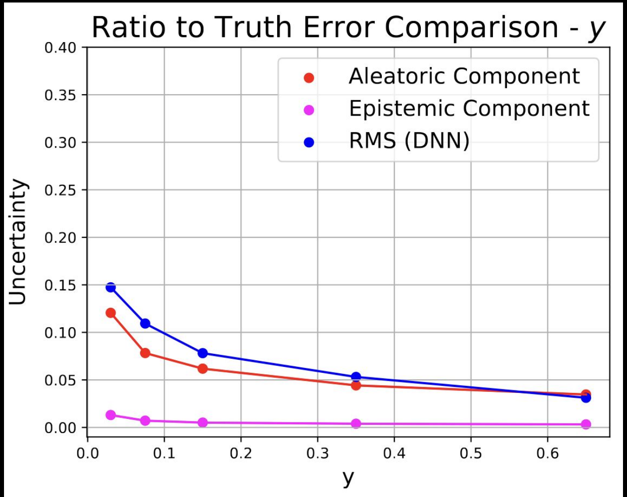
Y Bin	DA Method	DNN RMS	Aleatoric
(0.5, 0.8)	0.060537	0.031194	0.034643
(0.2, 0.5)	0.082115	0.053126	0.044249
(0.1, 0.2)	0.098631	0.078143	0.061840
(0.05, 0.1)	0.127276	0.109309	0.078276
(0.01, 0.05)	0.158493	0.147391	0.120546

Table 4: Aleatoric RMS Comparison Y

Comparison between DNN and BNN







- 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

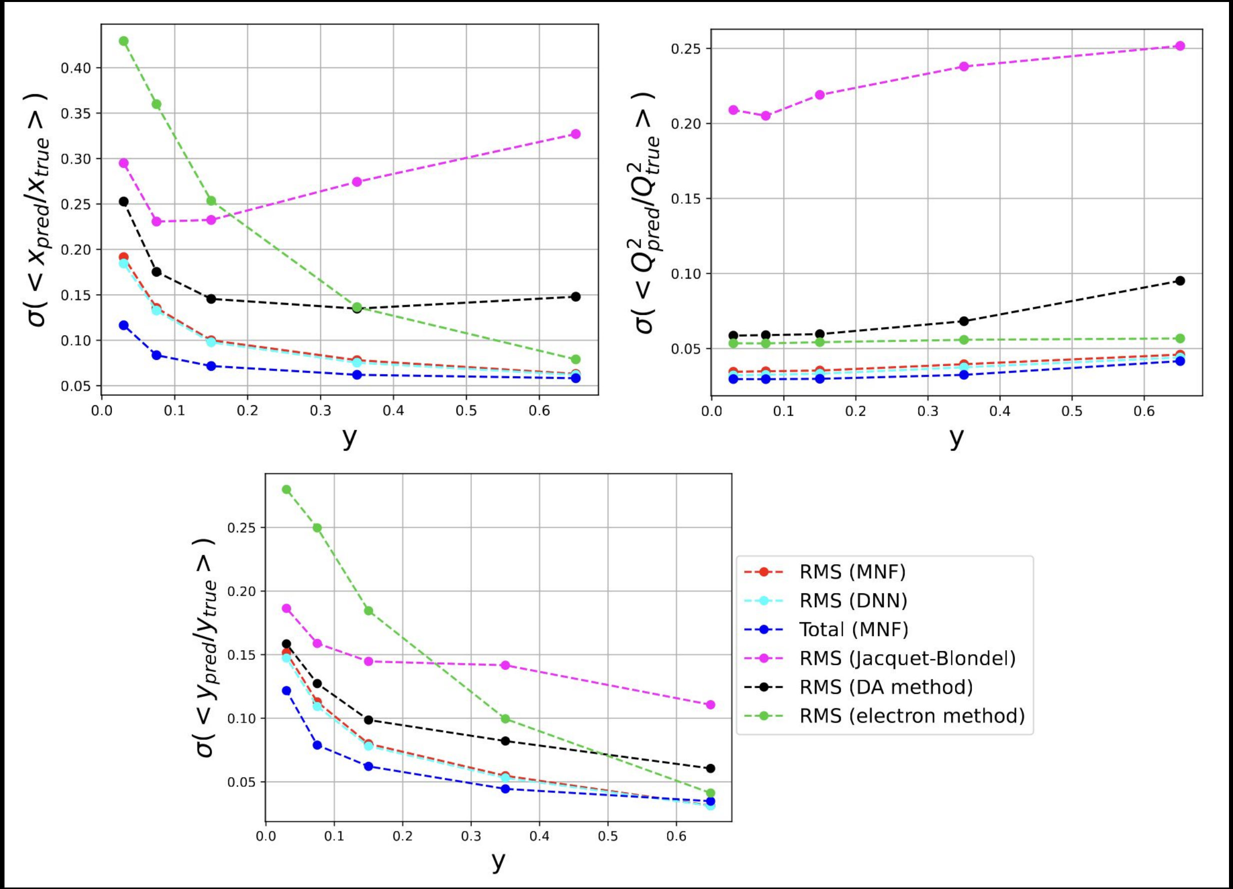
— Reporting uncertainty at the level of the event (e.g., RMS from other methods) —

All methods compared

- At low y, the RMS are typically larger due to “broader” distributions

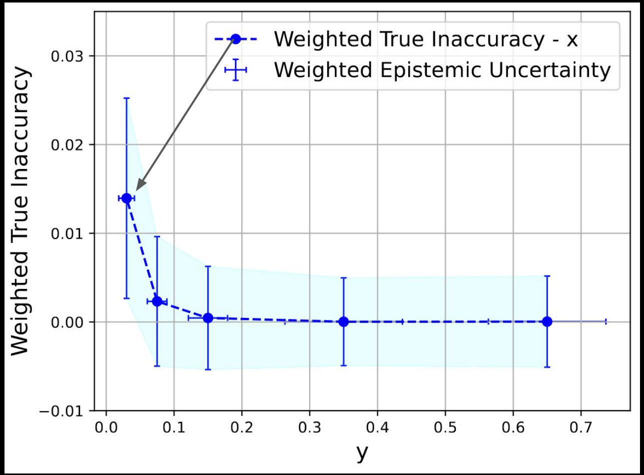
- DNN and MNF have smaller RMS over the whole y range compared to other methods (this was also the finding of NIM-A 1025 (2022): 166164) — “*our method outperforms other methods over a wide kinematics range*”

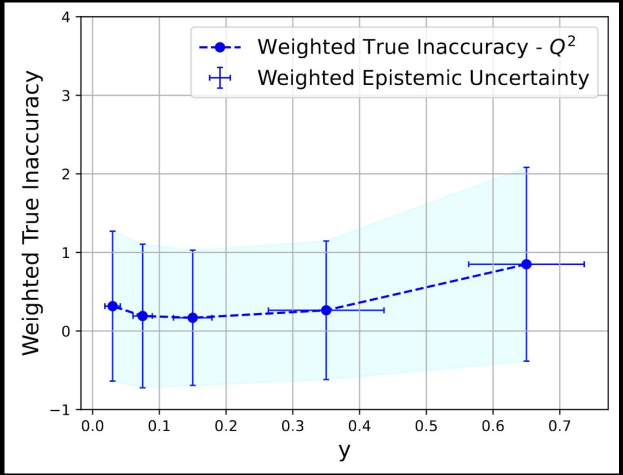
- “*The RMS resolution for y and x increase at lower y, even for the DNN reconstruction. … This results … may be attributed to further acceptance, noise, or resolution effects that deteriorates the measurement of the HFS*”

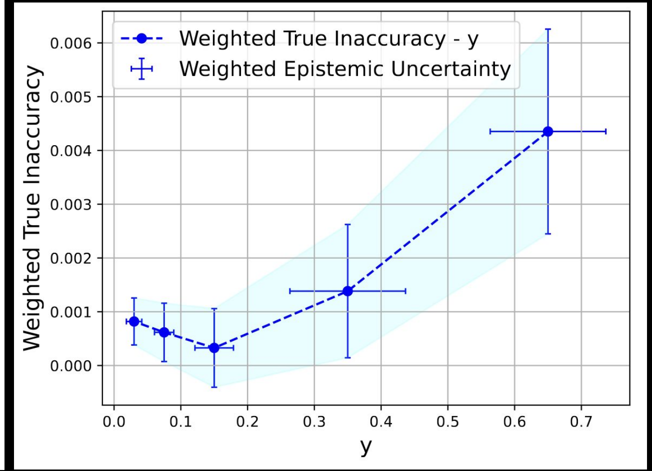


— Reporting uncertainty at the level of the event (e.g., RMS from other methods) —

Epistemic vs True Inaccuracy



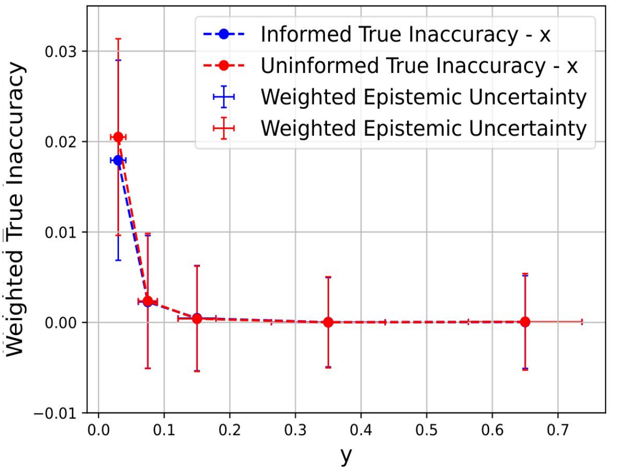


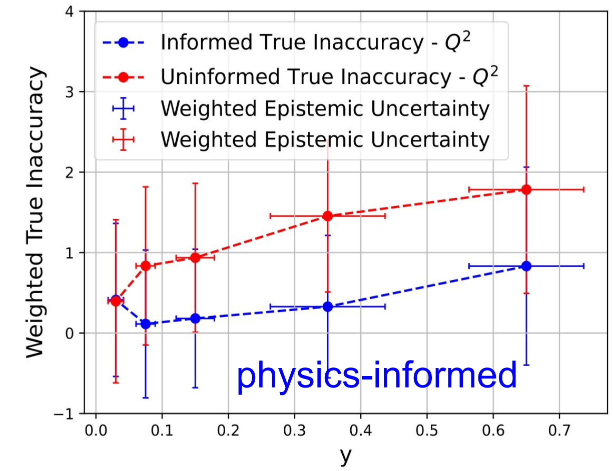


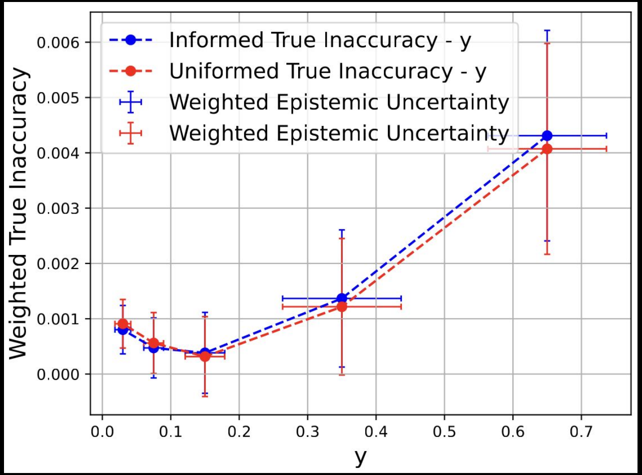
- The plots show that the epistemic uncertainty is larger when the true inaccuracy is larger

(N.b.: we are agnostic to the true inaccuracy)

Physics-informed term

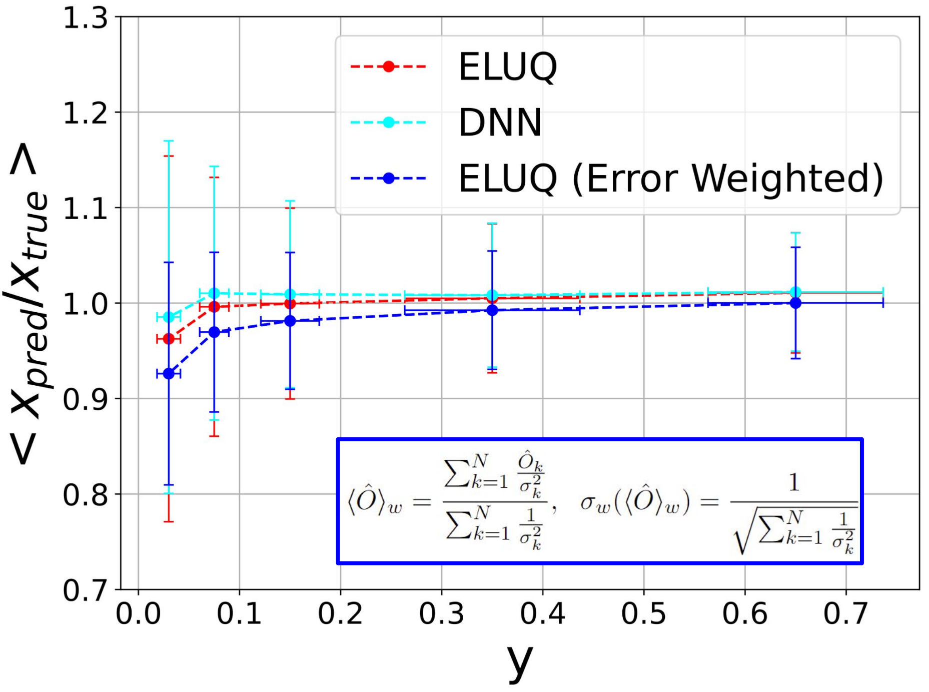






- The plots report the true inaccuracy, and the weighted epistemic uncertainty, which is larger the larger the true inaccuracy is
- The physics-informed term (blue) contributes to decrease the true inaccuracy.

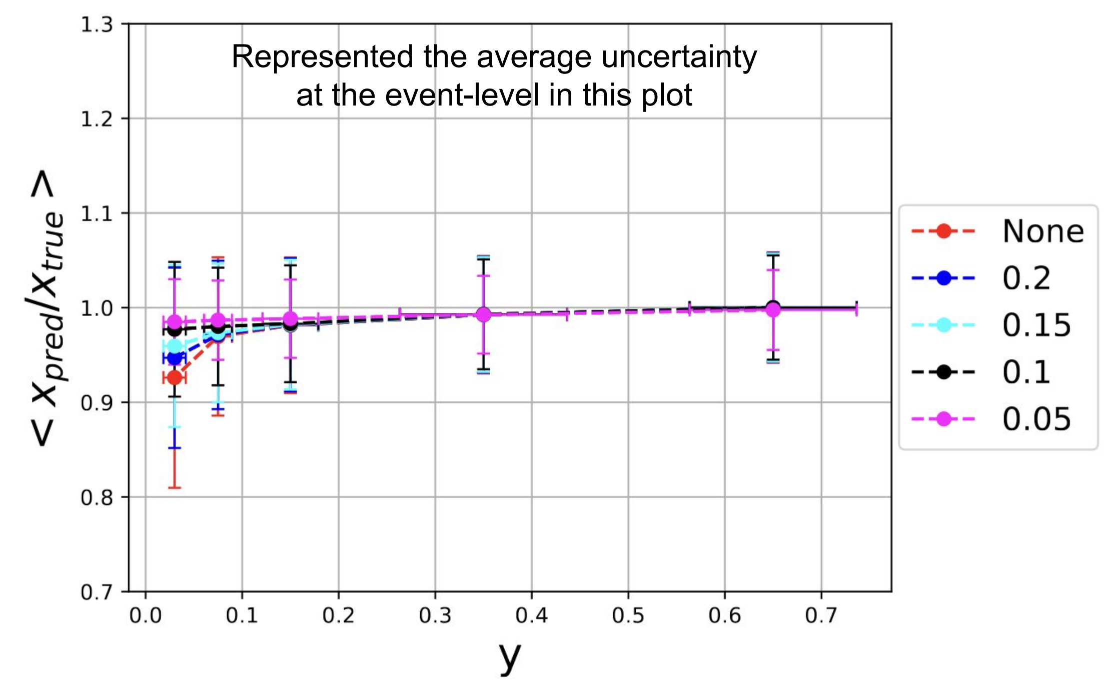
Leveraging event-level information



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

- In this plot, we are representing the average uncertainty at the event-level
- A “simple” DNN does not have per se uncertainty at the event level. In the plots we use the RMS from final distributions.
- ELUQ provides uncertainty on each event, individually. In the plots, we represent the average event calculated through a weighted average.

Leveraging event-level information

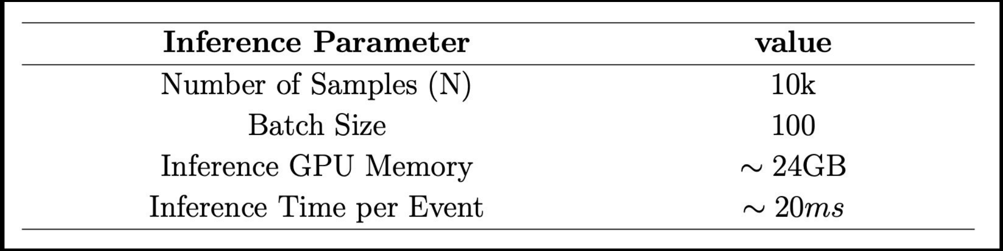


- Removing events with large relative event-level uncertainty (with respect to the network prediction) improve the ratio to truth and reduce inaccuracy
- Notice these cuts do not use any information at the ground truth level
- We know that ELUQuant is sensitive to anomaly detection. Performance studies are underway.

— N.b.: events with at least one among x,Q^2, y with relative uncertainty larger than a threshold are removed —

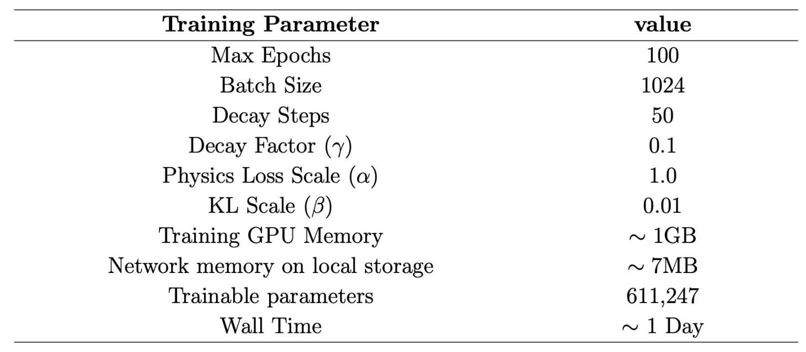
Time performance

- This is great, but what about computing time?



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

Inference specs of ELUQuant



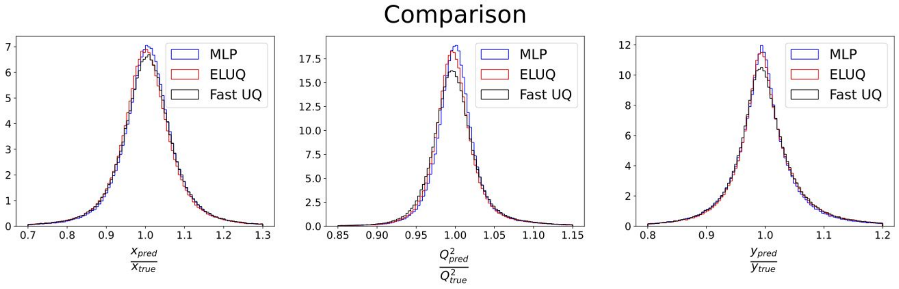
Training Parameter	value
Max Epochs	100
Batch Size	1024
Decay Steps	50
Decay Factor (γ)	0.1
Physics Loss Scale (α)	1.0
KL Scale (β)	0.01
Training GPU Memory	$\sim$ 1GB
Network memory on local storage	$\sim$ 7MB
Trainable parameters	611,247
Wall Time	$\sim$ 1 Day

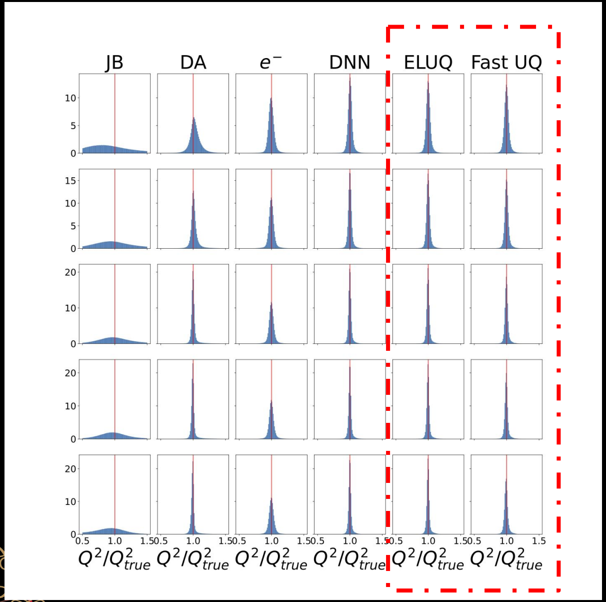
Inference specs of ELUQuant

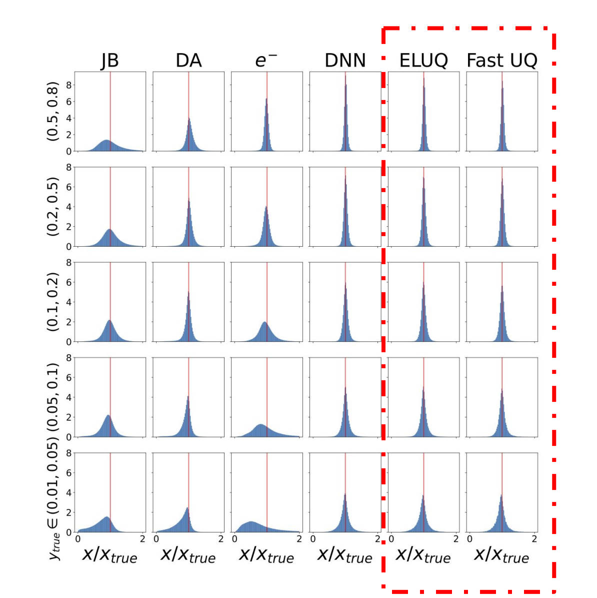
- In computational terms, ELUQuant at inference showed an impressive rate of 10,000 samples/event within a 20 milliseconds on an RTX 3090.

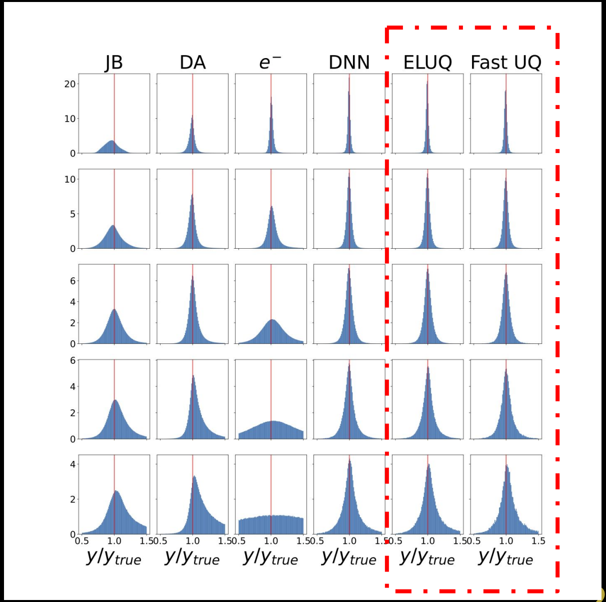
- Can we do faster than this?
  - Several ways. A rapid, streamlined approach is distilling this knowledge in a simpler but faster network (we explored a DNN with 450k parameters) called in the following “Fast UQ”, obtaining an effective inference time of 7-8us/event using batch ~0.5M events

Towards near real-time

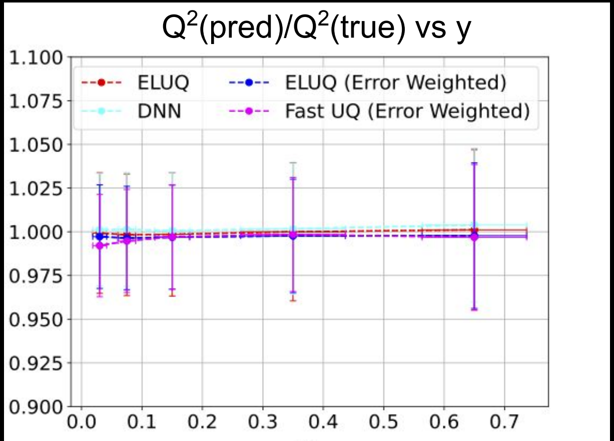


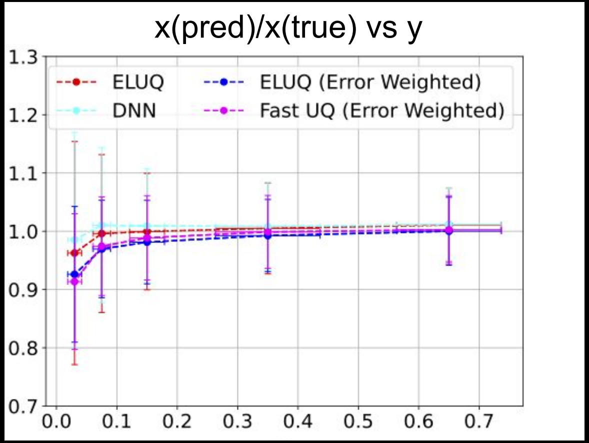


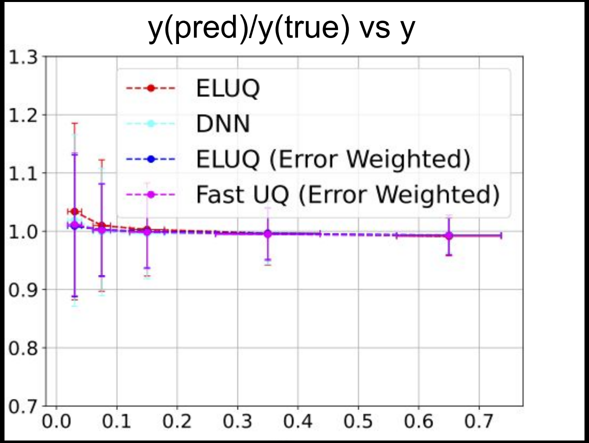


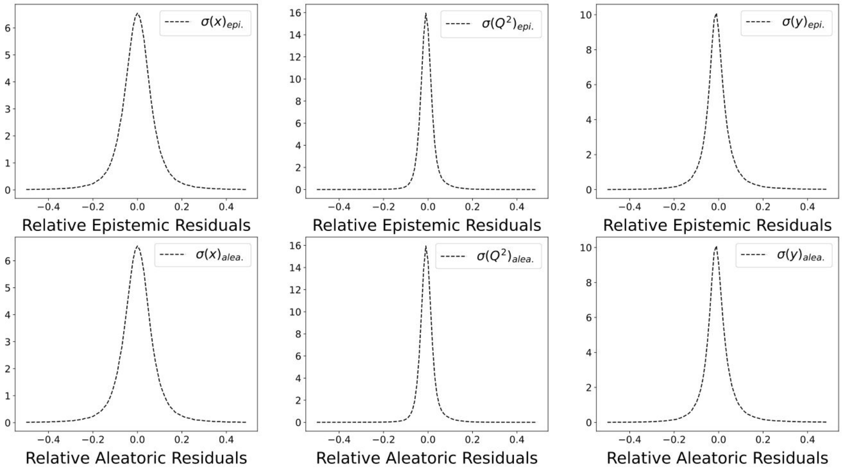


Towards near real-time


Q^2(pred)/Q^2(true) vs y


x(pred)/x(true) vs y


y(pred)/y(true) vs y



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

Summary

event(particle)-level, uncertainty quantification, near real-time

- The SRO approach unifies online and offline analyses, easing the integration of AI/ML for fast calibration and reconstruction, leading to rapid data processing and delivery of results.
- I highlighted the importance of UQ in ML/DL in general, and extended this argument to near real-time applications. This consideration is crucial for these models, especially at the event or particle level, and applies broadly to any ML/DL processing streamed data using lower-level features.
- I showed new results from ELUQuant, and show the possibility that UQ opens (accessing information we typically do not have at the event-level) in making decisions and predictions
- The inference performance of Bayesian architectures that address UQ improved in recent years with modern hardware (ELUQuant ~20ms/event on RTX 3090), and UQ can be (already) embedded in our data processing pipelines, in the larger scheme of having faster accurate data processing and analysis.
- We tried to speed up the computing time by distilling the knowledge of ELUQuant into a simpler and faster DNN architecture. We achieved accurate performance with effective inference times of 7-8 us/event