Arts & Sciences Articles

Arts and Sciences

ELUQuant: Event-level Uncertainty Quantification in Deep Inelastic Scattering

Cristiano Fanelli
William & Mary, cfanelli@wm.edu

James Giroux
William & Mary

MACHINE LEARNING
Science and Technology

PAPER

ELUQuant: event-level uncertainty quantification in deep inelastic scattering

C Fanelli[*] and J Giroux[*]

Department of Data Science, William & Mary, Williamsburg, VA 23185, United States of America
* Author to whom any correspondence should be addressed.

E-mail: cfanelli@wm.edu and jgiroux@wm.edu

Abstract

We introduce a physics-informed Bayesian neural network with flow-approximated posteriors using multiplicative normalizing flows for detailed uncertainty quantification (UQ) at the physics event-level. Our method is capable of identifying both heteroskedastic aleatoric and epistemic uncertainties, providing granular physical insights. Applied to deep inelastic scattering (DIS) events, our model effectively extracts the kinematic variables x, Q^2, and y, matching the performance of recent deep learning regression techniques but with the critical enhancement of event-level UQ. This detailed description of the underlying uncertainty proves invaluable for decision-making, especially in tasks like event filtering. It also allows for the reduction of true inaccuracies without directly accessing the ground truth. A thorough DIS simulation using the H1 detector at HERA indicates possible applications for the future electron–ion collider. Additionally, this paves the way for related tasks such as data quality monitoring and anomaly detection. Remarkably, our approach effectively processes large samples at high rates.

1. Introduction

In experimental nuclear physics (NP) and high-energy physics (HEP), data analyses typically regress fundamental quantities from observables measured in events produced and detected by experiments. A crucial aspect of these analyses is the corresponding event-level uncertainty quantification (UQ). The method introduced in this work ELUQuant (dubbed ELUQ in the figures), to our knowledge, pioneers this in NP/HEP by gleaning insights from computer vision [1] and multiplicative normalizing flows (MNFs) in Bayesian neural networks (BNNs) [2], effectively capturing both heteroskedastic aleatoric and epistemic uncertainties, which influence the regression of fundamental quantities from measured observables. Deep inelastic scattering (DIS) has recently benefited from deep learning techniques. An innovative study by Diefenthaler *et al* [3] employed deep neural networks (DNNs) to infer kinematic variables Q^2 and x of neutral current DIS from traditional reconstruction methods enhanced through correlations revealed in the simulated datasets of the ZEUS experiment. Successively, Arratia *et al* [4] applied DNNs, capitalizing on full kinematic information of both the scattered electron and the hadronic-final state (HFS), to reconstruct the kinematics of neutral–current DIS events, using H1 experiment simulations. Though both papers signify critical advancements in leveraging DNN for DIS, they did not delve into the domain of UQ. Our endeavor with ELUQuant distinctively bridges this aspect, and highlights the potential of detailed event-level UQ, a novelty among the referenced works. Our methodology harbors promise for any physics analysis demanding nuanced UQ.

This manuscript is structured as follows: section 2 introduces the DIS kinematics, the chosen case study, and both the quantified uncertainty sources (aleatoric and epistemic); section 3 delves into the ELUQuant architecture, detailing its loss function, training procedures, and inference performance; section 4 reports the results we obtained using H1's neutral current DIS Monte Carlo dataset also used in [4]. Conclusively, section 5 evaluates the broader impacts, emphasizing the effectiveness of ELUQuant in event-level UQ and its potential applications for data quality monitoring and anomaly detection.

2. Kinematic reconstruction of DIS

2.1. DIS

DIS is a reaction used to probe the internal structure of hadrons. In this process, high-energy leptons are scattered off hadrons, revealing intricate details about quarks and gluons [5]. Historically, the experiments conducted at the HERA collider, which remains the only electron–proton collider ever constructed, have been instrumental in DIS studies [6, 7]. The forthcoming electron–ion collider [8] promises to venture into previously uncharted regions of the DIS kinematic spectrum. Figure 1 depicts the DIS process, where k, k', and P are the four-momenta of the electrons and proton, respectively, and HFS is the hadronic final state.

DIS kinematics involve: squared four-momentum transfer $Q^2 = -q^2 = (k - k')^2$; inelasticity $y = \frac{q \cdot P}{k \cdot P}$, indicating the electron's energy fraction transferred to the nucleon; and Bjorken scaling $x = \frac{Q^2}{sy}$, showing the momentum fraction carried by the struck quark. The kinematic variables are related by $Q^2 = sxy$, where $s = (k + P)^2$ represents the squared center-of-mass energy. Momentum and energy conservation in DIS kinematics provide the ability to calculate x, Q^2, and y from measurements. Classical methods for their reconstruction differ (*see* [3, 4, 9]). We compare our results with methods such as electron (EL), double angle (DA), and Jacquet Blondel (JB). As highlighted in [3, 4], the DIS process can be influenced by several factors, such as initial-state and final-state radiation (ISR, FSR). Moreover, higher-order quantum electrodynamics (QED) and quantum chromodynamics (QCD) corrections can also manifest in the process. Each reconstruction method has its strengths across the phase space and sensitivities to radiative effects. For instance, EL uses only measurements of the scattered lepton and excels in high y scenarios but falters at low y. In contrast, JB uses only the HFS and performs better at low y. Hence, regression linking measured quantities to true kinematics is crucial. The true values of x, Q^2, and y in our data are derived from generator-level particle four-vectors, considering effects like ISR and FSR radiation.

2.2. Synthetic dataset and network input

We utilize full simulation from the H1 experiment that encompasses QED radiation and Lund hadronization model[1]. Table 1 summarizes the dataset statistics and size on disk. A total of 15 measured input features are used in our work and are sourced from [4]. These encompass seven features sensitive to QED radiation: $p_T^{bal} = 1 - \frac{p_{T,e}}{T}$ with T as the HFS transverse momentum and p_T the electron's; $p_z^{bal} = 1 - \frac{\Sigma_e + \Sigma}{2E_0}$, where $\Sigma_e = E - p_{z,e}$ and $\Sigma = \sum_i^{HFS}(E_i - p_{z,i})$; energy, η, and $\Delta\phi$ of the nearest photon to the electron beam direction, where $\Delta\phi$ is relative to the electron; E_{CAL}^{sum}/p_e, the ECAL energy sum within a cone of $\Delta R < 0.4$ around the scattered electron; and the count of ECAL clusters within $\Delta R < 0.4$. These seven are merged with another eight: scattered electron's $p_{T,e}, p_{z,e}, E$; the HFS four-vector components T, $P_{z,h}$, and E_h; $\Delta\phi(e, h)$, the angle between the scattered electron and HFS momentum; and the difference, $\Delta\Sigma = \Sigma_e - \Sigma$.

3. ELUQuant architecture

ELUQuant is applied to the DIS simulated dataset of H1, extracting x, Q^2, y, and their related epistemic and aleatoric uncertainties from 15 measured input features. Epistemic uncertainty stems from knowledge gaps, improving with more data and refined models. Aleatoric uncertainty, on the other hand, arises from inherent system variability and remains unaffected by additional data.

ELUQuant is a bicephalous regression network with Bayesian blocks characterized by MNF to approximate posteriors for event-level UQ. A representation of the architecture that enables building posteriors over the weights at each layer is shown in figure 2. Thus, when sampling, the result is a diverse combination of weights within the network. After training, each forward pass through the network yields a distinct set of weights drawn from the learned posterior.

To effectively handle uncertainty in Bayesian networks, the objective is to calculate a posterior, $q(\mathbf{W}|D)$, where $\mathbf{W}$ are the weights of the neural network. This allows predictions on the regressed quantities through a posterior distribution $q(\mathbf{y}|\mathbf{x}, D)$, integrated over the space of the weights. However, the formulation of such a posterior is intractable and therefore Bayesian inference must be employed. Typically, fully factorized Gaussians are assumed as an approximate posterior $q(\mathbf{W})$ such that we can minimize the evidence lower bound between the approximated posterior and the assumed prior. This is generally limiting and can underestimate true uncertainty. Another method is to utilize random auxiliary variables to improve the approximate posterior via a mixing density. In Louizos and Welling [1], they parameterize the mixing density in terms of auxiliary variables $\mathbf{z}$, which are in turn parameterized by normalizing flows to allow flexibility and local reparameterizations. They reduce the computational overhead of using a normalizing flow by allowing

---

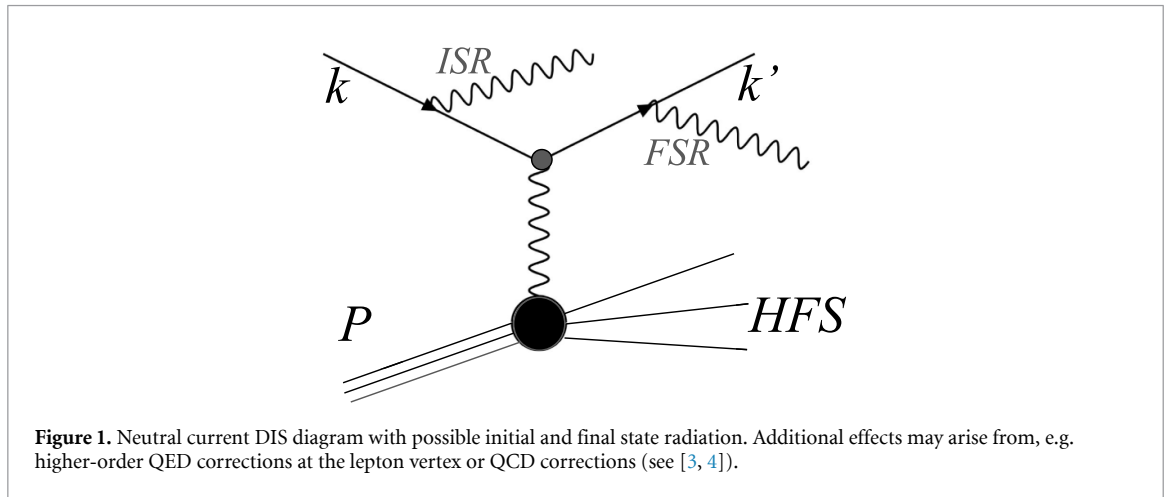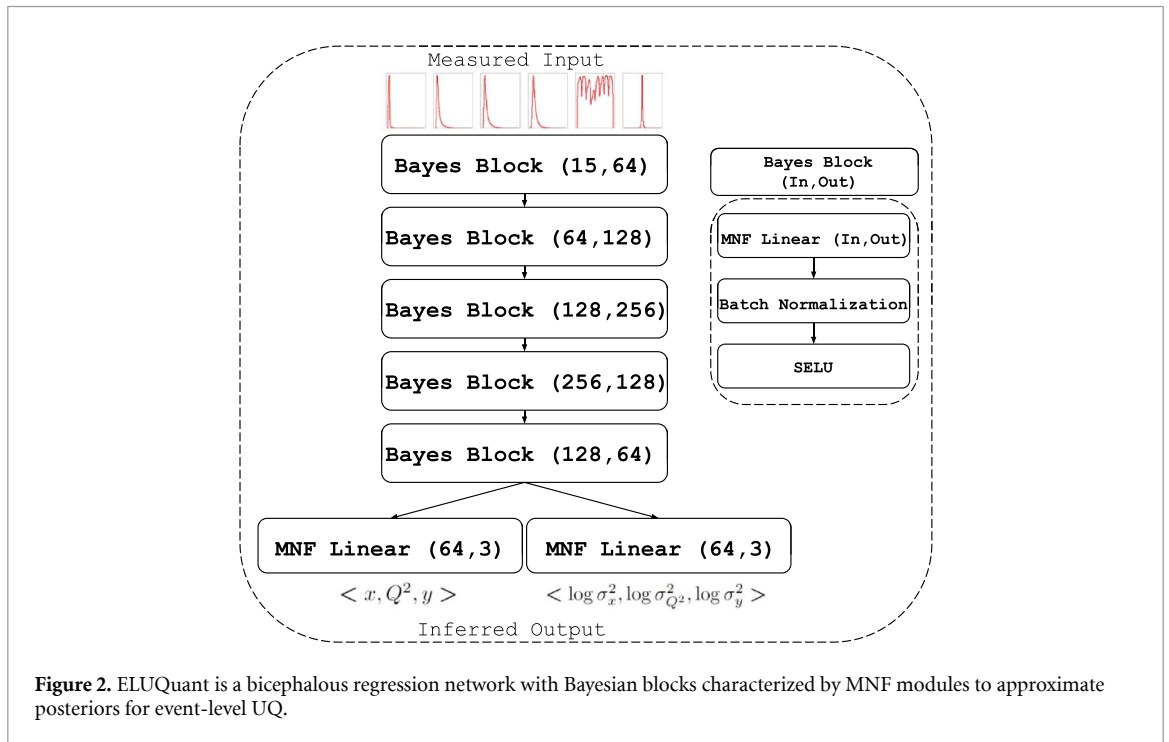[1] The same dataset has been utilized in [4].

Figure 1. Neutral current DIS diagram with possible initial and final state radiation. Additional effects may arise from, e.g. higher-order QED corrections at the lepton vertex or QCD corrections (see [3, 4]).

Table 1. Dataset statistics and size on disk.

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

Figure 2. ELUQuant is a bicephalous regression network with Bayesian blocks characterized by MNF modules to approximate posteriors for event-level UQ.

z to act multiplicatively on the means. Furthermore, the authors propose training the network under a variational inference paradigm, in contrast to more traditional BNNs which directly contain distributions of weights at each layer. Given a set of Gaussian weights, the pre-activation of neurons can be considered as a linear combination of the weights, which is in itself Gaussian. Louizos and Welling [1] further show that performing this sampling for each injection within a mini-batch results in a different set of weights, lower variance in gradients, and an overall more stable optimization. The sampling acts similarly to Gaussian dropout and effectively finds a distribution of optimal solutions within the space of the weights. An algorithmic description can be found in algorithm 1 of [1].

3.1. Loss function
The total loss function is the sum of different contributions:

$$\mathcal{L}_{\text{Tot.}} = \mathcal{L}_{\text{Reg.}} + \alpha \mathcal{L}_{\text{Phys.}} + \beta \mathcal{L}_{\text{KL.}} \tag{1}$$

The regression loss, equation (2), provides the DIS kinematic vector of observables we want to predict, namely $\mathbf{v} = (x, Q^2, y)$, as well as the corresponding heteroskedastic aleatoric term $\boldsymbol{\sigma} = (\sigma(x), \sigma(Q^2), \sigma(y))$:

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

The sum i runs over all vectors in the mini-batch, and the sum j runs over elements in the vector, where the epistemic term is captured by $\|\mathbf{v}_i - \hat{\mathbf{v}}_i\|$. The use of a logarithm at network output has been demonstrated in [2] to be more numerically stable than regressing σ^2, the variance[2]. Looking closely at equation (2), this is the logarithm of a multivariate normal distribution, see [2]. The physics-informed term, equation (3), is applied on the regressed observables which ideally should match the truth where $Q^2 = sxy$ holds:

$$\mathcal{L}_{\text{Phys.}} = \frac{1}{N} \sum_i \log \hat{Q}_i^2 - (\log s_i + \log \hat{x}_i + \log \hat{y}_i), \tag{3}$$

where the Mandelstam s is calculated at the ground truth level. The Kullback–Leibler (KL) term, equation (4), is adapted from [1], which employs MNF in variational BNNs to improve the posterior approximation

$$\begin{aligned}
\mathcal{L}_{\text{KL.}} &= -\text{KL}\left(q(\mathbf{W}) \| p(\mathbf{W}) \right) \\
&= \mathbb{E}_{q(\mathbf{W}, \mathbf{z}_T)} \left[-\text{KL}\left(q(\mathbf{W}|\mathbf{z}_{T_f}) \| p(\mathbf{W}) \right) + \log r(\mathbf{z}_{T_f}|\mathbf{W}) - \log q(\mathbf{z}_{T_f}) \right].
\end{aligned} \tag{4}$$

Given an assumed prior distribution over the weights, we wish to minimize the KL divergence between the prior and approximated posterior. However, the lower bound of the posterior is intractable, and therefore the entropy must be bounded via an auxiliary distribution $r(\mathbf{z}_{T_f}|\mathbf{W})$. The tightness of this bound depends directly on the auxiliary distribution's ability to approximate the posterior, and therefore an additional normalizing flow is used to allow flexibility [1]. We employ Gaussian priors and compute the posterior distribution as the product of a Gaussian and a mixing density parameterized by a normalizing flow. As shown in [1], such a parameterization is flexible, allowing nonlinear and multimodal dependencies between the weight elements.

The SELU activation functions, as presented by Klambauer *et al* [10], are employed for their inherent self-normalizing properties, which ensure non-vanishing gradients. Their self-normalization nature could provide cases in which batch normalization is not needed, although this is data-dependent. We utilize SELU along with batch normalization to improve network convergence [10]. We also note that [4] utilizes these activation functions.

3.2. Training

Training and inference are performed utilizing a Python 3.9.12 environment with PyTorch 1.12.1 and CUDA 11.3. The model is trained for a maximum of 100 epochs, utilizing a batch size of 1028, in which training is stopped early if validation loss has plateaued. The model is trained using the *Adam* optimizer with an initial learning rate of 5×10^{-4}, and is subject to a stepped learning rate function in which we decay (γ) by an order of magnitude every 50 epochs (step size). It was found that decreasing the learning rate in such a fashion allows the network to converge to a more stable lower value. The initial learning rate was optimized for faster convergence in the early epochs, reducing fluctuations in loss caused by instability in back-propagation from large weight updates. During training, it is important to correctly weight the KL loss contribution in such a way that it does not dominate, yet allows the convergence to informative posteriors. This also holds true in the relationship between the physics informed and regression losses. Optimal values of $\alpha = 1.0$ and $\beta = 0.01$ were found through simple grid search optimization schemes. The total dataset size is $\sim$12 million events, split into a standard $70\%, 15\%, 15\%$ training, validation and testing split, providing $\sim$2 million events for testing purposes. Data injected to the network is scaled on the interval $(-1, 1)$, and targets are also scaled into the same interval. Table 2 provides the specs for training.

3.3. Inference

At inference we sample each individual event 10k times, taking the mean value as our final prediction. The epistemic uncertainty component is given by the standard deviation in these predictions. Each individual inference will also provide the corresponding aleatoric uncertainty, in which we again take the mean as our final aleatoric component. We perform this in batches of size 100, which corresponds to an overall batch of 1 million, resulting in an event-level inference time of 20 ms. Note that the pipeline could be further vectorized to further decrease wall time. Table 3 provide the specs for inference.

---

[2] Note that we do not apply a logarithmic activation directly; instead, we interpret linear activations in terms of logarithms.

Table 2. Training specs of the ELUQuant architecture: training was performed with an Intel i7-12700k 12 core CPU, Nvidia RTX 3090 24 GB GPU, and 64 GB memory.

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$1 GB
Network memory on local storage	$\sim$7 MB
Trainable parameters	611 247
Wall time	$\sim$1 day

Table 3. Inference specs of the ELUQuant architecture. Same computing resources of table 2.

Inference parameter	Value
Number of samples (N)	10k
Batch size	100
Inference GPU memory	$\sim$24 GB
Inference time per event	$\sim$20 ms

3.4. Limitations

Identifying a suitable network structure for a Bayesian network is generally not straightforward given that the model, in essence, is attempting to optimize a distribution of weights at each layer. The problem has been alleviated by utilizing the network's deterministic counterpart (a DNN) to identify a minimal complexity model that provides acceptable performance. We then utilized the structure from this network in ELUQuant. Another aspect to consider, depending on the data, during the initial stages of training, is the possibility that a poorly calibrated model produces poor regression targets with small aleatoric components, resulting in unstable fluctuations. In light of this, we bounded the minimum and maximum variances to improve training stability. The numerical values of the bounds were set such that they do not influence the learned aleatoric component, i.e. the network does not default to the minimum or maximum allowed value.
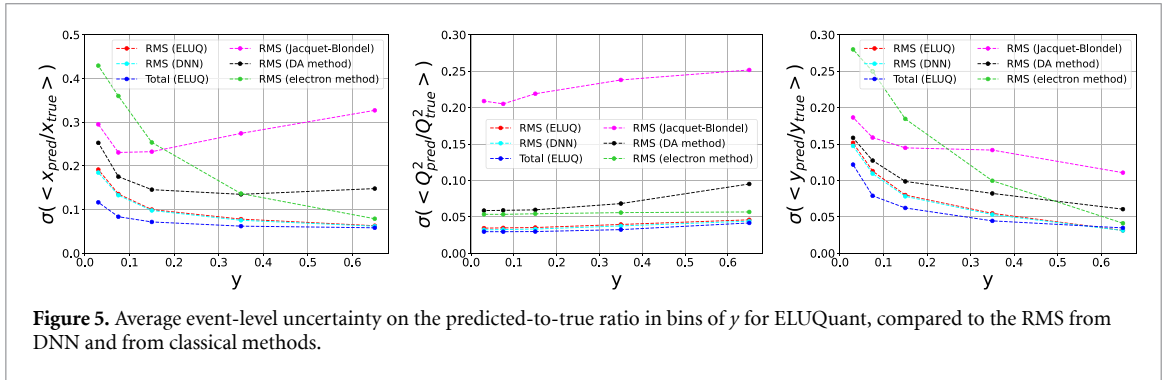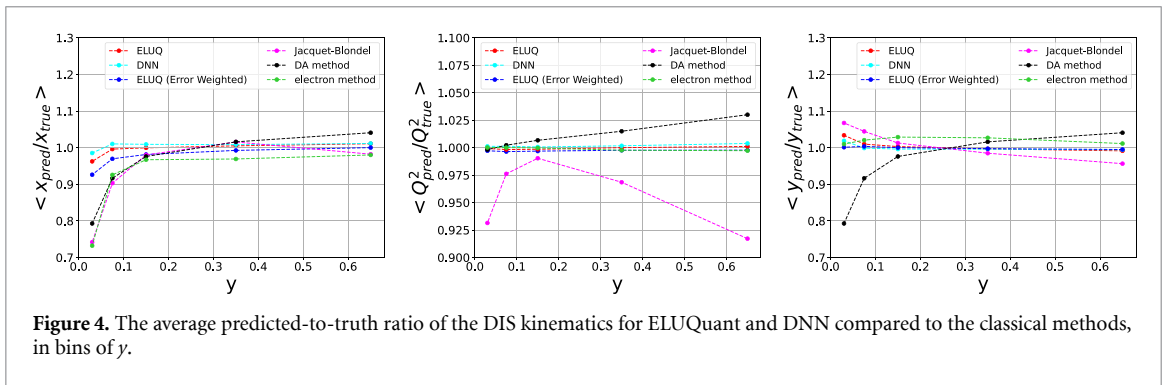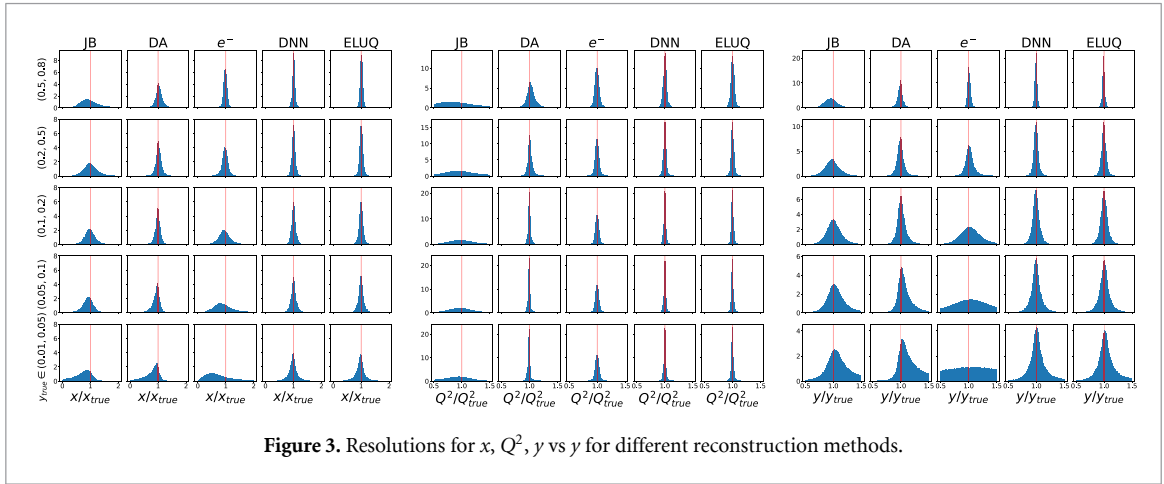
4. Analysis and results

Our strategy began with training a streamlined DNN that, despite its reduced complexity compared to [4] (150k parameters compared to 1.2M), achieved similar performance. While ELUQuant and the DNN share similar architectural layer sizes, ELUQuant stands out by offering enhanced UQ not possible with a basic DNN. We utilized ELUQuant to predict the DIS kinematic variables and their associated aleatoric and epistemic uncertainties. In what follows, we make comparisons with other traditional reconstruction methods, namely EL, JB, and DA, introduced in section 2. We will also incorporate DNN into the comparative visualizations. The section will be split into two parts. Section 4.1 will discuss the general performance of the architecture in relation to other methods, similar to what is done in [4], and section 4.2 will provide detailed studies on the uncertainties produced by our model, and how their utilization can result in increased performance.

4.1. Regression performance

Figure 3 shows resolutions for x, Q^2, y in bins of y and comparison among the various reconstruction methods.

We can immediately notice that the distributions of DNN and ELUQuant look alike over the whole range in y and for all the DIS kinematic variables. The choice of the binning in y is to reproduce and compare with the results in [4]. We also notice that DNN and ELUQuant outperform traditional methods. As expected, the traditional methods do perform differently as a function of y: for example, the methods that mostly rely on the scattered electron yield the best resolution in events with large y, but their resolution on x quickly diverges at low y. As already discussed, with ELUQuant we can calculate uncertainties at the event-level. Given an observable $\hat{O}_k$ for the kth event and its associated uncertainty σ_k, the weighted average of the observable over the entire dataset using uncertainty level information, and its associated uncertainty, are given by:

Figure 3. Resolutions for x, Q^2, y vs y for different reconstruction methods.

Figure 4. The average predicted-to-truth ratio of the DIS kinematics for ELUQuant and DNN compared to the classical methods, in bins of y.

Figure 5. Average event-level uncertainty on the predicted-to-true ratio in bins of y for ELUQuant, compared to the RMS from DNN and from classical methods.

$$\langle \hat{O} \rangle_{\mathrm{w}} = \frac{\sum_{k=1}^{N} \frac{\hat{O}_k}{\sigma_k^2}}{\sum_{k=1}^{N} \frac{1}{\sigma_k^2}}, \quad \sigma_{\mathrm{w}}\left(\langle \hat{O} \rangle_{\mathrm{w}}\right) = \frac{1}{\sqrt{\sum_{k=1}^{N} \frac{1}{\sigma_k^2}}}. \tag{5}$$

While other methods do not provide direct access to event-level uncertainty, comparisons between methods are still feasible. To facilitate this, the expected event-level uncertainty is approximated using the RMS as detailed in [4]. The RMS is then compared to the event-level equivalent derived from the weighted uncertainty, given by $\approx \sigma_{\mathrm{w}} \cdot \sqrt{N}$. Notice that due to the large statistics, the uncertainty on the averages will be exceedingly small and may be challenging to visually discern otherwise.

Figures 4 and 5 show the (weighted) average ratio of the predicted observables normalized to their ground truth, $< R_O > = < \hat{O}_{\mathrm{pred.}}/\hat{O}_{\mathrm{true}} >$, and the event-level uncertainties, respectively, in bins of the inelasticity y. In particular, figure 4 shows a drop in the $< R_x >$ at low y, where the RMS resolution for y and x increase, even for the DNN and ELUQuant reconstruction, as shown in figure 5. According to [4], these results may be attributed to further acceptance, noise, or resolution effects that deteriorate the measurement of the HFS. Notice that the weighted average is slightly more affected by this flaw in reconstruction performance than the arithmetic average.
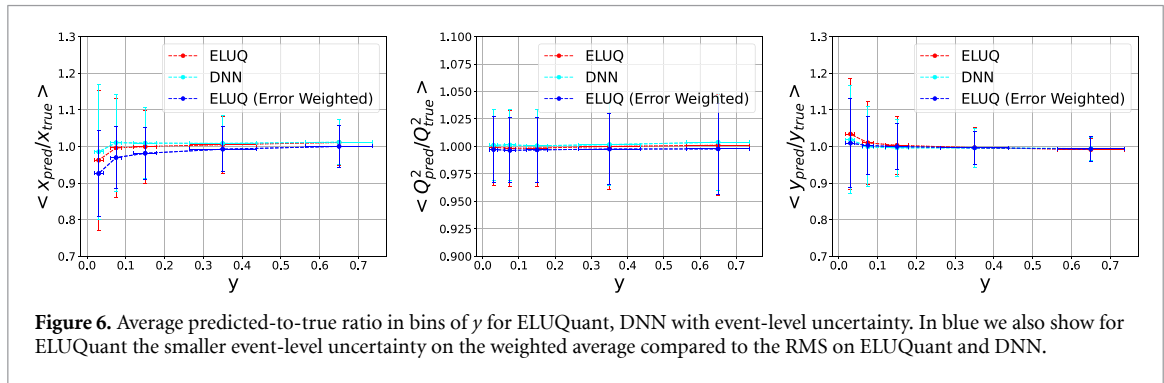
Figure 6. Average predicted-to-true ratio in bins of y for ELUQuant, DNN with event-level uncertainty. In blue we also show for ELUQuant the smaller event-level uncertainty on the weighted average compared to the RMS on ELUQuant and DNN.
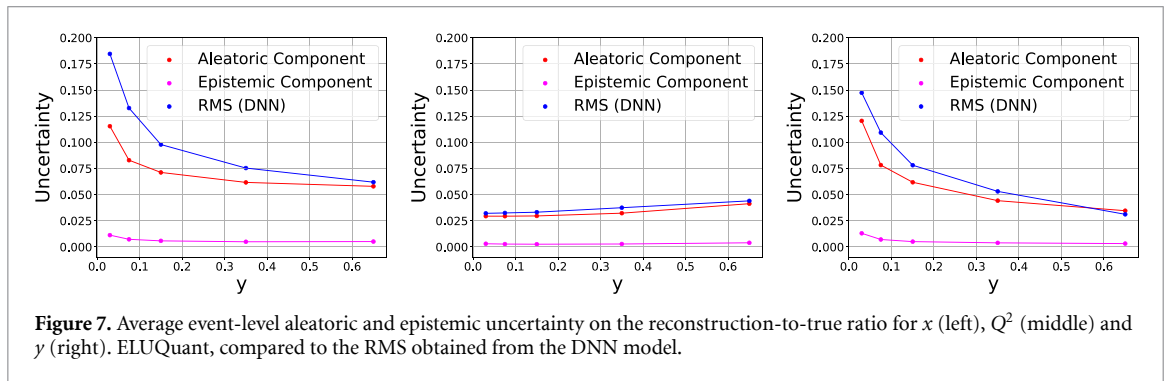

Figure 7. Average event-level aleatoric and epistemic uncertainty on the reconstruction-to-true ratio for x (left), Q^2 (middle) and y (right). ELUQuant, compared to the RMS obtained from the DNN model.

Figure 6 shows a comparison of the ratios between ELUQuant and DNN; for ELUQuant, we report both the RMS and the event-level equivalent weighted uncertainty. Notice that the total uncertainty at the event-level for ELUQuant is given by the sum in quadrature of the aleatoric and epsitemic components, *i.e.* $\sigma_{\text{tot}} = \sigma_{\text{ale}} \oplus \sigma_{\text{epi}}$.

4.2. Uncertainty analysis

The validation criteria of our model are two-fold. In the previous section, we validated regression performance in comparison to the model's deterministic counterpart (DNN), both with and without the inclusion of information from uncertainties. Showing the benefits of access to event-level uncertainty in relation to performance. In what follows we validate the event-level uncertainty components individually. We conduct a series of closure tests on the aleatoric component to show the event-level quantities propagate correctly at the histogram level. We also conduct closure tests on the epistemic component in which we show the uncertainty generated by our model decreases as a function of model calibration.

Figure 7 shows a comparison between σ_{ale}, σ_{epi} and the RMS from DNN, for the three regressed variates x, (left), Q^2 (middle), y (right). Figure 8 presents a detailed analysis of the histograms representing event-level occurrences of σ_{ale} and σ_{epi} uncertainties on x, Q^2, and y. These uncertainties are examined in bins of y.

Closure tests support the reliability of the aleatoric and epistemic uncertainties extracted. For instance, as shown in table 4, aleatoric uncertainties are consistent with the RMS of a DNN in bins of y (visually depicted in figure 7) where the regressed observables manifest as Gaussian distributions not affected by inaccuracy, that is, centered at the expected mean from ground truth. Notably, epistemic uncertainty—originating from the same multivariate normal distribution characterizing the aleatoric term in the loss function—amplifies in response to increased inaccuracy with respect to ground truth, see figure 9. UQ studies have also been conducted to demonstrate the effect of the physics-informed term on the inaccuracy $|\mathbf{v} - \hat{\mathbf{v}}|$. Figure 10 shows that equation (3) contributes to a decrease in the inaccuracy on Q^2; it also confirms that the epistemic increases if the inaccuracy gets larger. We also demonstrate how event-level UQ can be employed to assess the quality of events, retaining those with higher confidence and discarding events with more pronounced uncertainties. Figure 11 shows the effect of cutting events with large relative uncertainty using different thresholds. By excluding events with higher uncertainty, we mitigate the observed drop in the predicted-to-true ratio for the variable x. It is worth reminding that these cuts are agnostic to the ground truth.

However, this approach results in a reduction of statistics. Figure 12 illustrates the count of discarded events in relation to the severity of the cuts, segmented by bins of y. The loosest cut removes 40% of the events in the lowest bin in y, predominantly influenced by high aleatoric uncertainty in x.
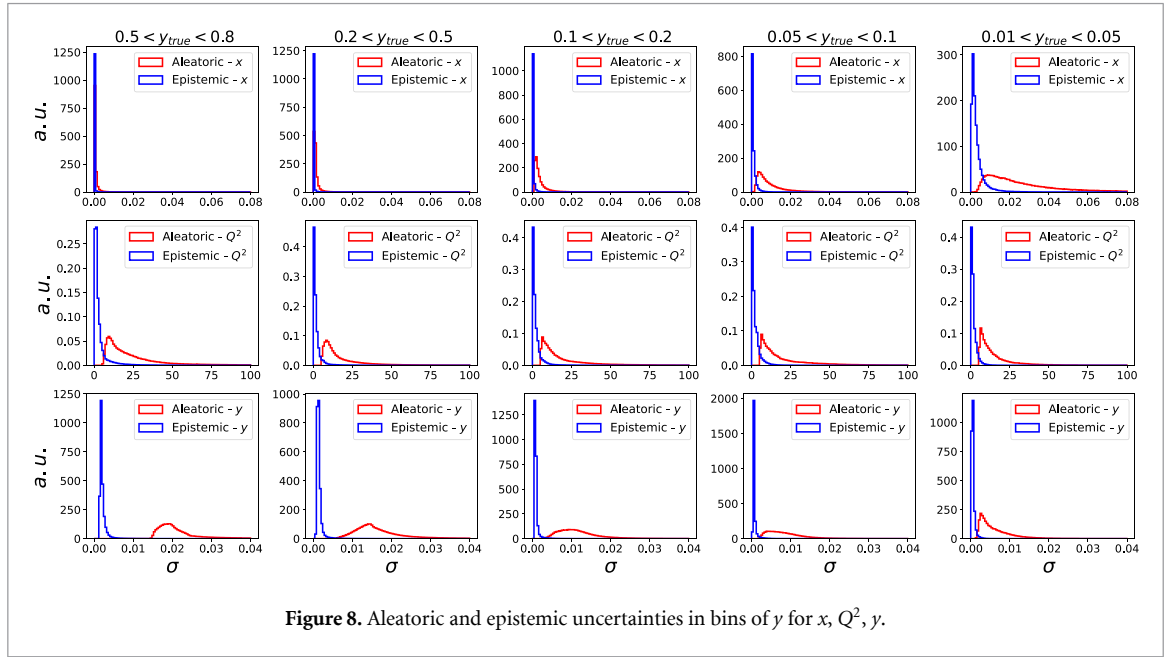
Figure 8. Aleatoric and epistemic uncertainties in bins of y for x, Q^2, y.

Table 4. Comparisons between the aleatoric uncertainty of ELUQuant with the RMS of other methods, for the DIS kinematic variables x, Q^2, y.

| y bin	RMS (x_{DA})	RMS (x_{ele})	RMS (x_{DNN})	$\sigma(x)$	RMS (Q^2_{DA})	RMS (Q^2_{ele})	RMS (Q^2_{DNN})	$\sigma(Q^2)$	RMS (y_{DA})	RMS (y_{ele})	RMS (y_{DNN})	$\sigma(y)$
(0.5, 0.8)	0.15	0.079	0.062	**0.058**	0.095	0.057	0.044	**0.041**	0.061	0.041	0.031	**0.035**
(0.2, 0.5)	0.13	0.14	0.075	**0.062**	0.068	0.056	0.038	**0.032**	0.082	0.100	0.053	**0.044**
(0.1, 0.2)	0.15	0.25	0.098	**0.071**	0.060	0.054	0.033	**0.030**	0.099	0.18	0.078	**0.062**
(0.05, 0.1)	0.18	0.36	0.13	**0.083**	0.059	0.053	0.033	**0.029**	0.13	0.25	0.11	**0.078**
(0.01, 0.05)	0.25	0.43	0.18	**0.12**	0.059	0.053	0.032	**0.029**	0.16	0.28	0.15	**0.12**
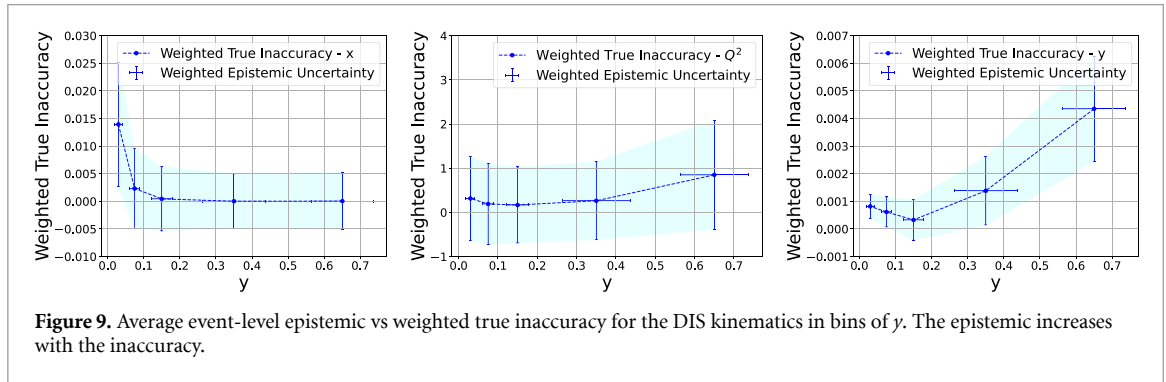
Figure 9. Average event-level epistemic vs weighted true inaccuracy for the DIS kinematics in bins of y. The epistemic increases with the inaccuracy.
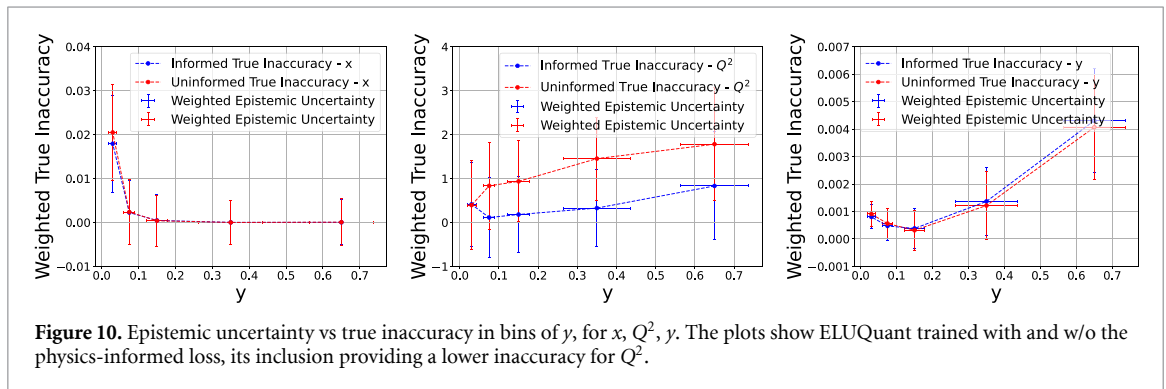

Figure 10. Epistemic uncertainty vs true inaccuracy in bins of y, for x, Q^2, y. The plots show ELUQuant trained with and w/o the physics-informed loss, its inclusion providing a lower inaccuracy for Q^2.
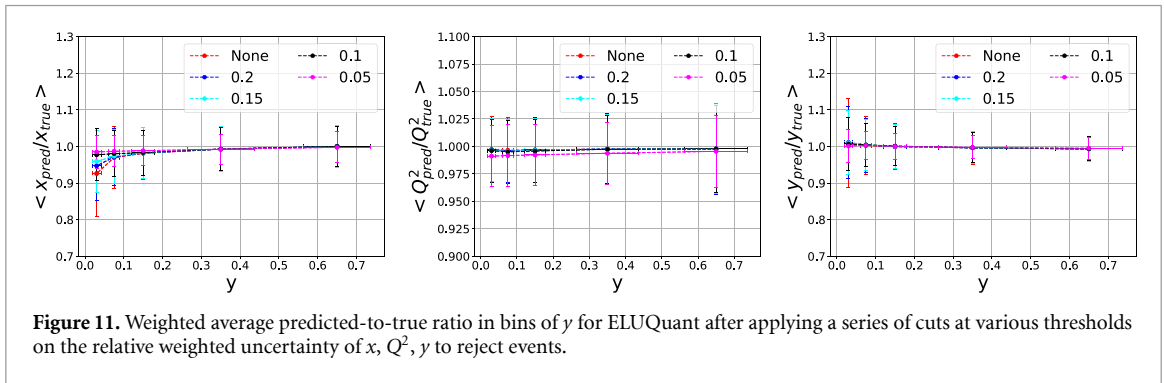
Figure 11. Weighted average predicted-to-true ratio in bins of *y* for ELUQuant after applying a series of cuts at various thresholds on the relative weighted uncertainty of *x*, Q^2, *y* to reject events.
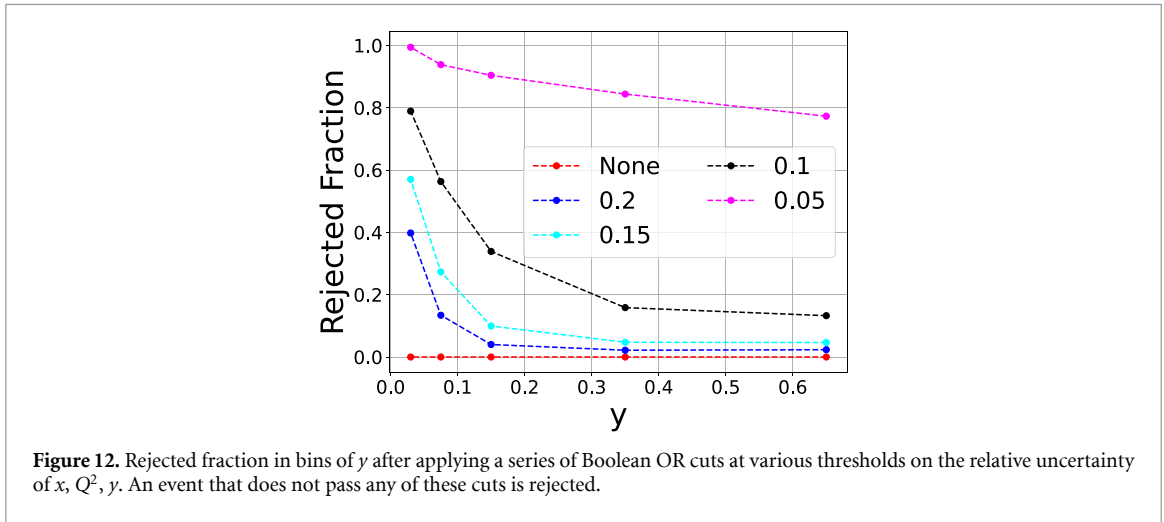

Figure 12. Rejected fraction in bins of *y* after applying a series of Boolean OR cuts at various thresholds on the relative uncertainty of *x*, Q^2, *y*. An event that does not pass any of these cuts is rejected.

5. Conclusions

We present ELUQuant, a novel network that integrates physics-informed BNNs with flow-approximated posteriors, marking a major advancement in physics analyses and uniquely providing insights into both heteroskedastic aleatoric and epistemic uncertainties on an event-level basis. To our knowledge, this is a pioneering achievement in the field, realizing a long-sought benchmark. Validated by results from the H1 detector's DIS simulation at HERA, our work suggests promising future extensions to the upcoming EIC for extracting essential kinematic observables, which could be affected by radiation effects, and their associated uncertainties. Closure tests support the reliability of the aleatoric and epistemic uncertainties extracted. For instance, aleatoric uncertainties align with the RMS of a DNN in *y*-regions where the regressed observables manifest as Gaussian distributions not affected by inaccuracy, centered at the expected mean from ground truth. Notably, epistemic uncertainty—originating from the same multivariate normal distribution characterizing the aleatoric term in the loss function—amplifies in response to increased inaccuracy with respect to ground truth. While the impact of ELUQuant for DIS data is evident, its versatility extends to a broader range of event-level physics analyses. The granularity ELUQuant offers can revolutionize event filtering decision-making. Informed by uncertainties, it can mitigate true inaccuracies, showing promise in both data quality monitoring and anomaly detection. In computational terms, our approach at inference showed an impressive rate of 10 000 samples/event within a mere 20 ms on an RTX 3090, emphasizing real-world application viability. In essence, ELUQuant's pioneering approach to event-level UQ sets a new standard for comprehensive analyses in NP and particle physics.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

ORCID iDs

C Fanelli ● https://orcid.org/0000-0002-1985-1329
J Giroux ● https://orcid.org/0000-0001-6487-7870

References

[1] Louizos C and Welling M 2017 Multiplicative normalizing flows for variational Bayesian neural networks *Proc. 34th Int. Conf. on Machine Learning* (Proc. Machine Learning Research vol 70) ed D Precup and Y W Teh (PMLR) pp 2218–27

[2] Kendall A and Gal Y 2017 What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems* vol 30 (Curran Associates, Inc.)

[3] Diefenthaler M, Farhat A, Verbytskyi A and Xu Y 2022 Deeply learning deep inelastic scattering kinematics *Eur. Phys. J. C* **82** 1064

[4] Arratia M, Britzger D, Long O and Nachman B 2022 Reconstructing the kinematics of deep inelastic scattering with deep learning *Nucl. Instrum. Methods Phys. Res.* A **1025** 166164

[5] Devenish R and Cooper-Sarkar A 2004 *Deep Inelastic Scattering* (Oxford University Press)

[6] Abt I *et al* 1997 The H1 detector at HERA *Nucl. Instrum. Methods Phys. Res.* A **386** 310–47

[7] Abramowicz H *et al* 2015 Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data: H1 and ZEUS Collaborations *Eur. Phys. J. C* **75** 1–98

[8] Khalek R A *et al* 2022 Science requirements and detector concepts for the electron-ion collider: EIC yellow report *Nucl. Phys. A* **1026** 122447

[9] Bassler U and Bernardi G 1995 On the kinematic reconstruction of deep inelastic scattering at HERA *Nucl. Instrum. Methods Phys. Res.* A **361** 197–208

[10] Klambauer G, Unterthiner T, Mayr A and Hochreiter S 2017 Self-normalizing neural networks *Advances in Neural Information Processing Systems* vol 30, ed I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (Curran Associates, Inc.)