MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™ Part of the MACHINE LEARNING Series

PAPER • OPEN ACCESS

Evidential deep learning for uncertainty quantification and out-of-distribution detection in jet identification using deep neural networks

To cite this article: Ayush Khot *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 035003

IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 035003 https://doi.org/10.1088/2632-2153/ade51b

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
9 January 2025

REVISED
23 May 2025

ACCEPTED FOR PUBLICATION
16 June 2025

PUBLISHED
8 July 2025

PAPER

Evidential deep learning for uncertainty quantification and out-of-distribution detection in jet identification using deep neural networks

Ayush Khot[1], **Xiwei Wang**[2], **Avik Roy**[3], **Volodymyr Kindratenko**[2,3,4] and **Mark S Neubauer**[1,2,3,*]

[1] Department of Physics, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States of America
[2] The Grainger College of Engineering, Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States of America
[3] Center for Artificial Intelligence Innovation, National Center for Supercomputing Applications, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States of America
[4] The Grainger College of Engineering, Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: msn@illinois.edu

Keywords: jet classification, machine learning, deep learning, evidential deep learning, uncertainty quantification, anomaly detection

Abstract

Current methods commonly used for uncertainty quantification (UQ) in deep learning (DL) models utilize Bayesian methods which are computationally expensive and time-consuming. In this paper, we provide a detailed study of UQ based on evidential DL (EDL) for deep neural network models designed to identify jets in high energy proton–proton collisions at the Large Hadron Collider and explore its utility in anomaly detection (AD). EDL is a DL approach that treats learning as an evidence acquisition process designed to provide confidence (or epistemic uncertainty) about test data. Using publicly available datasets for jet classification benchmarking, we explore hyperparameter optimizations for EDL applied to the challenge of UQ for jet identification. We also investigate how the uncertainty is distributed for each jet class, how this method can be implemented for the detection of anomalies, how the uncertainty compares with Bayesian ensemble methods, and how the uncertainty maps onto latent spaces for the models. Our studies uncover some pitfalls of EDL applied to AD and a more effective way to quantify uncertainty from EDL as compared with the foundational EDL setup. These studies illustrate a methodological approach to interpreting EDL in jet classification models, providing new insights on how EDL quantifies uncertainty and detects out-of-distribution data which may lead to improved EDL methods for DL models applied to classification tasks.

1. Introduction

Machine learning (ML) has become an indispensable tool in experimental high-energy physics (HEP), offering significant advancements in analyzing vast amounts of data obtained from complex detector systems. Over time, ML models have grown in complexity from simple regression and classification models into deep neural networks (DNNs) capable of performing sophisticated tasks to advance HEP. Despite the success of DNNs, they are often limited by their lack of explainability [1, 2] and ability to provide reliable uncertainties [3]. Uncertainty quantification (UQ) is crucial since uncertainties quantify the quality of predictive information and enable measurements to be contrasted or accurately combined. UQ also plays a crucial role in search for new physics (NP) signals, whether from specific NP phenomenological models or completely unexpected deviations from the standard model (SM) in the spirit of scientific exploration. The compatibility of extensions of the SM with data observations is constrained by the finite size of datasets as well as systematic uncertainties arising from detector performance and signal modeling.

Classification of jets, referred to as *jet tagging*, is a major application of ML and DL in the field of HEP. Jets are observed as conical sprays of hadronic showers originating from quarks and gluons produced in the high energy collisions at facilities like the Large Hadron Collider (LHC). Historically, the ATLAS and CMS collaborations using jet tagging algorithms in conjunction with classic statistical and ML models such as decision trees, played a pivotal role in jet tagging efforts (see [4–6] for instance in the context of top quark tagging). More recently, the advent of DNNs has ushered in a new era in jet classification algorithms for LHC physics. DNNs, with their ability to model complex, nonlinear relationships within data, have shown superior efficacy over traditional methods [7], particularly in scenarios with *boosted jets* where decay products of high-momentum heavy particles are highly collimated within a jet, requiring detailed analysis of jet substructures commonly employed in results using 13 TeV center-of-mass energy collisions at the LHC [8, 9].

A diverse range of deep learning (DL) models has been developed to optimize jet tagging. These approaches include DNNs [10], N-subjettiness taggers [11, 12], Deep Sets [13], CNNs [14, 15], and models that incorporate Lorentz symmetry [16–20]. There have also been networks that incorporate sequential models like recursive neural networks [21] and LSTMs [22], interaction networks [23], and Transformers [24]. There have been a variety of approaches to utilize the ability of DNNs to approximate arbitrary non-linear functions in high-dimensional data [25], and, as such, they have been successfully applied to the field of computer vision. Alternative models for jet tagging have been inspired by the underlying physics like jet clustering history [21], physical symmetries [16] and physics-inspired feature engineering [17]. These methods have inspired innovative model architectures and feature engineering by integrating or enhancing input feature spaces with physically meaningful quantities [17, 26, 27].

Despite the success of DL models for jet classification, UQ for these models remains a major challenge and an active area of research. Previous efforts have focused on uncertainties related to model deployment [28], nuisance parameters [29], theoretical modeling [30], and simulation systematics [31]. The black-box-like nature of DNNs obscures physical insight into the inner workings of these highly accurate classification machines, making it challenging to associate accurate and robust measures of uncertainty with these models. Traditional approaches to UQ in the context of DL models often utilize Bayesian inference models [32, 33], deep ensemble methods [34], and generative models like variational autoencoders [35].

A comprehensive review of these traditional approaches can be found in [36]. Many of these approaches pose significant challenges in terms of training complexity, convergence, and intuitive understanding of the associated uncertainty estimations. Additionally, some of these approaches are tied to specific models and cannot be easily adapted to other architectures. Recent advances in *explainable* artificial intelligence (XAI) [37] have made it possible to build intelligible relationships between an AI model's inputs, architecture, and predictions [1, 38, 39]. Additionally, UQ in association with ML models relies on developing robust explanations [3, 40] which are important for HEP algorithms such as jet tagging that require robust and interpretable models [41, 42] for high-quality physics results.

Expanding upon our previous work on interpretability of DL-based top quark taggers [42], we study *evidential deep learning* (EDL) for UQ [43] to develop a model-agnostic, robust, and interpretable approach towards UQ in jet tagging. EDL represents a novel and largely unexplored approach to UQ in HEP (But *cf* [44]), offering a method to evaluate the confidence of predictions made by DNN models. By treating the learning process as evidence acquisition and interpreting more evidence as increased predictive confidence, EDL provides a framework for models to express not just predictions but also the certainty of those predictions. It has a significantly lower computational cost than other DNN-based UQ methods like Ensemble or Bayesian networks. Allowing fast UQ, EDL opens up the possibility for applying UQ beyond the standard application of jet tagging in physics analyses. To translate the success of DNNs in jet and event classification into a fast and online jet tagger, recent work has placed emphasis on developing DNN-enabled FPGAs for trigger-level applications at the LHC [45–47]. As resource consumption and latency of FPGAs directly depend on the size of the network to be implemented, it is easier to embed simpler and faster networks on these devices. Hence, methods that quantify interpretable uncertainties without compromising performance can greatly benefit ML applications in both offline and real-time applications, especially for online event selection and jet tagging at current and future high energy colliders.

To demonstrate the application of EDL for UQ in jet tagging, we explore its integration with the Particle Flow Interaction Network (PFIN) model introduced in [42]. The PFIN model, originally developed to leverage the intricate details of particle flows for improved jet classification, is enhanced through the adoption of EDL to refine its predictive accuracy and provide new capability with regard to UQ. This adaptation represents a significant step towards rendering DNNs more interpretable and reliable for scientific research, particularly in fields where the precise understanding and handling of data uncertainty is required for data-driven discovery.

In this paper, we compare the uncertainties estimated by EDL with those from Ensemble and Bayesian methods and analyze the uncertainty distributions. The EDL structure and our chosen respective loss function is reviewed in section 2. To compare our results for existing benchmarks and different models, we use three publicly available datasets with varying number of jet classes to understand how the uncertainty shifts with different classes. The datasets were developed by the authors of [48–50] and are summarized in section 3. The EDL model hyperparameters, comparative Bayesian methods, dataset features, and their respective preprocessing are reviewed in section 3. The EDL-based uncertainties we analyzed for UQ are presented in section 4. We compare EDL uncertainties with those from Ensemble and Bayesian methods in section 5. We analyze and interpret the EDL-based uncertainty in section 6. In section 7, we explore the utilization of EDL for out-of-distribution (OOD) detection toward improved anomaly detection (AD) methods. We detail our outlook on EDL and the limitations of this method in section 8. Finally, section 9 summarizes our findings and illustrates new dimensions to explore in the conjunction of UQ and HEP.

2. Review of EDL

In jet tagging, UQ is crucial due to the complex nature of particle interactions, and the need for accurate, robust and interpretable jet classification. There are two main types of uncertainty: *aleatoric* and *epistemic*. Aleatoric uncertainty describes the noise in the training data, and epistemic uncertainty relates to insufficient training data [51]. Aleatoric uncertainty is often irreducible and can be estimated through neural networks [52]. On the other hand, epistemic uncertainty reduces with more data and is more difficult to approximate.

EDL introduces a novel approach to quantifying epistemic uncertainty, further referred to as uncertainty, in jet tagging. Unlike Ensemble and Bayesian methods, which rely on multiple inferences to approximate uncertainty, EDL directly models the uncertainty through sampling from a learned higher-order distribution. Grounded in the Dempster–Shafer theory of evidence [53] and implemented through Subjective Logic [54], EDL uses a Dirichlet distribution over class probabilities to interpret neural network outputs as subjective opinions, quantifying both confidence and uncertainty in predictions [43]. This approach reduces computational demands by eliminating the need of multiple network evaluations and offers a more detailed understanding of uncertainty, enabling networks to express a spectrum of potential outcomes and their respective confidence levels. This property of EDL is particularly advantageous in fields like particle physics that rely heavily on uncertainty estimation and statistical methods for interpreting large, complex data. In this paper, we present the first detailed study of EDL being applied to experimental HEP.

The foundational EDL approach [43] evaluates the epistemic uncertainty, or uncertainty mass, in classification tasks involving K exclusive class labels. Each class label has a corresponding belief mass b_k, $k = 1,\ldots,K$, and there is an overall uncertainty mass u. All of them are non-negative and sum to 1 as shown in equation (1):

$$\sum_{k=1}^{K} b_k + u = 1, \quad 0 \leqslant b_k \leqslant 1,\ 0 \leqslant u \leq 1,\ k = 1,\ldots,K. \tag{1}$$

The belief mass b_k of each class k is derived from a new concept, the evidence e_k. Evidence quantifies support gathered from data that advocates for categorizing a sample into a specific class. The relationship between b_k and e_k is shown in equation (2):

$$b_k = \frac{e_k}{S}, \quad S = \sum_{k=1}^{K} (e_k + 1)\,. \tag{2}$$

The uncertainty mass is then computed as shown in equation (3):

$$u = \frac{K}{S}. \tag{3}$$

The sum $S = \sum_{k=1}^{K}(e_k + 1)$ represents the Dirichlet strength, indicating the overall evidence strength supporting the classification. This is because the Dirichlet distribution, with parameters $\alpha_k = e_k + 1$, represents these belief mass assignments $b_1, b_2, \ldots, b_K$, which is also called a *subjective opinion*. The probability density function of the Dirichlet distribution with K parameters $[\alpha_1,\ldots,\alpha_K]$ is given by

$$D(\mathbf{x}|\alpha) = D(x_1, x_2, \ldots, x_K|\alpha_1, \alpha_2, \ldots, \alpha_K) = \frac{\prod_{i=1}^{K} x_i^{\alpha_i - 1}}{B(\alpha)},\ x_i \geqslant 0, \quad \sum_{i=1}^{K} x_i = 1 \tag{4}$$

where the normalizing constant $B(\alpha)$ can be defined in terms of the Gamma function $\Gamma(\cdot)$

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}. \tag{5}$$

For a given subjective opinion, the expected probability of the kth class is derived as the average value from the respective Dirichlet distribution, as shown in equation (6):

$$p_k = \mathbb{E}(x_k) = \frac{\alpha_k}{S} = \frac{e_k + 1}{S}. \tag{6}$$

The final stage in the EDL framework involves determining the evidence e_k. This can be accomplished by slightly modifying the outputs of traditional classification neural networks. Typically, classification neural networks utilize a SOFTMAX layer for output, which assigns probabilities to each class. In the EDL approach, the SOFTMAX layer is replaced with a ReLU activation layer. This ensures that the outputs are non-negative, which is necessary since these outputs are used as the evidence vector for the Dirichlet distribution that models the uncertainties and confidences in predictions. Let Θ denote the set of learnable parameters in the model. The outputs of the network, denoted as $\mathbf{f}(\mathbf{x}|\Theta)$, directly provide the evidence for the anticipated Dirichlet distribution through

$$e_k = f_k(\mathbf{x}|\Theta) \quad \text{and} \quad \alpha_k = f_k(\mathbf{x}|\Theta) + 1.$$

These modifications enable the network to not only predict outcomes but also provide a probabilistic assessment of these predictions, enriching the decision-making process in critical applications such as jet tagging.

To ensure the model learns these opinions, the optimal loss function for the EDL model is composed of two primary components, the reconstruction loss, $\mathcal{L}_{\text{MSE}}$, and the Kullback–Leibler (KL) Divergence, $\mathcal{L}_{\text{KL}}$. The reconstruction loss $\mathcal{L}_{\text{MSE}}$, is calculated as the mean squared error (MSE) between the predicted classification probabilities $\hat{\mathbf{y}}_i$ and actual targets $\mathbf{y}_i$ where $\hat{\mathbf{y}}_i = \frac{\alpha_i}{S} = \frac{f_i(\mathbf{x}|\Theta)+1}{\sum_{j=1}^{K}(f_j(\mathbf{x}|\Theta)+1)}$. Contrary to the traditional cross-entropy (CE) loss in a classification setting, using the MSE loss metric allows for simultaneous reduction of the prediction error and the variance of the Dirichlet distribution [43],

$$\mathcal{L}_{\text{MSE}}(\Theta)_i = \sum_{k=1}^{K} \mathbb{E}\left[(y_{ik} - \hat{y}_{ik})^2\right] = \sum_{k=1}^{K} \mathbb{E}\left[\left(y_{ik} - \frac{f_i(x_k|\Theta) + 1}{\sum_{j=1}^{K}\left(f_j(x_k|\Theta) + 1\right)}\right)^2\right]. \tag{7}$$

The second component of the loss function is a KL Divergence term defined as,

$$\begin{aligned}\mathcal{L}_{\text{KL}}(\Theta)_i &= \text{KL}\left[D(\hat{\mathbf{y}}_i|\tilde{\alpha}_i) \,\|D(\hat{\mathbf{y}}_i|\langle 1, \cdots, 1\rangle)\right] \\ &= \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\tilde{\alpha}_{ik}\right)}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha}_{ik})}\right) + \sum_{k=1}^{K}(\tilde{\alpha}_{ik} - 1)\left[\psi(\tilde{\alpha}_{ik}) - \psi\left(\sum_{j=1}^{K}\tilde{\alpha}_{ij}\right)\right]\end{aligned} \tag{8}$$

where

$$\tilde{\alpha}_i = \mathbf{y_i} + (1 - \mathbf{y_i}) \odot \alpha_i \quad \text{and} \quad \alpha_i = f_i(\mathbf{x}|\Theta) + 1$$

and $\psi(\cdot)$ is the digamma function.

As a key component in EDL to ensure that the model appropriately handles both in-distribution (ID) and OOD input data, equation (8) encourages the network to be more confident about correct predictions while allowing it to generously admit when it fails to do so. For OOD and hard-to-classify inputs, it ensures that the model outputs high uncertainty, effectively preventing overconfident and potentially erroneous predictions. For ID inputs, it encourages the model to exhibit a clear preference for one class over others by promoting one high evidence value e_k among the possible classes. This helps in sharpening the model's confidence in its predictions when faced with familiar data. This KL Divergence term is strategically integrated into the overall loss function as a regularization term, modulated by an annealing coefficient λ_t. The overall loss function is given by

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \mathcal{L}_{\text{MSE}}(\Theta)_i + \lambda_t \sum_{i=1}^{N} \mathcal{L}_{\text{KL}}(\Theta)_i. \tag{9}$$

The first term, $\mathcal{L}_{\text{MSE}}$, encourages the predicted class probabilities to match the true labels. The second term, $\mathcal{L}_{\text{KL}}$, encourages the model to remain uncertain when the prediction is not confident. In doing so, it prevents the model from making overconfident predictions on OOD and hard-to-classify inputs. The λ_t parameter is a hyperparameter of the EDL model which regulates the network's ability to assign uncertainties to model predictions. The authors of [43] proposed a dynamically scaled choice of λ_t to ensure a gradual increase during the training process, defined as $\lambda_t = \min(1.0, t/10) \in [0, 1]$, where t represents the epoch index. However, since the default choice did not always provide the most optimal solution in the applications we studied, we further adjusted its strength by parameterizing it as $\lambda_t(\zeta) = \zeta \times \min(1.0, t/10)$ with $\zeta \in [0, 1]$. This scaling allows the influence of KL Divergence term to be limited initially, avoiding overly harsh penalties that could lead to model convergence towards a uniform distribution prematurely. The annealing strategy ensures that as training progresses and the model stabilizes, the regularization effect of the KL Divergence becomes more important, guiding the model towards more accurate UQ.

3. Dataset and experimental setup

3.1. Datasets

In this paper, we consider three different datasets for UQ and AD using EDL: (1) top tagging, (2) JetNet, and (3) JetClass. The data details and cross validation setup for each of the datasets are summarized below:

(1) **Top tagging dataset** (`TopData`) [48, 55]: This dataset consists of 1 million top (signal) jets and 1 million QCD (background) jets generated with Pythia8 [56] with its default tune at 14 TeV center of mass energy for proton–proton collisions. The detector simulation was performed with Delphes [57] and jets were reconstructed using the $anti-k_t$ algorithm [58] with a jet radius of $R = 0.8$ using FastJet [59]. Only jets with transverse momenta within the range of 550 and 650 GeV are considered. For each jet, the dataset contains the four momenta of up to 200 constituents with zero-padded entries for missing constituents. The top tagging models are trained with transverse momentum (p_{T}), azimuthal angle (ϕ), and pseudorapidity (η) of the 60 most energetic particles. As part of data preprocessing, we standardized the constituents' η and ϕ by subtracting the jet's η and ϕ. The p_{T} values of the jets constituents are scaled by the inverse of the sum of constituents p_{T}, i.e. $1/\sum_i p_{\text{T},i}$. The dataset is divided into training, validation, and testing sets with a 6:2:2 split and trained in batches of 250. Some characteristic jet features from the dataset are shown in figure 1.

(2) **JetNet dataset** (`JetNet`) [49, 60]: This dataset consists of 880k particle jets originating from gluons (g), light quarks (q), top quarks (t), and bosons (W and Z). The parton-level events were generated using MadGraph5_aMC@NLO 2.3.1 [61] with its default tune at 13 TeV center of mass energy for proton–proton collisions. These parton-level events are then decayed and showered in Pythia8 [56]. Jets were reconstructed using the $anti-k_t$ algorithm [58] with a jet radius of $R = 0.8$ using the FastJet 3.13 and FastJet contrib packages [59, 62]. Only jets with transverse momenta within the window of 0.8 and 1.6 TeV are considered. For each jet, the dataset contains the four momenta of up to 30 constituents with zero-padded entries for missing constituents. Similar to the top tagging dataset, JetNet models are trained with p_{T}, ϕ, and η of jet constituents as input with the same preprocessing. The dataset is divided into training, validation, and testing sets with a 5:3:2 split and trained in batches of 250. Some characteristic jet features from the dataset are shown in figure 2.

(3) **JetClass dataset** (`JetClass`) [50, 63]: The dataset consists of 125 million particle jets of ten different types of jets initiated by gluons and quarks (q/g), top quarks (t), and bosons (W, Z, and H). As described in [64], jets initiated by a top quark or a Higgs boson are further categorized based on their different decay channels, resulting in the following ten categories: q/g, $t \to bqq'$, $t \to b\ell\nu$, $Z \to q\bar{q}$, $W \to qq'$, $H \to b\bar{b}$, $H \to c\bar{c}$, $H \to gg$, $H \to 4q$, and $H \to \ell\nu qq'$. The jets are extracted from simulated events that are generated with MadGraph5_aMC@NLO [61]. The parton showering and hadronization was performed withPythia8 [56] and the detector simulation was performed with Delphes [57]. Jets were reconstructed using the $anti-k_t$ algorithm [58] with a jet radius of $R = 0.8$ using the FastJet package [59]. Only jets with transverse momenta within the range of 550–1000 GeV and a pseudorapidity $|\eta^{\text{jet}}| < 2$ are considered. For each jet, the dataset contains 11 features for each particle, including information on kinematics, particle identification, and trajectory displacement. The particle features include the p_{T}, ϕ, and η of jet constituents, as well as the electric charge. Particle classification is represented using a five-class one-hot encoding to distinguish charged hadrons, neutral hadrons, electrons, muons, and photon. Additionally, the dataset includes measurements of the transverse and longitudinal impact parameters of particle trajectories, reported in mm. Each jet contains up to 60 constituents with zero-padded entries for missing constituents. The kinematic variables receive the same data preprocessing as in the other datasets. The dataset is divided into training, validation, and testing

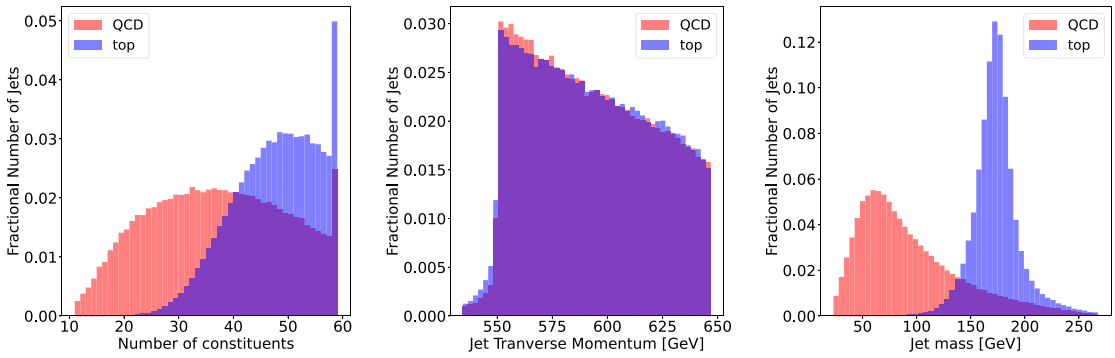

Figure 1. Distribution of (a) number of constituents, (b) jet transverse momentum ($p_{T,J}$), and (c) jet mass (m_J) for jets from QCD and top quarks.

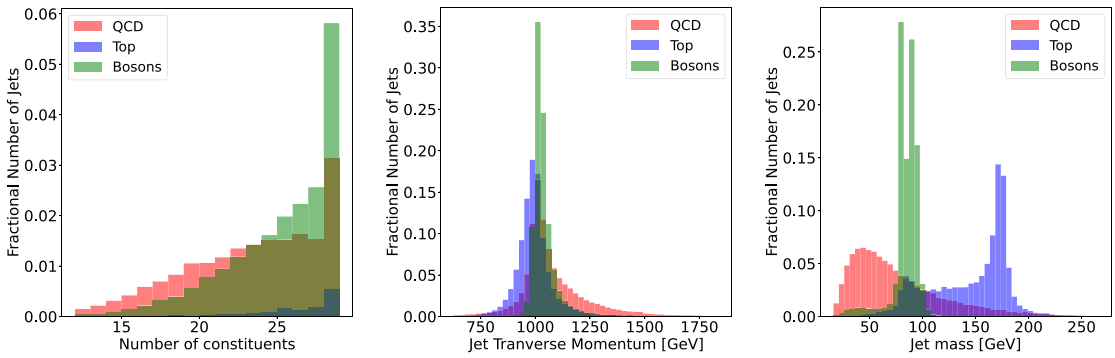

Figure 2. Distribution of (a) number of constituents, (b) jet transverse momentum ($p_{T,J}$), and (c) jet mass (m_J) for QCD (q,g), top (t), and boson (W,Z) jets.

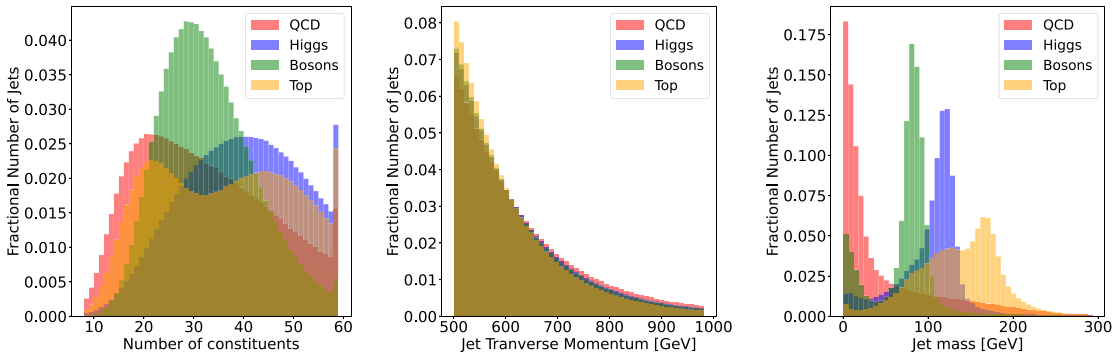

Figure 3. Distribution of (a) number of constituent particles, (b) jet transverse momentum ($p_{T,J}$), and (c) jet mass (m_J) for QCD (q/g), Higgs (H), boson (W,Z) and top (t) jets.

sets with a 100:5:20 split. In our work, we only use 20 M jets for training and 2 M jets for validation in batches of 2500 because there is an insubstantial increase in performance for larger training sizes. Some characteristic jet features from the dataset are shown in figure 3.

3.2. Model

The DNN tagger model we chose to integrate with the EDL model is the PFIN [42]. It is an augmentation of a Particle Flow Network (PFN) [13] with an interaction network (IN) [23, 65]. We chose this due to the

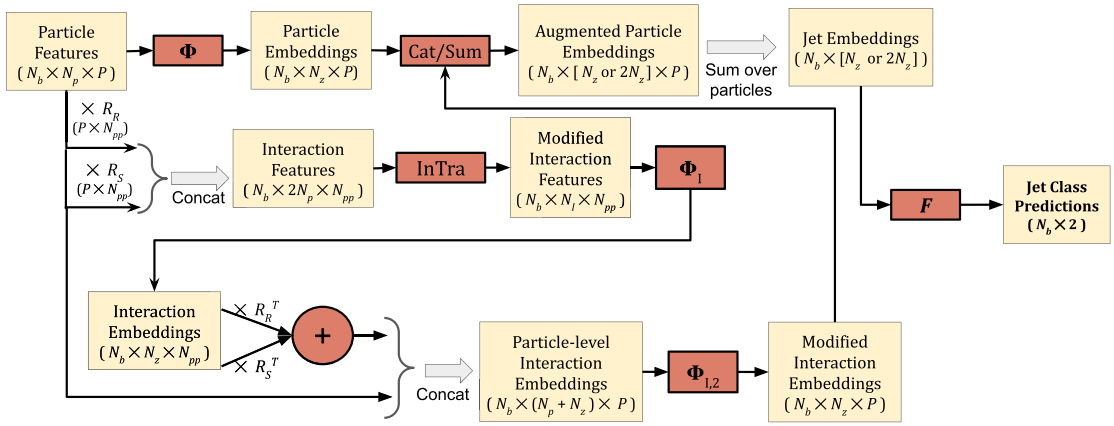

Figure 4. Model architecture and data flow for the PFIN model. N_b represents the batch size. The `InTra` block computes the pairwise particle interaction features given in equation (10). The `Cat/Sum` block creates the augmented particle embeddings by either concatenating or summing the outputs of $\Phi_{I,2}$ with Φ. The $\Phi, \Phi_I, \Phi_{I,2}$, and F networks are implemented as fully connected MLPs with ReLU activation. Reproduced from [42]. © The Author(s). Published by IOP Publishing Ltd. CC BY 4.0.

superior performance of the PFIN model on top tagging and its ability to learn from particle-level interactions in the latent space. These traits make it ideal for EDL to learn from particle-level features and investigate EDL's latent space representation.

As outlined in [42], the dataflow for the PFIN model is illustrated in figure 4. In PFIN, the particle interactions are encapsulated by formulating a fully connected undirected graph with $N_{pp} = \frac{P(P-1)}{2}$ edges where P represents the maximum number of constituent particles the model is trained with. Each particle within this graph is described by a set of N_p attributes. We have selected to use $N_p = 3$, using the triplet (p_t, η, ϕ) for each particle in `TopData` and `JetNet` datasets, following the same preprocessing steps. For the `JetClass` dataset, the number of attributes per particle was $N_p = 11$. For each edge in the graph, we combine the features of the two particles involved, resulting in an initial representation of $2N_P$ attributes for every edge. To assist in transforming these node-level features to edge-level attributes, we use two interaction matrices, R_R and R_S, each of which has dimensions $P \times N_{pp}$. The edge-level attributes are transformed by the Interaction Transformation (`InTra`) block to calculate a $N_I = 4$ dimensional representation for each edge by calculating the physics-inspired quantities $\ln \Delta$, $\ln k_T$, $\ln z$, and $\ln m^2$ [17, 24], where

$$\begin{aligned} \Delta &= \sqrt{(\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2} \\ k_T &= \min(p_{t,1}, p_{t,2})\,\Delta \\ z &= \frac{\min(p_{t,1}, p_{t,2})}{p_{t,1} + p_{t,2}} \\ m^2 &= (E_1 + E_2)^2 - ||\vec{p}_1 + \vec{p}_2||^2. \end{aligned} \tag{10}$$

The subscripts 1 and 2 denote the two particles associated with the edge and each variable within the relations refers to its unpreprocessed value. Since these quantities are symmetric with respect to the particles, the order of the particles does not impact PFIN's dataflow, maintaining the permutation-invariant property of PFN. These interaction features are transformed into N_z dimensional interaction embeddings by the trainable Φ_I network. These embeddings are propagated back to particle level using the interaction matrices, taking into account only those interactions where both particles are involved. These particle-level interaction embeddings are concatenated with the original particle features and further processed into N_z dimensional modified per-particle interaction embeddings through a trainable $\Phi_{I,2}$ network. The embeddings are then combined, either through concatenation or addition, with per-particle embedding from PFN's Φ network to obtain augmented particle embeddings. These augmented features are then summed over its constituents to obtain the jet-level latent representation. Finally, the F network obtains the output for each of the jet class based on these jet-level latent space features. At the end of the F network, a SOFTMAX layer is used for baseline models to output probabilities while a RELU layer is used for EDL models to output the Dirichlet parameters. The training for all models is done using the ADAM optimizer with minibatches. The model hyperparameters are chosen from the baseline PFIN summation model in [42].

3.3. Baseline methods

Traditionally, ensemble methods [34] and Monte Carlo (MC) Dropout [66] have been popular techniques for estimating uncertainty in DNNs. Ensemble methods involve training multiple models on the same task and using their varied outputs to evaluate uncertainty, providing a measure of confidence based on the diversity of the results. On the other hand, MC dropout leverages dropout layers during both training and inference phases to simulate the effect of Bayesian inference, thus providing a stochastic basis for uncertainty estimation [67]. Both methods are computationally intensive as they require multiple inferences to form a consensus on predictions, reflecting a significant trade-off between accuracy and computational efficiency. We use 10 independent estimates for each prediction in these methods. For the model ensemble, 10 instances of the same model are trained with different seeds to provide 10 independent models. For MC dropout, each sample is passed through the same model 10 times. Given that some of the datasets have more than two classes, minimizing the CE loss has been used as the cost function for all Ensemble and MC dropout models. The maximum of standard deviations of class-wise probability predictions has been used as an estimate of uncertainty for both ensemble and MC dropout methods.

3.4. Metrics

The models we trained for this analysis have been evaluated based on two underlying principles: (1) how confident a model is when it correctly predicts the class of a given jet and (2) how well the uncertainty estimate represents the ability of a model to identify misclassified or anomalous jets. Although this paper mostly focuses on the task of UQ, the metrics we propose in this section also allow us to assess the performance of AD models described in section 7. To simulate anomalous jets in AD models, we refer to two types of data: ID and OOD. ID jets refers to the type of data on which the model is trained, encompassing scenarios and characteristics that the model is expected to handle under normal operating conditions. Conversely, OOD data involves data points, or jet particle types, that are not represented during the training phase. Section 7 further explains the creation of ID and OOD datasets for the purposes of this study. The following metrics are critical for testing the model's robustness and its ability to handle unexpected or novel situations.

- **ID accuracy**: In our baseline models, ID accuracy is the same as model accuracy, measuring the ability of a model to correctly classify jets. In the case of AD models, this metric represents the accuracy of a model in correctly classifying ID jets, determined by the ratio of correct predictions on ID data to the total number of ID data. This metric ensures that a given model maintains high performance on familiar data and confirms that the enhancement in UQ or AD does not compromise its ability to handle expected scenarios.
- **Area under the receiver operating characteristic curve (AUROC)**: AUROC is commonly used metric to represent the overall quality of binary classification models. In our context, the AUROC represents how well the uncertainty estimate of a model correlates with an inability of the model to distinguish certain jet classes or identify anomalous jets. In a well-trained classification model, we want the model to be confident, i.e. assign low uncertainties, for correctly classified jets. On the other hand, large uncertainties should be associated with misclassified jets (in a UQ model) or anomalous jets (in an AD model). Figure 5(a) shows a typical ROC constructed from the results of a benchmark EDL model on the `JetNet` dataset. The vertical axis represents the fraction of misclassified jets that are assigned an uncertainty greater than a given threshold. The horizontal axis, on the other hand, represents the fraction of correctly classified jets that are assigned an uncertainty greater than a given threshold. The ROC is generated by varying the uncertainty threshold within the range of the observed uncertainties obtained by the model. A higher value of the AUROC would represent the model's superiority in projecting confidence for correctly classified jets while assigning larger uncertainties for incorrectly classified jets.

  A similar idea can also be constructed in the case of AD models. Figure 5(b) shows a typical ROC constructed from the results of a benchmark EDL model on the `JetNet-skiptop` dataset, a variant of the `JetNet` dataset that withholds the top jets from the training dataset but reintroduces them as OOD samples in the testing data. In this case, the vertical axis of the ROC represents the OOD detection rate, identified by the fraction of OOD jets assigned an uncertainty larger than the chosen threshold. The horizontal axis represents ID mis-tag rate, which is the fraction of ID jets assigned an uncertainty larger than the chosen threshold. Similar to what is done for UQ models, the ROC is generated by varying the uncertainty threshold within the range of the observed uncertainties obtained by the model. A larger value of the AUROC would imply a model's enhanced ability to tell apart OOD jets.

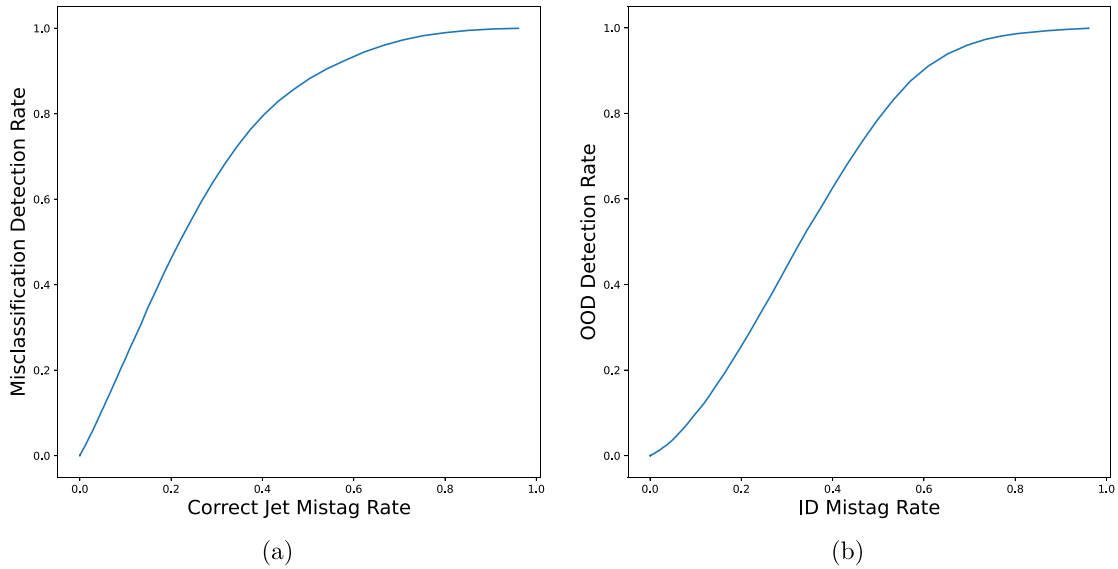

Figure 5. Receiver operative characteristic (ROC) curves for benchmark EDL models trained on the (a) baseline `JetNet` dataset for UQ and (b) `JetNet-skiptop` dataset for AD.

- **AUROC-STD:** For a Dirichlet distribution with K parameters $[\alpha_1, \ldots, \alpha_K]$, the Dirichlet standard deviation, D-STD (σ_k) for the kth class is given by

$$\sigma_k = \sqrt{\frac{\alpha_k (S - \alpha_k)}{S^2 (S+1)}} \tag{11}$$

where $S = \sum_{k=1}^{K} \alpha_k$ is Dirichlet strength defined in section 2. The quantity σ_k as introduced in equation (11) is a representative of uncertainty associated with the kth class prediction.

The AUROC, using D-STD as uncertainty (AUROC-STD), is similar to AUROC. However, it can only be used on EDL models because only they predict a Dirichlet distribution. We use this metric to compare with AUROC and determine if the D-STD or uncertainties from [43] are better estimates for UQ and AD. Since the D-STD predicts uncertainties per class, we use

$$u_{\text{D-STD}} = \sum_{k=1}^{K} \sigma_k \tag{12}$$

as a conservative estimate of total uncertainty associated with the classification. We chose to use linear summation of D-STD as uncertainties are correlated among various jet classes [68]. While quadrature summation was also studied for combining uncertainties, it yielded similar results. The AUROC-STD is then computed using the ROC curve constructed by varying the thresholds on $u_{\text{D-STD}}$. For EDL models, this metric is valuable for assessing the effectiveness of D-STD in representing uncertainty in comparison to equation (3).

4. Results on UQ

In this section, we examine how EDL models perform for UQ on jet classification tasks for the three datasets introduced in section 3. Ideally, uncertainties should be high for misclassified jets and low for correctly classified jets in a well-trained model. We examine multiple hyperparameter optimizations for the annealing coefficient $\lambda_t(\zeta)$ in equation (9), comparing fixed or gradually increasing approaches. We observe that gradually increasing λ_t, as proposed by the authors of [43], ensures faster convergence of accuracy. Additionally, we introduce a 'Confidence Tuned' variant of the EDL method (EDL-CT) in section 4.2. This variant initially converges without annealing ($\lambda_t = 0$), followed by parameter tuning through retraining with $\lambda_t > 0$. For EDL models, we also examine the use of the D-STD as uncertainty, referenced in equation (12).

4.1. Top tagging dataset

Since `TopData` only contains two classes, it is the simplest dataset to investigate the uncertainty generated from EDL. The performance of EDL model variants in the context of `TopData` is given in table 1. The model

Table 1. ID accuracy (Acc), AUROC (AUC), and AUROC-STD (STD) of the EDL and Ensemble methods on `TopData`, `JetNet`, and `JetClass` datasets. Within `TopData` and `JetNet` models, the Ensemble model has 970k parameters, while all other models have 97k parameters. The `JetClass` Ensemble model has 994k parameters, while all other `JetClass` models have 99k parameters. For each dataset, the entries marked in **bold** represent the EDL model with the highest central value for the corresponding metric. These measurements have uncertainties of $\mathcal{O}(0.001)$.

	`TopData`			`JetNet`			`JetClass`		
Model	Acc	AUC	STD	Acc	AUC	STD	Acc	AUC	STD
EDL $\lambda_t(0)$	**0.937**	0.723	0.894	**0.803**	0.550	0.792	**0.794**	0.602	0.816
EDL $\lambda_t(0.1)$	**0.937**	0.902	0.903	0.799	0.811	0.813	0.792	0.842	0.843
EDL $\lambda_t(0.5)$	0.936	0.902	0.902	0.796	0.815	0.816	—	—	—
EDL $\lambda_t(0.7)$	**0.937**	**0.904**	**0.904**	0.793	0.820	0.843	—	—	—
EDL $\lambda_t(1.0)$	**0.937**	**0.904**	**0.904**	0.790	0.822	0.823	0.776	**0.847**	**0.847**
EDL-CT $\lambda_t^{\mathrm{CT}}(0.1)$	—	—	—	0.801	0.814	0.815	—	—	—
EDL-CT $\lambda_t^{\mathrm{CT}}(0.5)$	—	—	—	0.788	0.831	0.832	—	—	—
EDL-CT $\lambda_t^{\mathrm{CT}}(0.7)$	—	—	—	0.776	**0.843**	**0.843**	—	—	—
Ensemble	0.937	0.890	—	0.806	0.772	—	0.805	0.782	—
MC dropout	0.933	0.887	—	0.797	0.743	—	0.793	0.745	—

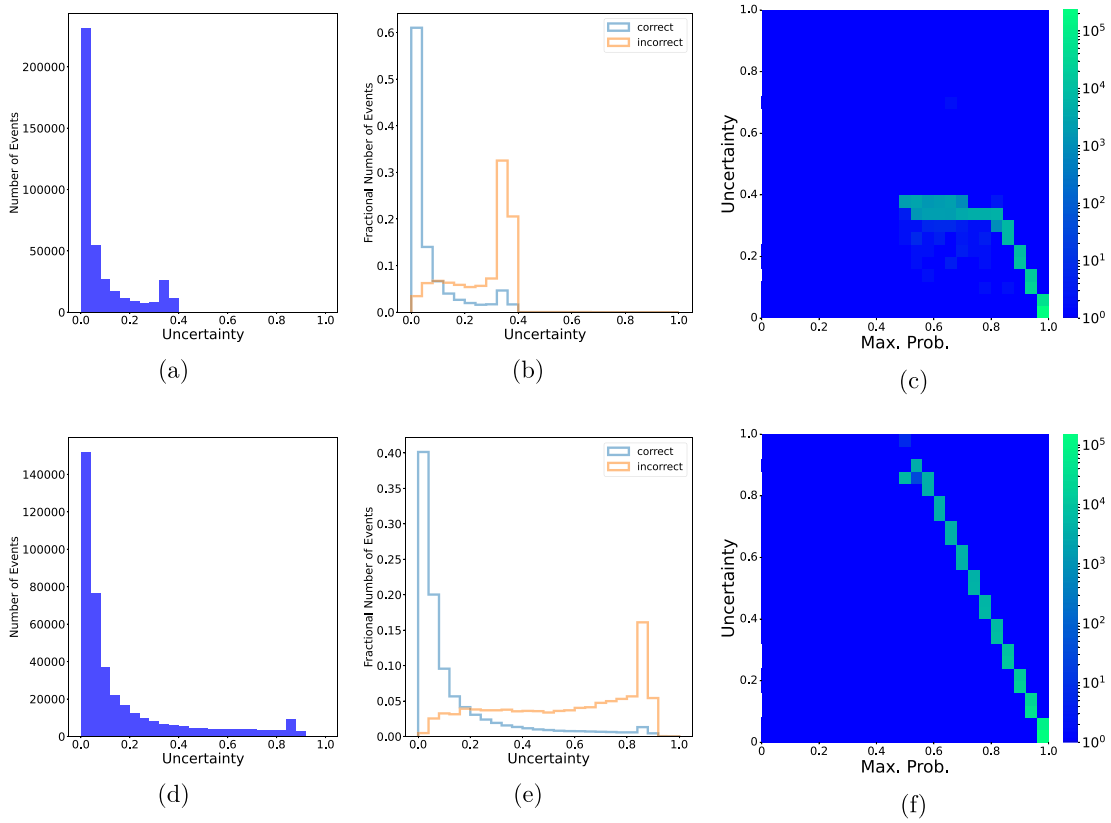

Figure 6. For the `TopData` dataset, on each row, from left to right, (a) and (d) uncertainty distribution, (b) and (e) uncertainty distributions separated by correct and incorrect jets, and (c) and (f) 2D histogram of maximum probability versus uncertainty, for baseline EDL $\lambda_t(0.1)$ (top row) and $\lambda_t(0.7)$ (bottom row).

accuracy is found to be very similar for different choices of EDL coefficients, depicting how the introduction of EDL for UQ does not interfere with the decision-making ability of the classifier model for this dataset. It is observed that higher ζ values result in a larger AUROC, signifying better discriminative ability between correct and incorrect predictions. The results are summarized in the `TopData` column of table 1. We find that EDL $\lambda_t(1.0)$ has the largest AUROC and EDL $\lambda_t(0.7)$ exhibits similar performance for the `TopData` dataset.

Figure 6 shows the impact of the choice of ζ as a hyperparameter for the choice of the model. As shown in figures 6(a) and (d), the distribution of the total uncertainty as obtained from these models shows a strong dependence on the choice of the regularization scale of the EDL model. Respectively choosing $\zeta = 0.1$ and $\zeta = 0.7$ for these two models, there are two distinct peaks in the uncertainty distribution. Figures 6(b) and (e) provide the uncertainty distributions separated for the correctly and incorrectly classified jets generated by the same models. In both instances, smaller uncertainties are attributed to correctly classified jets while the misclassified jets tend to be assigned larger uncertainties.

A general trend with EDL models is that as the ζ parameter increases in $\lambda_t(\zeta)$, there are more high-uncertainty jets, which is shown in figures 6(b) and (e). This corresponds to higher uncertainties in both correctly classified and misclassified jets. The cause can be attributed to the loss function. The ζ parameter is used to regulate the magnitude of the KL-divergence loss, which diverges away from a uniform Dirichlet distribution when misclassification takes place. When $\zeta = 0$, the Dirichlet parameters of the correct label keep increasing whenever the prediction is correct to minimize the loss resulting in an overly confident prediction. However, as ζ increases, the regularizing KL-divergence term takes more priority, penalizing the divergences from the 'I do not know' state. Then, the EDL model with larger ζ will have smaller Dirichlet parameters and high uncertainties as opposed to an EDL model with $\lambda_t(0)$. This can also be seen in the distribution of uncertainty as a function of the largest assigned probability (i.e. Max. Prob) in figures 6(c) and (f). As seen in figure 6(c), the uncertainty distribution hits a plateau close to the value of 0.4 as an artifact of the training with a weaker constraint on the KL-divergence term in the EDL loss function. On the other hand, EDL $\lambda_t(0.7)$ in figure 6(f) conforms with the general expectations from a well-trained uncertainty-aware classifier, that is (a) a general inverse relationship between Max. Prob and uncertainty and (b) a high concentration of correctly classified events in the low uncertainty bins. Since EDL $\lambda_t(0.7)$ has the

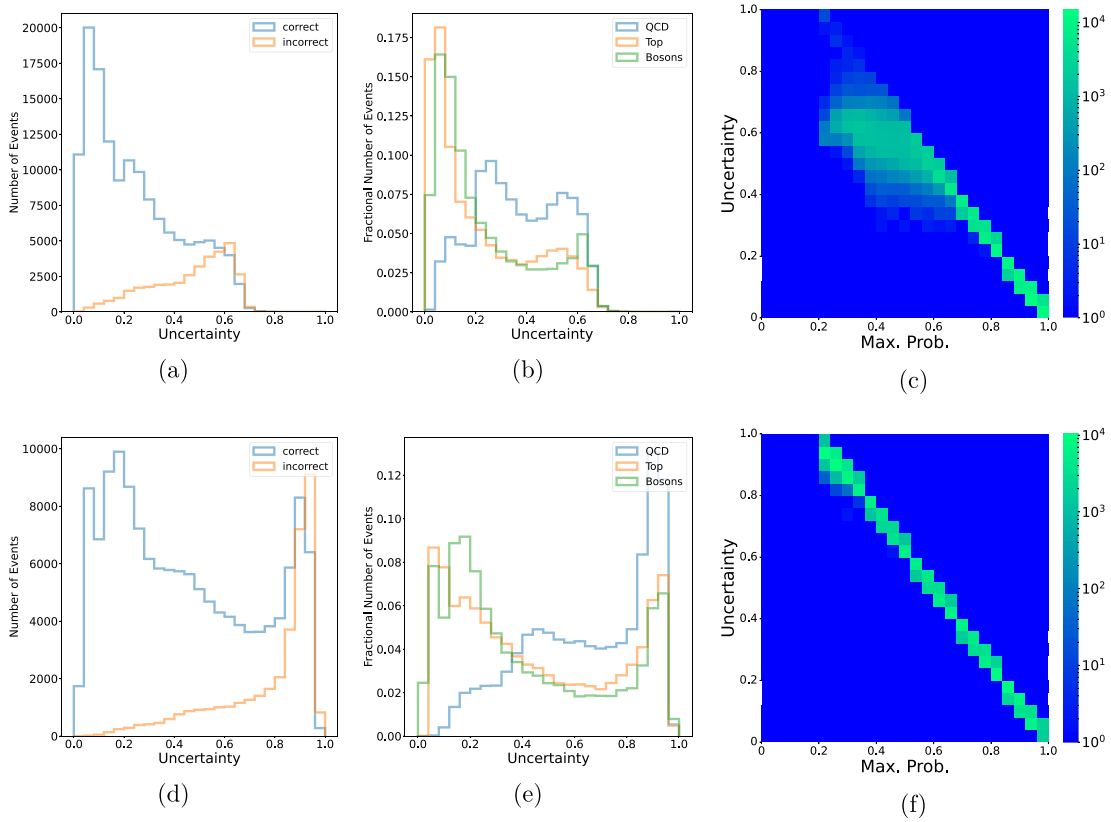

Figure 7. For the `JetNet` dataset, on each row, from left to right, (a) and (d) uncertainty distribution, separated by correctly and incorrectly classified jets, (b) and (e) uncertainty distribution for correctly classified jets, separated by initiating particle jet type, and (c) and (f) 2D histogram of maximum probability versus uncertainty for baseline EDL $\lambda_t(0.1)$ (top row) and $\lambda_t(0.7)$ (bottom row).

highest AUROC of any EDL model, this log-linear relationship indicates better misclassification prediction for this simple binary classification dataset.

Since the EDL model predicts the parameters of a Dirichlet distribution, we can also examine the D-STD as a measure of uncertainty in the top tagging dataset. As stated previously, the AUROC-STD is AUROC but with D-STD uncertainty. As shown in the `TopData` column of table 1, there is no significant difference between the AUROC and AUROC-STD scores for $\zeta > 0$.

4.2. JetNet dataset

In contrast to the binary classification of the `TopData` dataset, `JetNet` has five distinct classes of jets, giving a more comprehensive overview of how the EDL uncertainty behaves in a multiclass scenario. The `JetNet` dataset contains the following jets with their corresponding class labels: quarks (0), gluons (1), top quarks (2), W bosons (3), and Z bosons (4). As shown in the `JetNet` column in table 1, EDL models with higher ζ tend to have marginally lower accuracy but higher AUROC and AUROC-STD. This implies that EDL models with higher ζ make more incorrect predictions but tend to assign commensurately larger uncertainties to them.

To understand why ID accuracy decreases and AUROC increases as ζ increases, we examine the uncertainties of baseline `JetNet` EDL $\lambda_t(0.1)$ and $\lambda_t(0.7)$ in figures 7(a) and (d), respectively. Similarly to our observations for the EDL models applied to the `TopData` dataset, the range of uncertainties and proportion of high uncertainty jets for `JetNet` EDL models grows as the ζ increases. The uncertainties for EDL $\lambda_t(0.1)$ still have a bimodal distribution associated with correctly classified jets at low uncertainties and misclassified jets at high uncertainties. But for EDL $\lambda_t(0.7)$, there are a large number of correctly classified jets with higher uncertainties.

We visualize the uncertainties for each label and prediction through the Uncertainty Aware Confusion Matrix (UACM), as displayed in figure 8. The UACM is an extension of the traditional confusion matrix that incorporates uncertainty information for each prediction. The y-axis represents a binned distribution of predicted label plus uncertainty, which has a maximum of one, so it can display the general uncertainty distributions for correctly classified and misclassified jets. The x-axis represents the true labels for each jet.

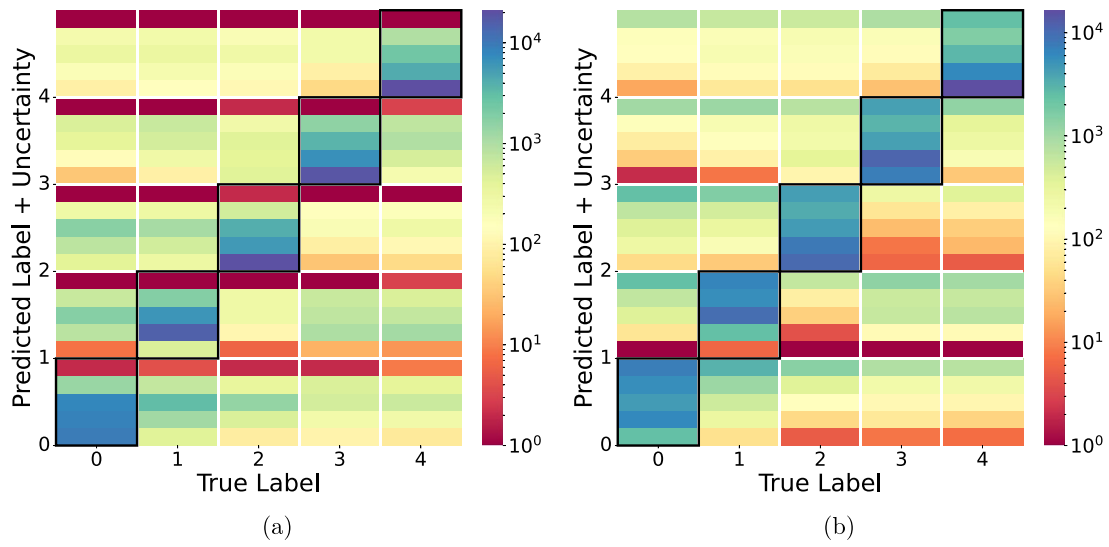

Figure 8. Uncertainty Aware Confusion Matrix for, respectively, baseline `JetNet` EDL (a) $\lambda_t(0.1)$ and (b) $\lambda_t(0.7)$.

Each cell is subdivided into five horizontal bins, corresponding to evenly spaced uncertainty bins in the range $[0, 1]$. These subdivisions allows us to determine the class-specific patterns for the distribution of uncertainty. The color intensity in each cell reflects the logarithm of the count of jets in that bin. For both choices of ζ, correctly classified quark and gluon jets with respective labels of 0 and 1 tend to have higher uncertainties.

As depicted in figures 7(b) and (e), the high-uncertainty, correctly-classified jets are dominated by QCD jets, while the heavier jets usually have lower uncertainties. This gives us an interesting insight into how the EDL models behave when two or more classes within the training dataset have similar physical characteristics. It is well known that jets initiated by quarks (q) and gluons (g) have very similar characteristics (being from the fragmentation of particles with color charge) and are generally regarded as *hard-to-tell-apart* (HTA) [69]. In fact, many LHC physics analyses either combine them together as a single jet class of light or QCD jets or employ sophisticated taggers developed specifically for q/g separation [70, 71]. This challenge of telling apart quark and gluon jets from their observed characteristics is large uncertainties assigned to these jets for higher values of ζ even when the model learns to correctly classify them. By increasing ζ, the model penalizes divergences from the 'I do not know' state. In both models, as shown in figures 7(c) and (f), the relationship between uncertainty and maximum probability is similar as found in case of EDL models applied the `TopData` dataset. However, unlike what we observed for the `TopData` dataset, the performance of the model does not necessarily improve with larger ζ. Models with large ζ show high uncertainty association for incorrectly classified jets at the expense of reduced confidence in correctly classified jets.

In an effort to circumnavigate this issue of large penalties for HTA jets, we introduce an alternative (hybrid) training paradigm. We refer EDL models trained with this paradigm as EDL-CT models. For the first 30 epochs, this model was trained with $\lambda_t = 0$ and the EDL regularization is restored with a nonzero constant after 30 epochs:

$$\lambda_t^{\text{CT}}(\zeta) = \begin{cases} 0, & \text{for first 30 epochs of training} \\ \zeta, & \text{for the remaining epochs} \end{cases} . \tag{13}$$

As shown in table 1, the EDL-CT model with $\lambda_t^{\text{CT}}(0.1)$ has an accuracy comparable with the EDL model with $\lambda_t(0)$ while its AUROC is much higher than the EDL model with $\lambda_t(0.1)$. This is an encouraging result, since it shows that this EDL-CT model can retain its classification performance while large uncertainty assignments correlate with misclassification more strongly. The method of confidence tuning results in smoother uncertainty distributions for correctly classified jets, as seen in figures 9(a) and (d). Both choices for EDL-CT models in figure 10 tend to show a softer uncertainty assignment for the q/g jets while most misclassified jets have larger uncertainties as compared with the models in figure 8. Similar to what we have observed before, a larger choice of ζ makes the model more conservative: its uncertainties are better at distinguishing misclassified jets at the expense of model accuracy.

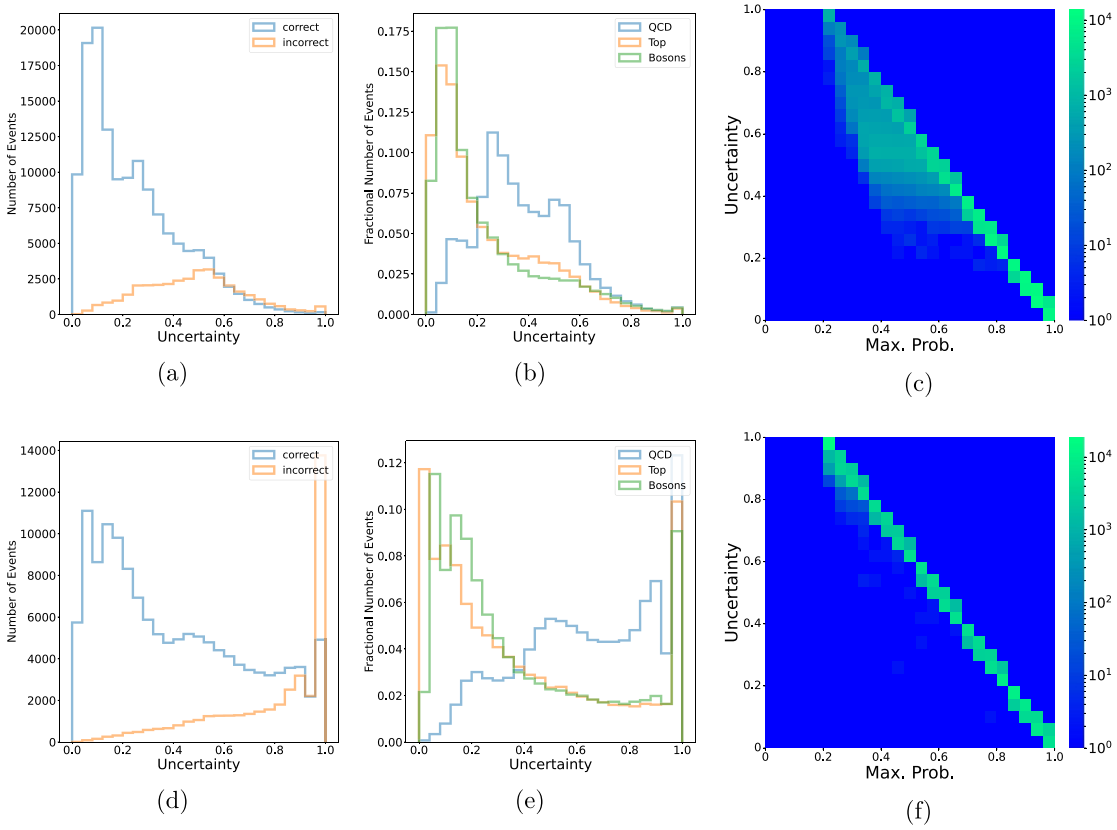

Figure 9. For the `JetNet` dataset, on each row, from left to right, (a) and (d) uncertainty distribution, separated by correctly and incorrectly classified jets, (b) and (e) uncertainty distribution for correctly classified jets, separated by initiating particle jet type, and (c) and (f) 2D histogram of maximum probability versus uncertainty for baseline EDL-CT $\lambda_t^{\text{CT}}(0.1)$ (top row) and $\lambda_t^{\text{CT}}(0.7)$ (bottom row).

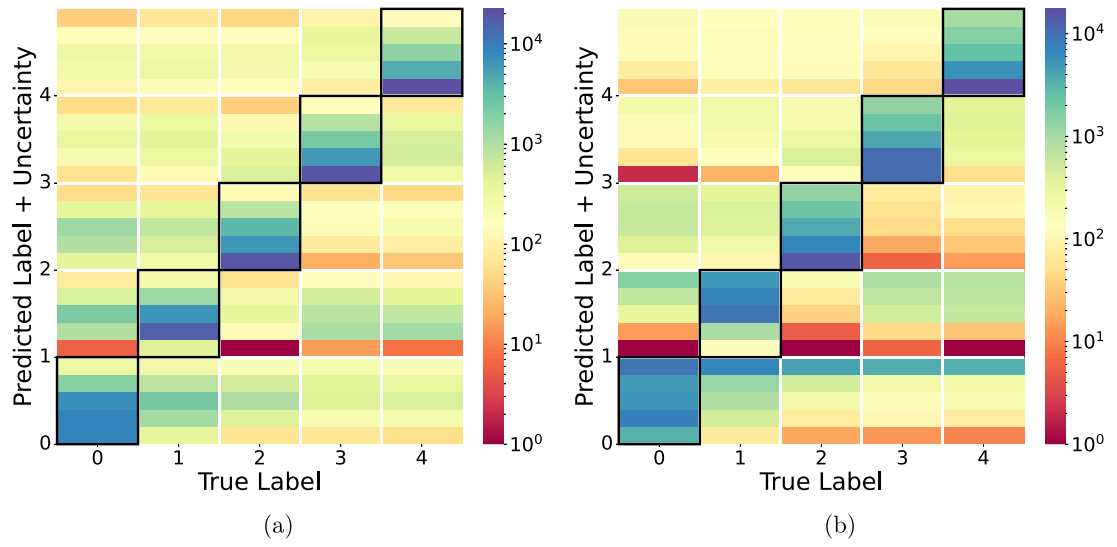

Figure 10. Uncertainty Aware Confusion Matrix for, respectively, baseline `JetNet` EDL-CT (a) $\lambda_t^{\text{CT}}(0.1)$ and (b) $\lambda_t^{\text{CT}}(0.7)$.

4.3. JetClass dataset

This dataset is much larger than the `TopData` and `JetNet` datasets and further subdivides jet classes by particle structure, allowing us to fully explore the extent of EDL-based UQ on jet tagging. The classes of the dataset and their indices are: q/g (0), $H \to b\bar{b}$ (1), $H \to c\bar{c}$ (2), $H \to gg$ (3), $H \to 4q$ (4), $H \to \ell\nu qq'$ (5), $Z \to q\bar{q}$ (6), $W \to qq'$ (7), $t \to bqq'$ (8), $t \to b\ell\nu$ (9). Building upon our experience with the smaller datasets,

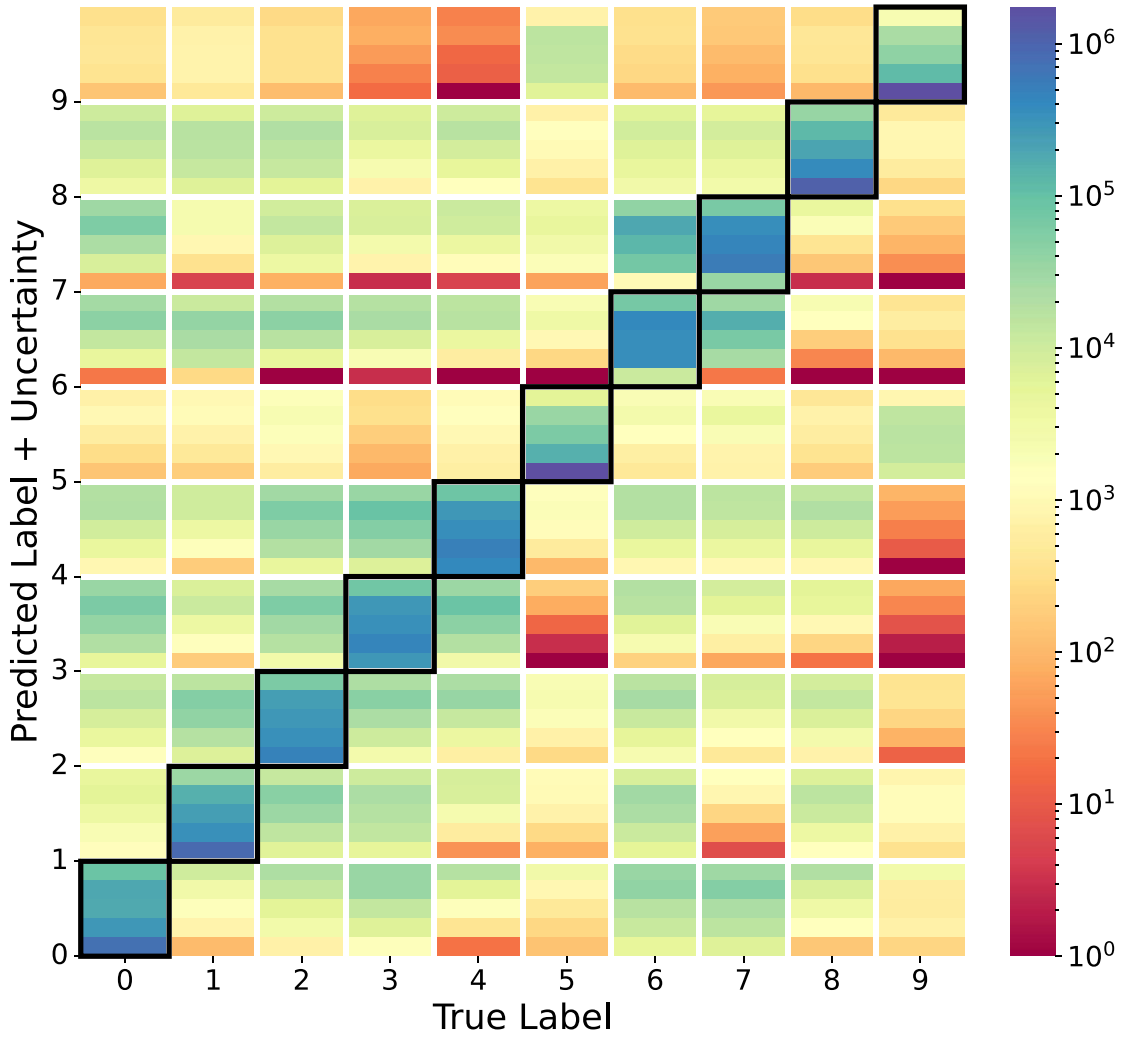

Figure 11. Uncertainty Aware Confusion Matrix for baseline `JetClass` EDL $\lambda_t(0.1)$.

we only trained the `JetClass` EDL models with two different choices of non-zero annealing coefficients: $\lambda_t(0.1)$ and $\lambda_t(1.0)$, to illustrate the impact of smaller and larger values of ζ[5].

The `JetClass` column in table 1 shows an evaluation of the predictive performance of the `JetClass` EDL models we studied. Similar to the EDL models applied to the `JetNet` dataset, as ζ increases, the classification accuracy decreases and AUROC increases, representing a tradeoff between predictive performance and conservative UQ. The EDL model with $\zeta = 0.1$ has a marginally smaller accuracy but the AUROC improves significantly when compared with the $\zeta = 0$ model. The uncertainty distributions across different classes are shown in the UACM in figure 11. The uncertainty distributions show the desirable characteristics with large uncertainties being attributed to misclassified jets while the correctly classified jets typically have softer uncertainties. This is also evident from the uncertainty distributions given in figure 12(a).

Figure 12(b) provides a detailed overview of uncertainties associated with different jet classes, where jet classes are combined according to the originating particle. While most classes show a smoothly declining uncertainty profile, the bosons class, comprising the $Z \to q\bar{q}$ and $W \to qq'$ classes, show a bimodal distribution with the second peak close to the mode of the uncertainty distribution of the incorrectly classified jets. This is also seen in the UACM in figure 11. These two classes were also found to be most likely misclassified as one another, which can be attributed to the similarity in their invariant masses and final states. Two correctly classified jet categories have low uncertainties: $H \to \ell\nu qq'$ (4) and $t \to b\ell\nu$ (9). These are also the only jets that contain leptons, suggesting that the model has confidently learned to exploit final state characteristics such as the particle-type information and decay topology.

[5] Training the PFIN model for the JetClass dataset showed a somewhat enhanced sensitivity to model initialization, sometimes requiring multiple iterations to reach a converging state.

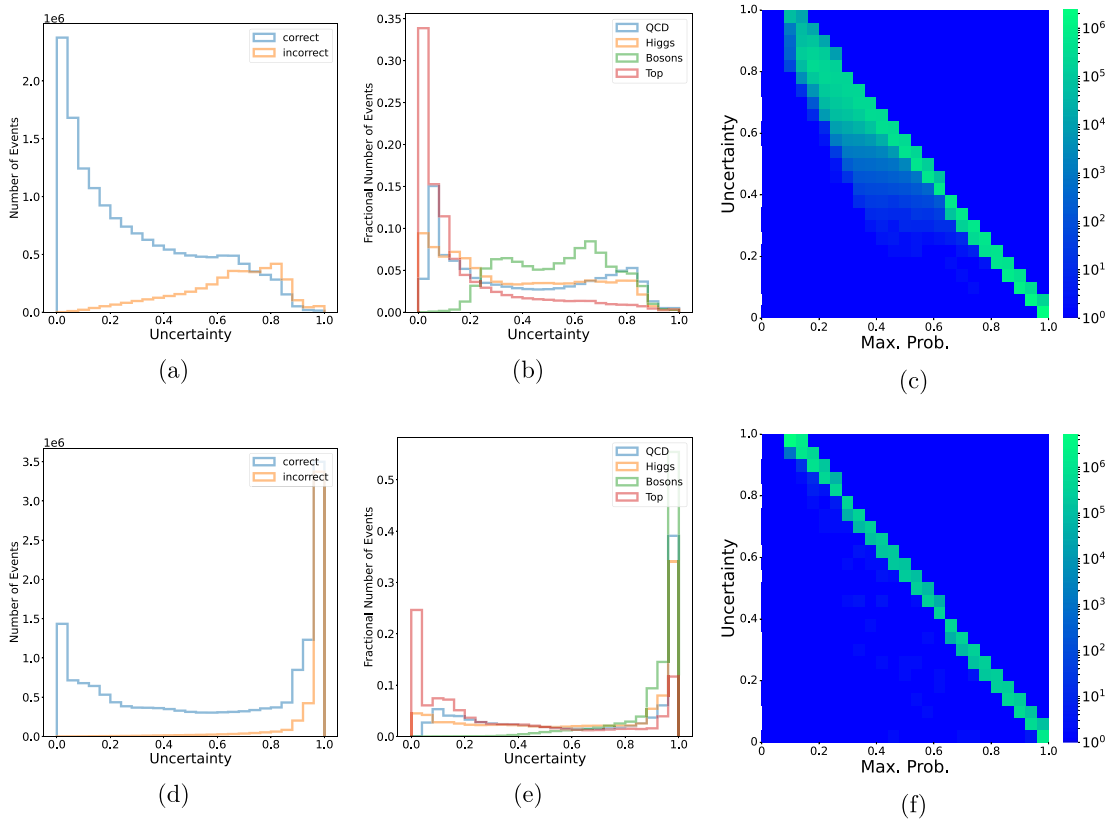

Figure 12. For the `JetClass` dataset, on each row, from left to right, (a) and (d) uncertainty distribution, separated by correctly and incorrectly classified jets, (b) and (e) uncertainty distribution for correctly classified jets, separated by initiating particle jet type, and (c) and (f) logarithmic 2D histogram of maximum probability versus uncertainty for baseline EDL $\lambda_t(0.1)$ (top row) and $\lambda_t(1.0)$ (bottom row).

As ζ increases, we observe a significant difference in the uncertainty profile of the different jet classes. Both correctly and incorrectly classified jets tend to show very large uncertainties (figure 12(d)) and all jet classes show strong bimodal distributions with a large peak near $u = 1.0$ (figure 12(e)). The uncertainty distributions can be further investigated from the UACM in figure 13. With increasing ζ, the model leverages the larger contribution of the DL divergence term in the loss function to assign high uncertainties to most of the jets. This again relays the importance of considering the accuracy alongside AUROC to determine the performance of an EDL model.

As we conclude this section, we note that even EDL $\lambda_t(1.0)$ successfully distinguishes the jet classes with leptonic decay modes with low uncertainties. We also observe confusion of the model to distinguish the $W \to qq'$ and $Z \to q\bar{q}$ classes, misclassifying one into the other while assigning relatively large uncertainties on these class determinations.

5. Comparison with ensemble methods for UQ

To determine the efficacy of EDL, we compare this method with two different Bayesian methods: Ensemble training and MC dropout. We analyze the uncertainties for Ensemble and MC dropout, as detailed in appendix A. Both Bayesian methods took ten times longer than EDL models on inference passes to estimate the uncertainty due to the nature of these methods. As such, EDL models are preferable for systems with limited computational resources. However, the choice of model is subject to optimization and depends on the dataset and training set.

When benchmarking against the best-performing EDL model chosen from the optimal combination of accuracy and AUROC scores, we observe that the Ensemble methods typically provide better or comparable accuracies. On the other hand, MC dropout performs similarly or worse than EDL in terms of accuracy. Both methods show worse performance in terms of AUROC. This indicates that a well-trained EDL model can provide similar performance in terms of accuracy but does a better job at assigning larger uncertainties to misclassified jets.

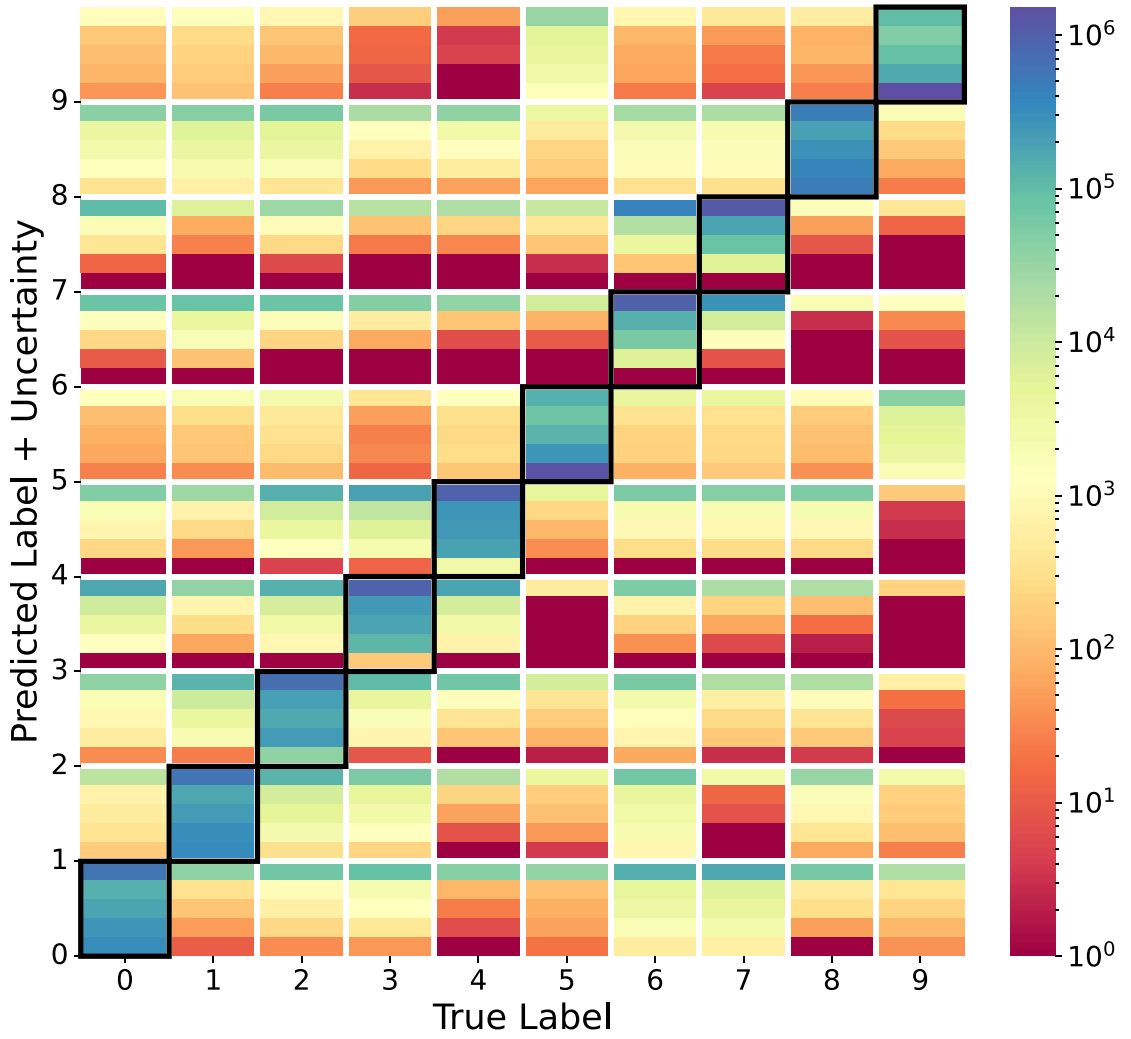

Figure 13. Uncertainty Aware Confusion Matrix for baseline `JetClass` EDL $\lambda_t(1.0)$.

The results that compare EDL models with Bayesian methods for `TopData`, `JetNet`, and `JetClass` datasets are summarized in table 1. For `TopData`, both the EDL and Ensemble models achieve the same ID accuracy, but MC dropout has slightly worse prediction performance. The EDL $\lambda_t(0.7)$ and $\lambda_t(1.0)$ models outperform both Ensemble models on AUROC, suggesting that EDL is a better UQ method for the top tagging dataset.

The performance of the baseline EDL models on the `JetNet` dataset exhibits a different trend. All EDL models with non-zero ζ perform better than the Bayesian methods in UQ at the expense of classification accuracy, although the degradation is modest. The best performing EDL model is the EDL-CT model with $\zeta = 0.1$ which has a slightly worse accuracy but a big improvement in AUROC compared to the Ensemble method.

The performance of the baseline `JetClass` models is similar to that of the `JetNet` models. We note that both accuracy and AUROC improve in the Ensemble model for `JetClass` when compared with the benchmark of $\lambda_t(0)$. The best performing EDL model is the model with $\lambda_t(0.1)$ which has a slightly lower classification accuracy but a significantly larger AUROC.

6. Interpretation of EDL uncertainty estimation

Since the EDL model is a deterministic DNN that directly predicts a Dirichlet distribution, the model must also encode some information on the evidence gained for each class in the latent space. To understand how the model learns the uncertainty, we examine the distribution of variances in the latent space representation using Principal Component Analysis [72]. As shown in our previous work in [42], PCA reveals how the model reorganizes useful correlations with highly discriminative features. We perform similar studies on the PFIN latent space for all three datasets studied in this paper. We use the best performing model for each

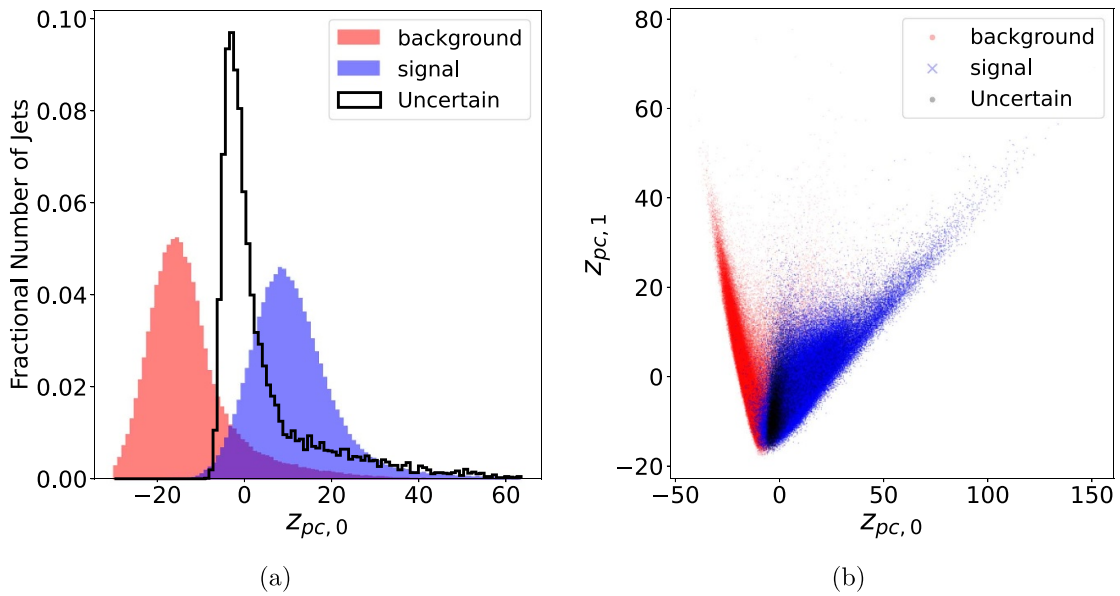

Figure 14. For `TopData` EDL $\lambda_t(1.0)$, (a) distribution of first principal component for background QCD jets and signal top jets along with incorrectly predicted jets with uncertainties greater than 0.8, also known as 'Uncertain' jets and (b) pairwise distribution between first and second principal component for background, signal, and 'Uncertain' jets.

dataset, namely EDL $\lambda_t(1.0)$ for `TopData`, EDL-CT with $\lambda_t^{\text{CT}}(0.1)$ for `JetNet`, and EDL $\lambda_t(0.1)$ for the `JetClass` dataset.

For the `TopData` dataset, we found that 99% of the observed variance in the test data was described by the top 37 principal components. Along with this, we set an uncertainty threshold at 0.8 and examine the distributions of the first principal component of the misclassified jets with an uncertainty higher than this threshold. We identify high-uncertainty misclassified jets as *uncertain* jets. Figure 14(a) shows the distribution of the top principal component, $z_{\text{pc},0}$ for the two jet classes along with the uncertain jets for EDL $\lambda_t(1.0)$. We can readily see how the large-uncertainty misclassified jets lie right at the overlap region, where discrimination is the hardest. We can also examine how the correlation between these PCA-transformed latent features further displays large uncertainty at the intersection of the distributions in figure 14(b).

For the larger `JetNet` and `JetClass` datasets, we can group jet classes based on initiating particle types and analyze the latent space in figure 15. They show similar patterns in how high-uncertainty misclassified jets are near the intersection of principal components and class types. The distribution of the principal components for top quarks are much further other class, which is likely why they usually have lower uncertainty as shown in figures 7 and 12.

Having examined how the uncertainty maps onto the principal components of the latent space, it is also instructive to investigate if learning of uncertainty impacts the ability of a model to incorporate information about physical jet characteristics. As stated in the original PFIN paper, jet-class information is found to be embodied in the distribution of correlations among latent space features [42]. We repeated those studies in the context of our current experiments to find if the model still manages to embody jet class information in such correlations. We chose to examine jet features such as jet mass and the number of constituents which, as shown in figures 1–3, can have moderate-to-strong discriminative power and give estimates for uncertainty.

For the `TopData` dataset, the first principal component, $z_{\text{pc},0}$, shows a strong correlation with jet mass for both jet categories with correlation coefficients of 0.9 and 0.8 for background and signal jets, respectively. Similarly, in the EDL models applied to the `JetNet` dataset, the correlation coefficient between $z_{\text{pc},0}$ and QCD jets is 0.9 (with a similar level of correlation found for top jets) while for boson jets the correlation is weaker at 0.7. The first principal component shows comparable correlation with jet mass in the `JetClass` dataset with correlation coefficients being 0.9 for QCD jets and 0.6 for all other jet categories. Despite the larger size of the `JetClass` dataset as compared with `TopData` and `JetNet` datasets, EDL models do not diminish in their ability to construct expressive distributions in the latent space.

PFIN also allows us to explore the impact of pairwise particle interaction matrices on UQ and jet classification. As explained in [42], we calculated the ΔAUC and Mean Absolute Differential Relevance (MAD relevance) score for each pair of particles by masking the corresponding input to the Φ_l network and calculating the deviation in model prediction with respect to the baseline model result. Additionally, we calculate the deviation in the model prediction probabilities and uncertainty using the `TopData` dataset.

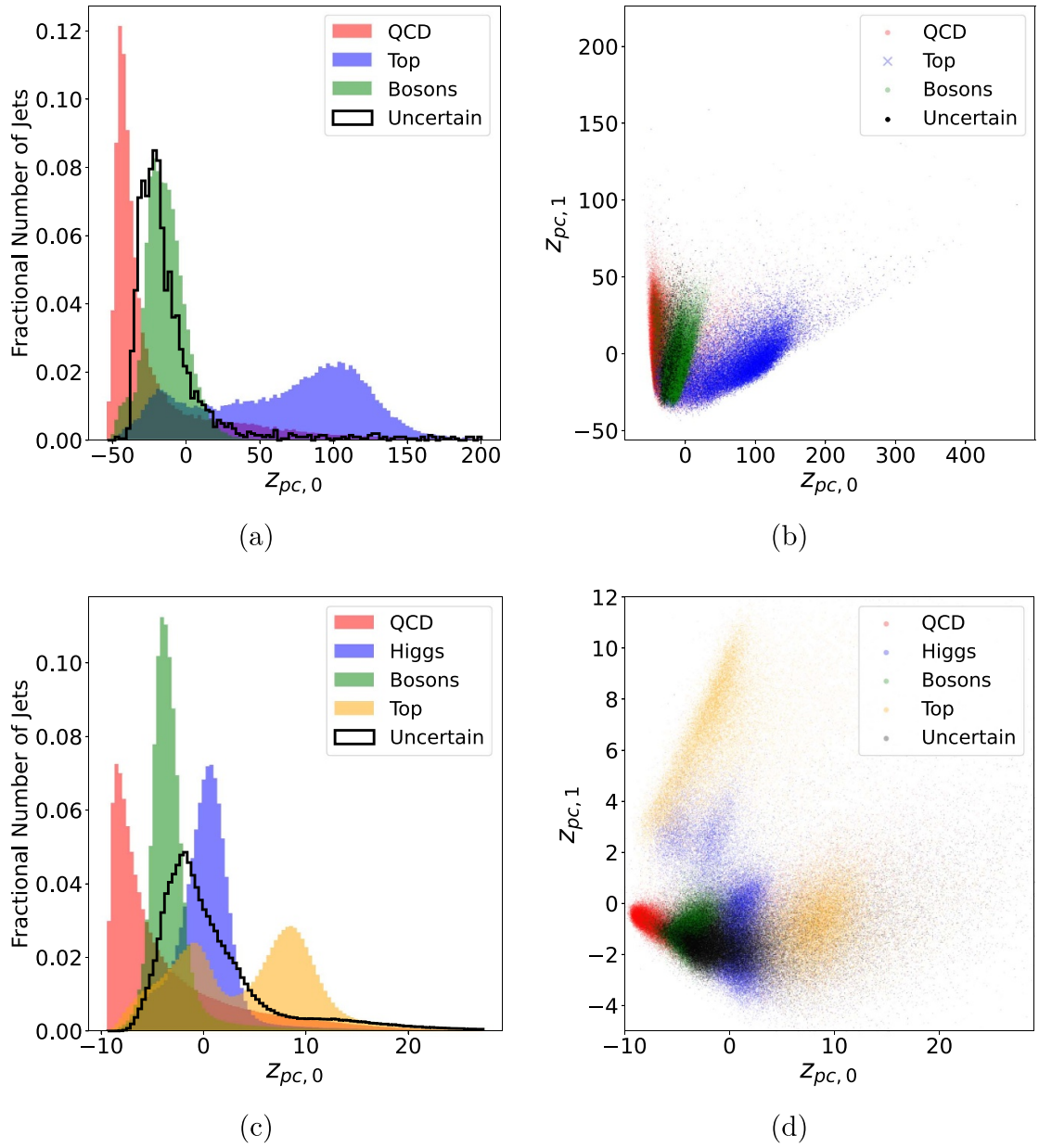

Figure 15. For (a) `JetNet` EDL-CT $\lambda_t^{\mathrm{CT}}(0.1)$ and (c) `JetClass` EDL $\lambda_t(0.1)$, distribution of first principal components for jets grouped based on initiating particle types along with incorrectly predicted jets with uncertainties greater than 0.8, which we refer to as 'Uncertain' jets and (b) and (b) pairwise distribution between first and second principal component for jets grouped on initiating particle types and 'Uncertain' jets.

These quantities are useful for evaluating the contribution of individual features by examining how the model performs when we mask a particle interaction. The results for the EDL baseline model with $\lambda_t(1.0)$ on the `TopData` dataset are shown in figure 16.

The pairwise particle interactions play a particularly important role in identifying the signal jets. The mean deviations in the background jet class probabilities are barely impacted by masking interaction features. However, for the signal jets, this impact is found to be rather large, with the mean prediction probability reduced by almost 20% when the interaction between the two most energetic jets is masked. In addition, the uncertainty slightly increases when masking interaction features, which is expected due to the removal of important information from the model.

7. EDL for AD

There is a compelling and potentially powerful connection between UQ for ML models and detection of data anomalies having characteristics not seen in model training, such as under/overdensities or OOD data. The foundational EDL paper [43] demonstrated this capability using rotated handwritten numbers from the

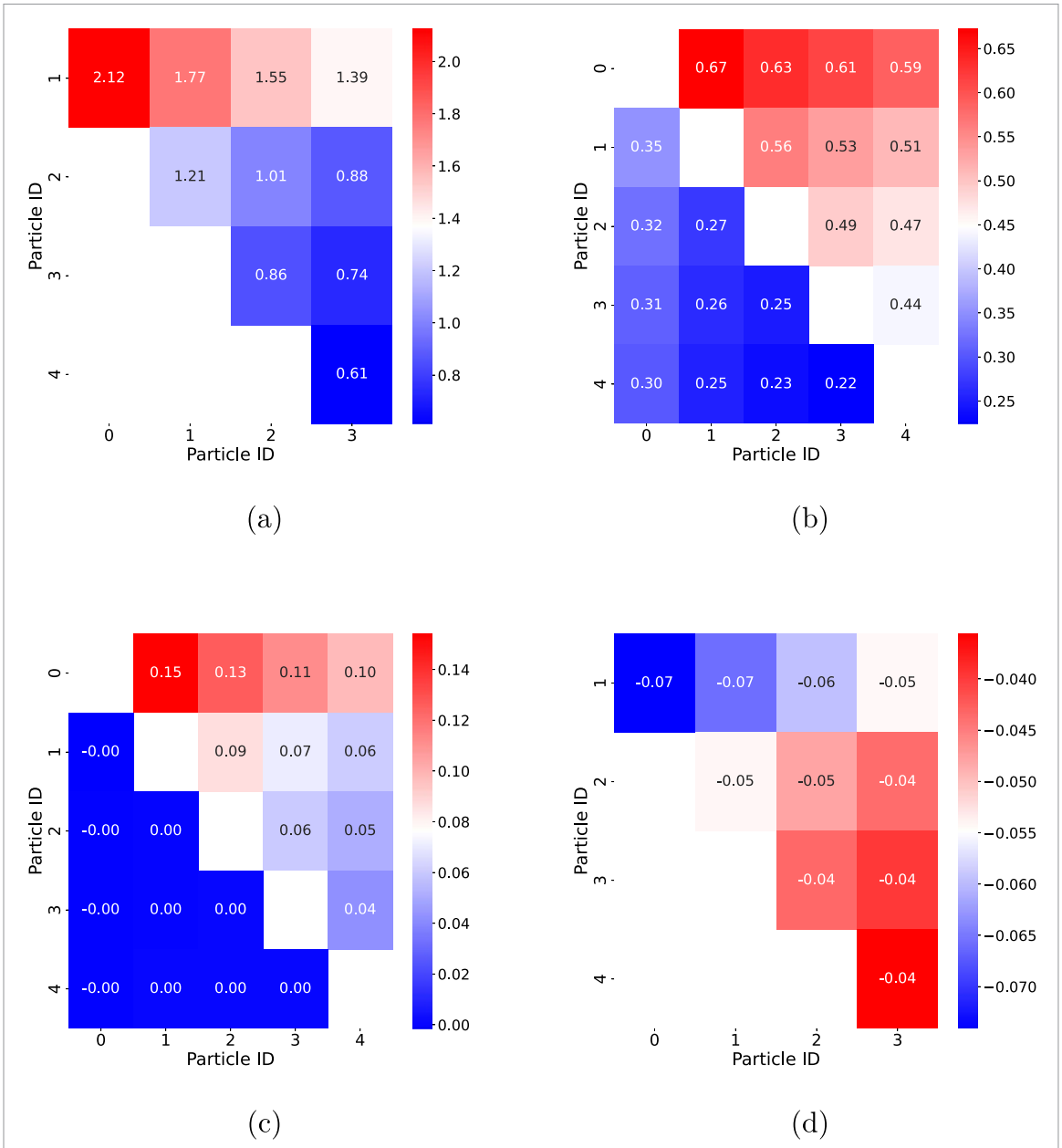

Figure 16. For the `TopData` baseline EDL $\lambda_t(1.0)$ model, (a) ΔAUC, (b)normalized MAD relevance score, (c) mean deviation in jet class prediction probabilities, and (d) mean deviation in uncertainties for pairwise masking of particle interaction features for the five most energetic constituents. Particle constituents are arranged in decreasing order of their energies. For (b)and (c), the upper and lower diagonal entries represent the corresponding scores for the signal and background jet classes, respectively.

Table 2. Experimental configurations within the `JetNet` dataset highlighting different training scenarios by excluding specific particle classes.

`JetNet` training configurations				
Names	baseline	`skiptop`	`skipwz`	`skiptwz`
In-distribution jets	g, q, t, W, Z	g, q, W, Z	g, q, t	g, q
Out-of-distribution jets		t	W, Z	t, W, Z

MNIST dataset [73]. EDL has been applied to AD in numerous settings, for example the detection of maritime anomalies due to unusual vessel maneuvering [74].

In this section, we examine how the EDL-based uncertainty behaves with OOD data. We analyze Ensemble and MC dropout-based uncertainty in appendix B. To create 'anomalies' or OOD jets, we omit certain classes from the training dataset and analyze the uncertainties for both ID and OOD jets from the test dataset. We examine three different AD models for the `JetNet` dataset shown in table 2: `skiptop`, `skipwz`, `skiptwz`. The models are evaluated based on the same metrics as the baseline models.

In JetNet-skiptop EDL networks, we skip jets from top quarks during training and analyze how well the uncertainties identify these 'anomalies' in the test set. The results are summarized in table 3. As shown in the skiptop column of table 3, the best performing EDL model is EDL $\lambda_t(0.5)$ with an AUROC peaking at 0.754. Increasing ζ further decreases both the ID accuracy and AUROC. As shown in figure 17(a), JetNet-skiptop EDL $\lambda_t(0.5)$ assigns high uncertainties to most OOD jets, which serves as an indicator of the predictive limitations of the model on OOD data. There are many high-uncertainty QCD jets from misclassifications, making it difficult to differentiate between the ID QCD jets and OOD top jets. This points to a fundamental challenge in using EDL for OOD jet detection. The EDL uncertainty, in its simplest form, fails to distinguish between the jets that are hard to tell-apart and the jets that are unknown from the training data. This limitation makes Ensemble models remain competitive and a relatively simpler alternative, despite their high computational cost.

We observe similar situations when trying to detect OOD jets using EDL for the skipwz and skiptwz datasets, as shown in figures 17(b) and (c) respectively. The skipwz dataset considers the W and Z boson categories as anomalies while in the skiptwz dataset, all jets coming from the heavy bosons and the top quark are considered OOD. In both cases, the uncertainty distribution of the QCD jets has a peak near the tail of the respective distribution, close to where the uncertainty assigned to most OOD jets is concentrated.

We note that choosing the best EDL model for the skipwz dataset was trickier than the other cases. As shown in the skipwz column of table 3, as ζ increases, both accuracy and AUROC decrease, with EDL $\lambda_t(0)$ having the highest AUROC. The EDL $\lambda_t(0)$ model performs better than even Ensemble and MC dropout. This is significantly different from the previous JetNet models where there was increased AUROC for non-zero ζ. However, the physical range of uncertainties associated with the $\zeta = 0$ model is very narrow and close to zero, so uncertainty attributions are rather sporadic and noisy. Hence, uncertainty estimates are better characterized in the model with $\lambda = 0.1$. Both categories of ID jets show a strong peak near $u = 0$, and a second peak close to the tail of the distribution. The peak near the larger uncertainties is much pronounced for QCD jets, which is a somewhat expected behavior based on our observations of the EDL model performance in the baseline case in section 4.2.

Since skiptwz models skip both top quarks and bosons during training, the model is now a binary classifier for quarks and gluons. In terms of ID accuracy, the skiptwz EDL models perform similarly to the binary top tagging models described in section 4.1, with the accuracy barely decreasing as ζ increases. However, the AUROC score improves with non-zero ζ, and peaks at $\zeta = 0.6$. In figure 17(c), there is a bimodal distribution for ID jets associated with low-uncertainty correctly-identified jets and high-uncertainty misclassified jets. The OOD jets have high uncertainties, but many of the ID jets are also assigned large uncertainties due to misclassification, which often occurs for HTA jets.

Overall, it is difficult to differentiate between HTA and OOD jets. In a way, this behavior is expected from EDL models. The EDL-assigned uncertainty to each jet instance reflects the level of confidence in the classification scores the model predicts. As a singular metric, we expect this quantity to be large whenever the model comes across a jet that is unlike anything it has seen before. We also expect this quantity to be large when this model encounters a jet with characteristics making it difficult to confidently place it in a single category. Misclassifications are likely to happen in such cases and though it is promising that OOD jets are associated with high uncertainties.

8. Outlook on model selection and limitations of the EDL method

As we have discussed in section 2 and demonstrated through our study of jet classification, the method of EDL provides a valuable method to faithfully assign epistemic uncertainties to deep classifiers. The uncertainty (equation (3)) associated with EDL can be interpreted as the lack of statistical evidence in favor of the classification score obtained from a parameterized Dirichlet distribution. As pointed out in [43, 75], the uncertainty mass is associated with the network's lack of knowledge in classifying certain events. A larger value of the uncertainty can be qualitatively interpreted as the network's decision in declaring either it is hard to tell or it is different. This interpretation is realized by implementing the evidential paradigm allowing the network to model a Dirichlet distribution (and not a categorical distribution) over an n-dimensional simplex. The loss function has been carefully designed to (a) align the expectation values of the n-dimensional probability vector with the truth labels of the training class (and hence reduce bias), (b) control the overall variance of the distribution, while (c) penalizing the network for accumulating evidence in favor of any misclassification. On the other hand, equation (11) is a more direct measure of uncertainties associated with the class predictions. Though both equations are ubiquitously termed as uncertainty in standard ML literature, their usage in the context of a physics analysis requires a proper examination of what these quantities represent. In the context of a jet classifier, the former would represent the quality of the classification, so a proper use-case of this uncertainty can be, for instance, using this quantity as a threshold

Table 3. ID accuracy (Acc), AUROC (AUC), and AUROC-STD (STD) of the EDL and Ensemble methods on variants of `JetNet`: `skiptop`, `skipwz`, `skiptwz`. The Ensemble model has 970k parameters, while all other models have 97k parameters. For each of the `JetNet` variants, the entries marked in **bold** represent the EDL model with the highest central value for the corresponding metric. These measurements have uncertainties of $\mathcal{O}(0.001)$.

	`skiptop`			`skipwz`			`skiptwz`		
Model	Acc	AUC	STD	Acc	AUC	STD	Acc	AUC	STD
EDL $\lambda_t(0)$	**0.815**	0.380	0.511	**0.818**	**0.824**	**0.753**	**0.837**	0.386	0.614
EDL $\lambda_t(0.1)$	0.814	0.701	0.713	0.816	0.701	0.697	0.836	0.713	**0.709**
EDL $\lambda_t(0.5)$	**0.815**	0.754	0.754	0.816	0.697	0.696	0.833	0.682	0.682
EDL $\lambda_t(0.6)$	0.813	**0.756**	**0.757**	0.814	0.690	0.690	**0.837**	**0.714**	0.687
EDL $\lambda_t(1.0)$	0.811	0.743	0.744	0.815	0.666	0.666	0.836	0.690	0.690
EDL-CT $\lambda_t^{\mathtt{CT}}(0.1)$	0.814	0.724	0.729	0.817	0.676	0.676	0.836	0.635	0.681
EDL-CT $\lambda_t^{\mathtt{CT}}(0.5)$	0.808	0.746	0.745	0.816	0.681	0.680	0.835	0.694	0.694
EDL-CT $\lambda_t^{\mathtt{CT}}(0.7)$	0.808	0.743	0.743	0.816	0.692	0.691	0.835	0.697	0.697
Ensemble	0.822	0.766	—	0.824	0.741	—	0.824	0.741	—
MC dropout	0.810	0.656	—	0.817	0.717	—	0.833	0.693	—

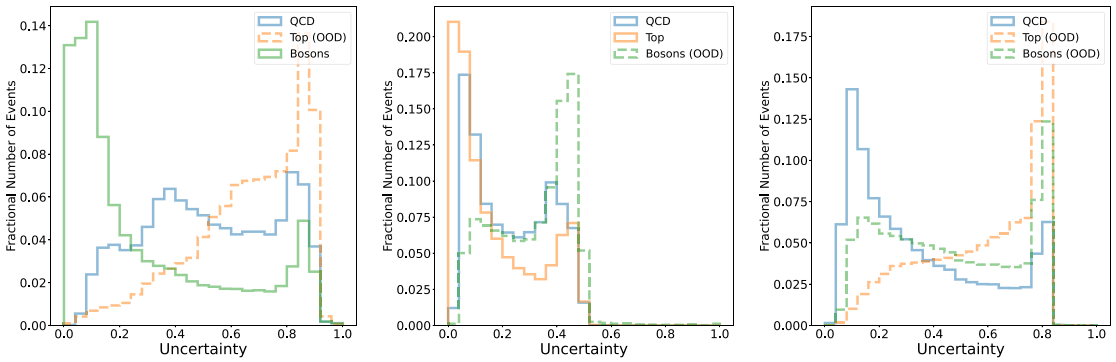

Figure 17. Uncertainty distributions, with jets separated by initiating particle jet type, for (a) `JetNet-skiptop` EDL $\lambda_t(0.5)$, (b) `JetNet-skipwz` EDL $\lambda_t(0.1)$, and (c) `JetNet-skiptwz` EDL $\lambda_t(0.6)$ models.

for jet selection. The latter would be a more appropriate quantity to be assigned to physical distributions associated with jets, and can be incorporated as a systematic uncertainty in the context of likelihood optimization for a search or a precision measurement analysis.

As our analyses have demonstrated, the performance of the EDL mechanism is subject to (1) the choice of the hyperparameter ζ and (2) the choice of training methodology as illustrated in the differences between the standard EDL and EDL-CT methods. It is an artifact of the nature of the EDL loss function. Hence, it is important to define a systematic procedure to make the right choice of ζ. The observations we made in section 4 suggest that model accuracy is typically the largest with $\zeta = 0$ with a small AUROC. A small increase in ζ yields in a significant increase in the AUROC with a marginal degradation in accuracy. Even with the EDL-CT models, smaller values of ζ tend to give better performance. These findings are also in line with the observations made in [76]. As a result, for the choice of right ζ, model accuracy should be benchmarked against the accuracy obtained with $\zeta = 0$ while the AUROC should show a significant improvement over the $\zeta = 0$ case. In most use cases, a small, non-zero value of ζ for either EDL or the EDL-CT method would be most appropriate choice for the model. Additionally, epistemic uncertainty for ID data should converge towards zero as the training data increases. However, the authors of [76], have theoretically and empirically confirmed that the learned epistemic uncertainty in EDL is constant for ID data with increasing number of observations. Consistent with prior findings, we independently observe that epistemic uncertainty remains constant with increasing `TopData` training set sizes.

Finally, we point out a potential limitation of the EDL method. The uncertainty that EDL assigns is an effectively discriminative estimate of model uncertainty for a given choice of model parameters. However, as argued by some authors (e.g. in [76]), this is a conditional but incomplete estimate of model uncertainty as it does not take into account uncertainties arising from variations in model parameters. In that sense, the EDL uncertainties might be complementary to the uncertainties obtained from the Ensemble method since the latter attempts to capture the systematic variations in the model's predictions arising from variations in the model parameters. In the context of experimental analyses, a conservative account of both types of epistemic uncertainty could be made by incorporating both Ensemble-based variances with EDL-estimated uncertainties as independent and uncorrelated systematic uncertainties. However, the applicability and effectiveness of such a strategy might be dependent on the nature of the analysis itself. A more complete account of evidential uncertainties might require employing an ensemble of EDL models. That study is beyond the scope of this paper and is left for future work.

9. Conclusions and outlook

This paper presents a comprehensive study of EDL in the context of UQ in jet tagging datasets. Our work has unveiled a number of important aspects regarding how the uncertainty and performance vary with the corresponding datasets. We have observed the EDL-based uncertainty and its comparable performance to Bayesian methods. The convergence and performance of the EDL method strongly depend on the choice of the annealing coefficient, ζ. Larger annealing coefficients result in lower accuracy, higher AUROC, wider ranges of uncertainties, and a larger number of high-uncertainty jets. Hence, model selection of a robust EDL-based classifier relies on the proper choice of ζ. Our empirical insights, as summarized in section 8, suggest that in most use cases a model with a small nonzero ζ value would give a desirable AUROC while maintaining an accuracy close to the benchmark of $\zeta = 0$.

As a method of UQ, EDL provides direct estimates of uncertainties on class-wise predictions expressed as standard deviations of a parametric Dirichlet distribution. Given the physical range of uncertainties associated with EDL-based UQ varies with each choice of the annealing hyperparameter, the uncertainties predicted by this model must be regarded as *post-hoc* uncertainties associated with a given instance of the model. In other words, EDL uncertainties express the confidence a model projects for a given choice of model parameters. These predictions should not be regarded as representative uncertainties distributed over a class of potential parameter and hyperparameter choices.

We also observe how the EDL-based uncertainty maps onto the latent space of the PFIN model. We demonstrate that high-uncertainty misclassified jets populate (based on the first principle component) at the intersection of jet distributions in latent space embeddings in all datasets. This bridges an important gap between our previous studies on model interpretability and the current work on UQ. It is evident from our studies of the latent space embeddings that a well-tuned EDL model can show strong uncertainty associations for misclassified and HTA jets. Finally, although the method of EDL shows promise leveraging UQ for the detection of OOD jets, AD using EDL can be limited in telling apart the OOD jets from the HTA ID jets. Any attempt to reliably detect OOD jets can definitely benefit from additional degrees of freedom to identify anomalous jets from ID jets.

This work establishes a methodology to evaluate and optimize application of EDL for UQ and AD, using jet classification at the LHC as an important case study. While the results presented in this work rely exclusively on the PFIN model, the EDL method by itself remains model-agnostic. Post-hoc EDL uncertainties are reliable and highly discriminative estimates of model uncertainty, but it requires some effort to obtain performance optimization through hyperparameter tuning and training strategies. This paper also lays out the primary optimization criteria for selecting the best model for a given use case. As EDL uncertainties can be obtained in a single pass on the data during inference stage with minimal additions and modifications to a neural network model for classification or regression, it opens up potential applications for uncertainty-aware algorithms and hardware co-design for edge and low-latency applications, such as fast data reduction, detector triggering, and AD. In regard to AD, there is potential to leverage EDL to improve the performance and model independence of traditional approaches such as autoencoders, which we leave to future work.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/FAIR4HEP/PFIN4UQAD [77].

Acknowledgments

The authors would like to thank the Center for Artificial Intelligence Innovation at the NCSA for support through our affiliation. This research is part of the Delta research computing project, which is supported by the National Science Foundation (Award OCI 2005572), and the State of Illinois. Delta is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, Grant 1725729, as well as the University of Illinois at Urbana-Champaign. This work was supported by the FAIR Data program of the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under Contract Number DE-SC0021258, the U.S. Department of Energy, Office of Science, High Energy Physics, under Contract Number DE-SC0023365, and the National Science Foundation Cooperative Agreement PHY-2117997. The authors also acknowledge the use of OpenAI's ChatGPT to improve clarity and readability.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authorship contribution statement

Ayush Khot: Methodology, Analysis, Software, Visualization, Validation, Writing—original draft & editing. **Xiwei Wang**: Methodology, Analysis, Software, Visualization, Validation. **Avik Roy**: Conceptualization, Methodology, Analysis, Software, Visualization, Validation, Writing—original draft, review & editing. **Volodymyr Kindratenko**: Conceptualization, Resources, Supervision, Writing—review & editing. **Mark S. Neubauer**: Conceptualization, Resources, Supervision, Writing—original draft, review & editing.

Appendix A. Uncertainties from ensemble methods for UQ

In this section, we provide the results of additional examinations of Ensemble and MC dropout methods for UQ on jet classification tasks for the three datasets introduced in section 3. In the context of `TopData` in table 1, EDL outperforms both approximated Bayesian methods. Ensemble also slightly outperforms MC dropout in the AUROC metric. To understand the performance, we plot the uncertainties generated from Ensemble and MC dropout methods in figure A1. The range of uncertainties produced by these ensemble methods is significantly narrower compared to those from EDL, as shown in figure 6. In particular, in figures A1(b) and (e), the difference between uncertainties of correctly and incorrectly identified jets is much smaller for the Ensemble and MC dropout methods than EDL, as shown in figure 6(e). As a result, EDL outperforms both ensemble methods by providing more discriminative uncertainties. A significant difference between the uncertainties of both ensemble methods and EDL is their relationship with Max. Prob. They all have a high concentration of correctly classified events with low uncertainties, aligning with the expected behavior of a well-trained uncertainty-aware classifier. However, figure A1(c) and (f) indicates a weaker correlation compared with EDL models. This suggests that the epistemic uncertainty is more associated with OOD detection rather than misclassification detection. Both `JetNet` and `JetClass` uncertainties from ensemble methods have similar such properties.

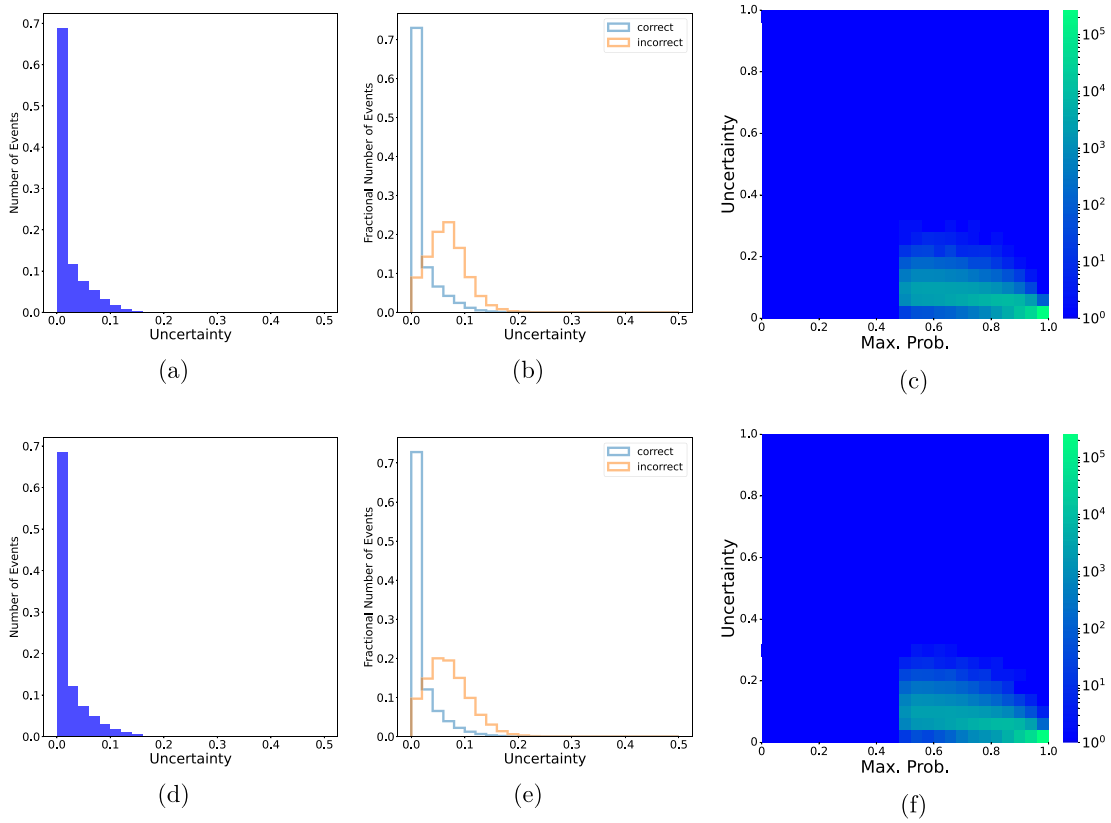

Figure A1. For the `TopData` dataset, on each row, from left to right, (a) and (d) uncertainty distribution, (b) and (c) uncertainty distributions separated by correct and incorrect jets, and (c) and (f) 2D histogram of maximum probability versus uncertainty, for baseline Ensemble (top row) and MC dropout (bottom row).

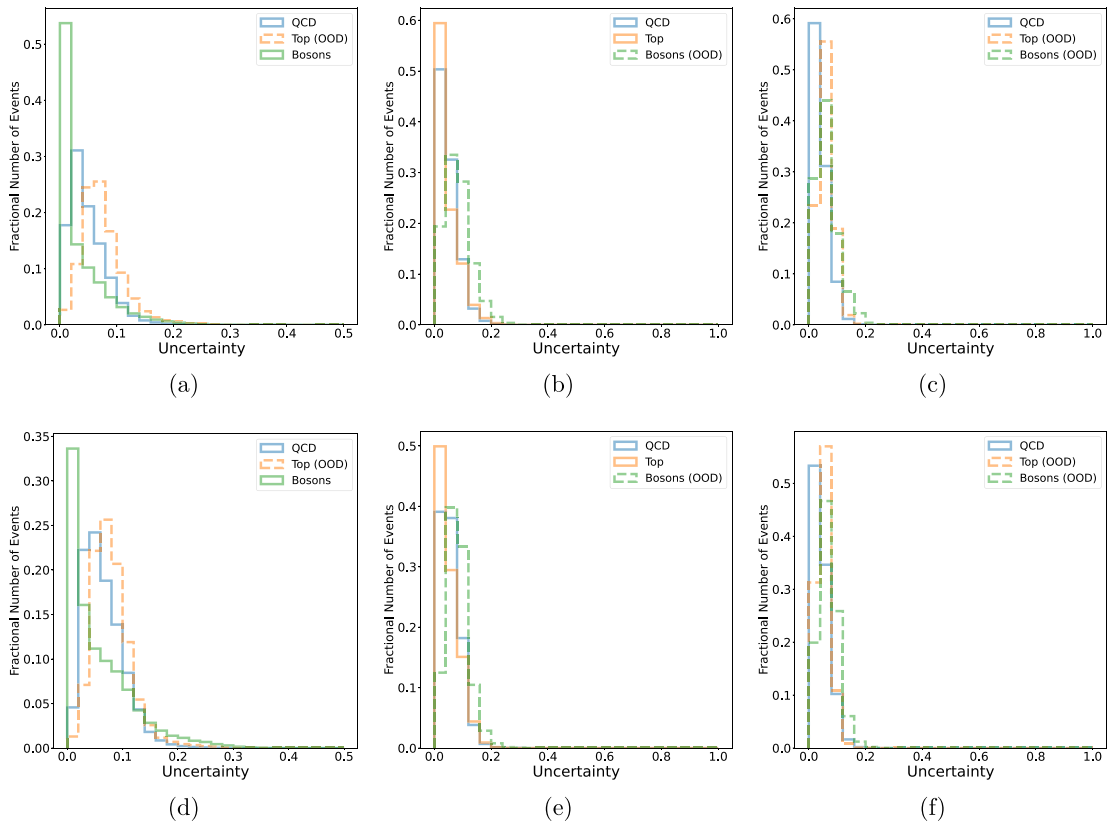

Figure B1. For the `JetNet` dataset, one each row, from left to right, uncertainty distributions, with jets separated by initiating particle jet type, for (a) and (d) `JetNet-skiptop`, (b) and (e) `JetNet-skipwz` and (c) and (f) `JetNet-skiptwz` for Ensemble (top row) and MC dropout (bottom row).

Appendix B. Uncertainties from ensemble methods for AD

In this section, we examine how ensemble-based uncertainties behave with OOD data. As previously shown in table 2, Ensemble outperforms EDL in `JetNet-skiptop` and `JetNet-skiptwz`, and MC dropout performs very poorly. We plot the uncertainties from Ensemble and Bayesian methods in figure B1 to understand this performance. Although the uncertainty range is much lower, OOD jets generally have higher uncertainties than ID jets. This is likely due to the relationship between Max.Prob. and uncertainty displayed in figures A1(c) and (f). The uncertainties are more concerned with how likely the data is OOD rather than with misclassification. On the other hand, EDL in its basic formulation is incapable of distinguishing between hard-to-classify jets and OOD jets. This leads to a better performance compared to EDL, which ends up containing many ID jets with high uncertainties due to there hard-to-classify nature. In addition, as shown in figures B1(d)–(f), MC dropout is worse than the Ensemble method in this regard because the OOD and ID uncertainties overlap greatly. Overall, the Ensemble method is still a simpler and, based on our studies, a more accurate OOD detector as compared to EDL.

References

[1] Linardatos P, Papastefanopoulos V and Kotsiantis S 2020 Explainable AI: a review of machine learning interpretability methods *Entropy* **23** 18
[2] Neubauer M S and Roy A 2022 Explainable AI for high energy physics Contribution to the 2021 U.S. Community Study on the Future of Particle Physics (Snowmass 2021) (arXiv:2206.06632)
[3] Shanahan P, Terao K and Whiteson D, Snowmass 2021 computational frontier CompF03 topical group report: Machine learning (arXiv:2209.07559)
[4] ATLAS collaboration 2016 Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s} = 8$ tev with the ATLAS detector *J. High Energy Phys.* JHEP06(2016)093
[5] CMS Collaboration 2009 A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging *Technical Reports* (CMS-PAS-JME-09-001)
[6] CMS Collaboration 2014 *Boosted Top Jet Tagging at CMS Technical Reports* CMS-PAS-JME-13-007 CERN
[7] Baldi P, Bauer K Eng C, Sadowski P and Whiteson D 2016 Jet substructure classification in high-energy physics with deep neural networks *Phys. Rev. D* **93** 094034
[8] ATLAS Collaboration 2019 Performance of top-quark and *W*-boson tagging with ATLAS in Run 2 of the LHC *Eur. Phys. J.* C **79** 1

[9] CMS collaboration 2020 Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques *J. Instrum.*
[10] Pearkes J, Fedorko W, Lister A and Gay C 2017 Jet constituents for deep neural network based top quark tagging (arXiv:1704.02124)
[11] Moore L, Nordström K, Varma S and Fairbairn M 2019 Reports of my demise are greatly exaggerated: *N*-subjettiness taggers take on jet images *SciPost Phys.* **7** 036
[12] Datta K and Larkoski A 2017 How much information is in a jet? *J. High Energy Phys.* JHEP06(2017)073
[13] Komiske P T, Metodiev E M and Thaler J 2019 Energy flow networks: deep sets for particle jets *J. High Energy Phys.* JHEP01(2019)121
[14] Qu H and Gouskos L 2020 Jet tagging via particle clouds *Phys. Rev. D* **101** 056019
[15] Macaluso S and Shih D 2018 Pulling out all the tops with computer vision and deep learning *J. High Energy Phys.* JHEP10(2018)121
[16] Butter A, Kasieczka G, Plehn T and Russell M 2018 Deep-learned top tagging with a Lorentz layer *SciPost Phys.* **5** 028
[17] Erdmann M, Geiser E, Rath Y and Rieger M 2019 Lorentz boost networks: autonomous physics-inspired feature engineering *J. Instrum.* **14** 06006
[18] Bogatskiy A, Anderson B, Offermann J, Roussi M, Miller D and Kondor R 2020 Lorentz group equivariant neural network for particle physics *Int. Conf. on Machine Learning* (PMLR) pp 992–1002
[19] Gong S, Meng Q, Zhang J, Qu H, Li C, Qian S, Du W, Ma Z-M and Liu T-Y 2022 An efficient lorentz equivariant graph neural network for jet tagging (arXiv:2201.08187)
[20] Bogatskiy A, Hoffman T, Miller D W and Offermann J T 2022 Pelican: permutation equivariant and lorentz invariant or covariant aggregator network for particle physics (arXiv:2211.00454)
[21] Louppe G, Cho K, Becot C and Cranmer K 2019 QCD-aware recursive neural networks for jet physics *J. High Energy Phys.* JHEP01(2019)05
[22] Egan S, Fedorko W, Lister A, Pearkes J and Gay C 2017 Long short-term memory (LSTM) networks with jet constituents for boosted top tagging at the LHC (arXiv:1711.09059)
[23] Moreno E A, Cerri O, Duarte J M, Newman H B Nguyen T Q Periwal A, Pierini M, Serikova A, Spiropulu M and Vlimant J-R 2020 JEDI-net: a jet identification algorithm based on interaction networks *Eur. Phys. J. C* **80** 1
[24] Qu H, Li C and Qian S 2022 Particle transformer for jet tagging (arXiv:2202.03772)
[25] Hornik K, Stinchcombe M and White H 1989 Multilayer feedforward networks are universal approximators *Neural Netw.* **2** 359
[26] Chakraborty A, Lim S H and Nojiri M M 2019 Interpretable deep learning for two-prong jet classification with jet spectra *J. High Energy Phys.* JHEP07(2019)135
[27] Agarwal G, Hay L, Iashvili I, Mannix B, McLean C, Morris M, Rappoccio S and Schubert U 2021 Explainable AI for ML jet taggers using expert variables and layerwise relevance propagation *J. High Energy Phys.* JHEP05(2021)208
[28] Nachman B 2020 A guide for deploying deep learning in lhc searches: how to achieve optimality and account for uncertainty *SciPost Phys.* **8** 090
[29] Dorigo T and de Castro P 2021 Dealing with nuisance parameters using machine learning in high energy physics: a review (arXiv:2007.09121)
[30] Ghosh A and Nachman B 2022 A cautionary tale of decorrelating theory uncertainties *Eur. Phys. J. C* **82** 43
[31] Viren B, Huang J, Huang Y, Lin M, Ren Y, Terao K, Torbunov D and Yu H 2022 Solving simulation systematics in and with Ai/Ml Contribution to the 2021 U.S. Community Study on the Future of Particle Physics (Snowmass 2021) (arXiv:2203.06112)
[32] Vadera M P, Cobb A D, Jalaian B and Marlin B M 2020 Ursabench: Comprehensive benchmarking of approximate bayesian inference methods for deep neural networks Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning (arXiv:2007.04466)
[33] Blei D M, Kucukelbir A, McAuliffe J D and 2017 Variational inference: a review for statisticians *J. Am. Stat. Assoc.* **112** 859
[34] Lakshminarayanan B, Pritzel A and Blundell C 2017 Simple and scalable predictive uncertainty estimation using deep ensembles *Advances in Neural Information Processing Systems* p 30
[35] Kingma D P and Welling M 2022 Auto-encoding variational bayes (arXiv:1312.6114)
[36] Abdar M *et al* 2021 A review of uncertainty quantification in deep learning: techniques, applications and challenges *Inf. Fusion* **76** 243
[37] Miller T 2019 Explanation in artificial intelligence: insights from the social sciences *Artif. Intell.* **267** 1
[38] Gunning D, Stefik M, Choi J, Miller T, Stumpf S and Yang G-Z 2019 XAI-explainable artificial intelligence *Sci. Robot.* **4** eaay7120
[39] Vilone G and Longo L 2020 Explainable artificial intelligence: a systematic review (arXiv:2006.00093)
[40] Seuß D 2021 Bridging the gap between explainable AI and uncertainty quantification to enhance trustability (arXiv:2105.11828)
[41] Grojean C, Paul A, Qian Z and Strümke I 2022 Lessons on interpretable machine learning from particle physics *Nat. Rev. Phys.* **4** 284–86
[42] Khot A, Neubauer M S and Roy A 2023 A detailed study of interpretability of deep neural network based top taggers *Mach. Learn.: Sci. Technol.* **4** 035003
[43] Sensoy M, Kaplan L and Kandemir M 2018 Evidential deep learning to quantify classification uncertainty *Proc. 32nd Int. Conf. on Neural Information Processing Systems, NIPS'18,* (Curran Associates Inc.) pp 3183–93
[44] Kriesten B and Hobbs T J 2024 Anomalous electroweak physics unraveled via evidential deep learning (arXiv:2412.16286)
[45] Duarte J *et al* 2018 Fast inference of deep neural networks in FPGAs for particle physics *J. Instrum.* **13** 07027
[46] Iiyama Y, Cerminara G, Gupta A, Kieseler J, Loncar V, Pierini M, Qasim S R, Rieger M, Summers S and Van Onsem G 2021 Distance-weighted graph neural networks on fpgas for real-time particle reconstruction in high energy physics *Front. Big Data* 44
[47] Heintz A et al 2020 Accelerated charged particle tracking with graph neural networks on FPGAs (arXiv:2012.01563)
[48] Kasieczka G *et al* 2019 The machine learning landscape of top taggers *SciPost Phys.* **7** 14
[49] Kansal R, Duarte J, Su H, Orzari B, Tomei T, Pierini M, Touranakou M, Vlimant J-R and Gunopulos D 2021 Particle cloud generation with message passing generative adversarial networks *Advances in Neural Information Processing Systems* vol 34, ed M Ranzato, A Beygelzimer, Y Dauphin, P Liang and J W Vaughan (Curran Associates, Inc.) pp 23858–71
[50] Qu H, Li C and Qian S 2022 Particle transformer for jet tagging *Proc. 39th Int. Conf. on Machine Learning* (*Proc. of Machine Learning Research*) vol 162 K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and S Sabato (PMLR) (https://proceedings.mlr.press/v162/qu22b.html) pp 17–23
[51] Kendall A and Gal Y 2017 What uncertainties do we need in bayesian deep learning for computer vision? *Proc. 31st Int. Conf. on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA)* (Curran Associates Inc.) pp 5580–90
[52] Gawlikowski J *et al* 2023 A survey of uncertainty in deep neural networks *Artif. Intell. Rev.* **56** 1513–89
[53] Dempster A P 2008 Classic works of the Dempster-Shafer theory of belief functions *Stud. Fuzz. Soft Comput.* **219** 73

[54] Jøsang A 2016 *Subjective Logic: A Formalism for Reasoning Under Uncertainty* 1st edn (Springer)
[55] Top tagging dataset (available at: https://desycloud.desy.de/index.php/s/llbX3zpLhazgPJ6)
[56] Sjöstrand T, Ask S, Christiansen J R, Corke R, Desai N, Ilten P, Mrenna S, Prestel S, Rasmussen C O and Skands P Z 2015 An introduction to PYTHIA 8.2 *Comput. Phys. Commun.* **191** 159
[57] De Favereau J, Delaere C, Demin P, Giammanco A, Lemaitre V, Mertens A and Selvaggi M 2014 DELPHES 3: a modular framework for fast simulation of a generic collider experiment *J. High Energy Phys.* JHEP02(2014)057
[58] Cacciari M, Salam G P and Soyez G 2008 The anti-k_t jet clustering algorithm *J. High Energy Phys.* JHEP04(2008)
[59] Cacciari M, Salam G P and Soyez G 2012 FastJet user manual *Eur. Phys. J. C* **72** 1
[60] Kansal R, Duarte J, Su H, Orzari B, Tomei T, Pierini M, Touranakou M, Vlimant J-R and Gunopulos D 2022 JetNet: a Python package for accessing open datasets and benchmarking machine learning methods in high energy physics *J. Open Source Softw.* **8** 5789
[61] Alwall J *et al* 2014 The automated computation of tree-level and next-to-leading order differential cross sections and their matching to parton shower simulations *J. High Energy Phys.* JHEP07(2014)079
[62] Cacciari M and Salam G P 2006 Dispelling the n3 myth for the kt jet-finder *Phys. Lett.* B **641** 57
[63] Qu H, Li C and Qian S 2022 JetClass: A Large-Scale Dataset for Deep Learning in Jet Physics (https://doi.org/10.5281/zenodo.6619768)
[64] Birk J, Buhmann E, Ewen C, Kasieczka G and Shih D 2025 Flow matching beyond kinematics: generating jets with particle-id and trajectory displacement information *Phys. Rev. D* **111** 052008
[65] Moreno E A, Nguyen T Q, Vlimant J-R, Cerri O, Newman H B Periwal A, Spiropulu M, Duarte J M and Pierini M 2020 Interaction networks for the identification of boosted $h \to b\bar{b}$ decays *Phys. Rev. D* **102** 012010
[66] Gal Y and Ghahramani Z 2016 Dropout as a bayesian approximation: Representing model uncertainty in deep learning *International Conference on Machine Learning* (PMLR) pp 1050–9
[67] Gal Y and Ghahramani Z Dropout as a bayesian approximation: representing model uncertainty in deep learning *Proc. 33rd Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 48) *(20–22 Jun 2016)* ed M F Balcan and K Q Weinberger (PMLR) pp 1050–9 (https://proceedings.mlr.press/v48/gal16.html)
[68] Taylor B N and Kuyatt C E 1994 *Guidelines for Evaluating and Expressing the Uncertainty of Nist Measurement Results* (Nist Technical Note 1297) (National Institute of Standards and Technology)
[69] Gallicchio J and Schwartz M D 2013 Quark and gluon jet substructure *J. High Energy Phys.* JHEP04(2013)090
[70] Gallicchio J and Schwartz M D 2011 Quark and Gluon Tagging at the LHC *Phys. Rev. Lett.* **107** 172001
[71] ATLAS collaboration 2024 Performance and calibration of quark/gluon taggers using 140 fb^{-1} of pp collisions at $\sqrt{s} = 13$TeV with the atlas detector *Chin. Phys.* C **48** 023001
[72] Jolliffe I T and Cadima J 2016 Principal component analysis: a review and recent developments *Phil. Trans. R. Soc. A* **374** 20150202
[73] LeCun Y and Cortes C 2010 MNIST handwritten digit database (available at: http://yann.lecun.com/exdb/mnist/)
[74] Singh S K, Fowdur J S, Gawlikowski J and Medina D 2022 Leveraging graph and deep learning uncertainties to detect anomalous trajectories (arXiv:2107.01557)
[75] Gao J, Chen M, Xiang L and Xu C 2024 A comprehensive survey on evidential deep learning and its applications (arXiv:2409.04720)
[76] Shen M, Ryu J J, Ghosh S, Bu Y, Sattigeri P, Das S and Wornell G W 2024 Are uncertainty quantification capabilities of evidential deep learning a mirage? *The 38th Annual Conf. on Neural Information Processing Systems*
[77] FAIR4HEP 2025 PFIN4UQAD (available at: https://github.com/FAIR4HEP/PFIN4UQAD)