Disentangling quark and gluon jets with normalizing flows

Matthew J. Dolan[*] and Ayodele Ore[†]

ARC Centre of Excellence for Dark Matter Particle Physics, School of Physics, The University of Melbourne, Victoria 3010, Australia

(Received 11 December 2022; accepted 21 April 2023; published 1 June 2023)

The isolation of pure samples of quark and gluon jets is of key interest at hadron colliders. Recent work has employed topic modeling to disentangle the underlying distributions in mixed samples obtained from experiments. However, current implementations do not scale to high-dimensional observables as they rely on binning the data. In this work we introduce TopicFlow, a method based on normalizing flows to learn quark and gluon jet topic distributions from mixed datasets. These networks are as performant as the histogram-based approach, but since they are unbinned, they are efficient even in high dimension. The models can also be oversampled to alleviate the statistical limitations of histograms. As an example use case, we demonstrate how our models can improve the calibration accuracy of a classifier. Finally, we discuss how the flow likelihoods can be used to perform outlier-robust quark/gluon classification.

DOI: 10.1103/PhysRevD.107.114003

I. INTRODUCTION

The primary objects of study in hadronic physics at the Large Hadron Collider (LHC) are jets. A fundamental probabilistic question that particle physicists often ask regarding jets is what is the parton-level of origin of a given jet? Unfortunately for the question of quark versus gluon jet-tagging, there is no clear hadron-level definition of the distinction between a quark-jet and a gluon-jet [1,2]. While one can nonetheless train classifiers based on signal and background samples with "truth" parton-level information obtained from Monte-Carlo simulation [3–10], these definitions are subject to large modeling uncertainty [11,12]. As such, recent research has explored the utility of operational definitions. These are often based on data-driven mixture models, such as jet topics [13–15], latent Dirichlet allocation (LDA) [16–18] and others [19]. Using these methods one can disentangle samples containing multiple underlying components, with applications including anomaly detection, weakly supervised classification and studies of quantum chromodynamics (QCD). As these definitions are operational, they also have the advantage of avoiding issues such as systematic errors in the modeling of QCD and detector effects. On the other hand the extracted quark and gluon distributions may change when different algorithms are used, motivating the development of further techniques for disentangling quark and gluon distributions in LHC data.

To date, data-driven disentangling has been performed primarily in single observables, such as jet multiplicity. While jet topics and LDA technically generalize to higher dimension, their dependence on binning causes increased computational cost and statistical error in this regime. On the other hand, deep generative networks can model probability distributions in an unbinned manner and have not yet been explored for topic modeling in jet physics. Normalizing flows are examples of such generative models and have proven to be a powerful tool for many tasks in collider physics, including speeding up various parts of the event generation pipeline [20–30], detecting anomalous events [31–36] and more [37–42]. Recent reviews of LHC event generation in the context of machine learning are Refs. [43,44].

In this work, we build on the jet topics framework by leveraging normalizing flows. We train models to fit the topic distributions directly from mixed datasets, eliminating the need to subtract histograms which compounds statistical uncertainty. The benefits of our approach are therefore twofold. First, since the flows are unbinned, they can be applied efficiently in large dimension which allows high-order correlations to be captured. This opens the door for combined use with other machine learning models that are sensitive to such correlations. Second, the flows can be oversampled to obtain smooth distributions, mitigating statistical error. This is particularly relevant for high purity quark/gluon mixtures, which can only be obtained with small cross section [45]. An additional feature of the training procedure is that it explicitly encourages separation

[*]matthew.dolan@unimelb.edu.au
[†]ayodeleo@student.unimelb.edu.au

of the classes in the latent space. This can be exploited to define an outlier-robust classifier using the likelihood ratio defined by two learned topic distributions.

In the remainder of this section, we provide background on the jet topics framework before motivating a generative approach and describing the training procedure for our normalizing flows. Section II details our specific datasets, network architecture and training parameters. In Sec. III we demonstrate the quality with which our normalizing flows can distill underlying jet topic distributions, comparing to the histogram-based construction. We also present two use cases: improving data-driven classifier calibration and performing generative classification. We provide concluding remarks and discuss further applications of these models in Sec. IV.

A. Jet topics

We consider samples of jets as unlabeled statistical mixtures of quarks and gluons. Assuming that two mixtures M_1 and M_2 consist of identical quark and gluon component distributions p_Q and p_G, the probability densities for an observable x can be written as

$$\begin{aligned} p_{M_1}(x) &= f_1 p_Q(x) + (1-f_1) p_G(x) \\ p_{M_2}(x) &= f_2 p_Q(x) + (1-f_2) p_G(x), \end{aligned} \tag{1}$$

with f_1 and f_2 being the unknown quark fractions of the mixtures. In the jet topics framework [13–15], one further assumes that the component distributions are *mutually irreducible*, meaning that there exist regions of phase space which can be identified as being either pure-quark or pure-gluon. The mixtures can be disentangled using the DEMIX algorithm [46], which proceeds by calculating *reducibility factors* defined as

$$\kappa_{ij} = \min_x \frac{p_{M_i}(x)}{p_{M_j}(x)}, \tag{2}$$

for $i, j \in \{1, 2\}$. These factors correspond to the maximal amount by which p_{M_j} can be subtracted from p_{M_i} while ensuring the result is positive-definite. Reference [15] provides a number of methods for calculating these factors, the simplest of which amounts to binning the mixtures in the observable x and performing the minimization of Eq. (2) over these bins. The underlying "topic" distributions may then be reconstructed as

$$\begin{aligned} p_Q(x) &= \frac{p_{M_1}(x) - \kappa_{12} p_{M_2}(x)}{1 - \kappa_{12}} \\ p_G(x) &= \frac{p_{M_2}(x) - \kappa_{21} p_{M_1}(x)}{1 - \kappa_{21}}. \end{aligned} \tag{3}$$

Importantly, the observable x in the above reconstruction need not be the same as that which was used to extract the reducibility factors. This is because the mixture fractions are collective properties of the samples, not the individual jets. The jet topics framework has proven to be powerful for examining quarks and gluons separately, both from theoretical [14,47–49] and experimental perspectives [15,50–53].

In existing practical studies that use jet topic disentangling, the pure distributions of Eq. (3) are constructed from histograms of the training mixtures. In principle one could use these histograms to estimate likelihoods according to the bin occupancy or generate samples using the inverse cumulative probability. However as one considers larger dimension, the computational cost[1] of these tasks increases and the density of the dataset within the space decays exponentially, which would lead to large errors. In addition to this curse of dimensionality, statistical errors are amplified when the bin values are subtracted during construction of the topics.

We propose to alleviate these issues with deep generative networks, which are able to model high-dimensional probability distributions including correlations. Since these models are unbinned, they have more efficient scaling to high dimension. Further, if an appropriate training procedure is used, one can learn the topic distributions directly from the quark/gluon mixtures, without the need for explicit subtraction.

B. Normalizing flows

In this work, we model the topic distributions using normalizing flows (NFs) [54–56]. An NF is a density estimation model that aims to fit an unknown probability distribution from finite data samples. Being an unsupervised method, these models have gained popularity in collider physics for anomaly detection [32–36]. Furthermore, a key property of NFs is that they are invertible, and so they can also be used as generative models. This mode of operation has also been explored in a particle physics context [23–31,40].

In essence, an NF represents a mapping $g\colon Z \to X$ from a known latent distribution p_Z (often a Gaussian) in a latent space Z, to a distribution q in the data space X of the same dimension. If g is invertible, then the density of a data point $x \in X$ is given by the change of variables formula,

$$q(x) = p_Z(g^{-1}(x)) \left| \det \frac{\partial g^{-1}}{\partial x} \right|. \tag{4}$$

The map g is usually parametrized by a neural network with weights θ. The network can be trained to model an arbitrary data distribution p_X by minimizing the Kullback-Leibler (KL) divergence between p_X and q_θ. One can show that this is equivalent to minimizing the negative log-likelihood of data samples under the model:

[1]Both computation time and memory requirements scale with the number of histogram bins.

$$\begin{aligned}\mathrm{KL}(p_X||q_\theta) &= \int \mathrm{d}x\, p_X(x) \log\left(\frac{p_X(x)}{q_\theta(x)}\right) \\ &= \left\langle \log\left(\frac{p_X(x)}{q_\theta(x)}\right)\right\rangle_{x\sim p_X} \\ &= -\langle \log q_\theta(x)\rangle_{x\sim p_X} + \text{const.} \end{aligned} \tag{5}$$

Training against the above objective requires that one can sample directly from the target distribution. In the case of quark and gluon jets, we instead have access only to mixed datasets M_i. However, Eq. (3) describes the pure distributions as linear combinations of these mixture distributions, which can be substituted into the equation above in order to split the integral. Taking the quark distribution as an example, this leads to

$$\begin{aligned}\mathrm{KL}(p_Q||q_\theta) = \kappa_{12}\langle \log q_\theta(x)\rangle_{x\sim p_{M_2}} - \langle \log q_\theta(x)\rangle_{x\sim p_{M_1}} \\ + \text{const}, \end{aligned} \tag{6}$$

and similarly for the gluon distribution. Thus we can use batches from the mixed datasets in order to learn the correct underlying components. Intuitively, the NF is trained to increase the likelihood of one mixture while decreasing the likelihood of the other. Since the only statistical difference between the datasets is their mixture fraction, the flow will learn the distribution of whichever class the first mixture is enriched with compared to the second.

This approach to training a normalizing flow is not limited to jet topics. Indeed, the expansion that lead to Eq. (6) can be applied to any distribution defined as a linear combination. For example, this loss function can be used to train flows to perform event subtraction, which has been explored using generative adversarial networks (GANs) [57].

II. TRAINING DETAILS

A. Datasets

To emulate impure samples, we use the quark/gluon dataset of Refs. [58,59] which consists of 2M light quark (uds) and gluon jets generated with PYTHIA 8.2 [60]. We use 15% for testing, 10% for validation and up to 75% for training. The jets come clustered by the anti-k_T algorithm [61] with radius $R = 0.4$ and satisfy $p_T \in [500, 550]$ GeV and $|y| < 1.7$. Mixed datasets are constructed by collecting the appropriate number of quarks and gluons from each subset given the desired quark purity of each sample. As such, we have statistically identical underlying quark/gluon distributions in all mixtures as required by the jet topics framework.[2] We also have access to the exact reducibility factors of the mixtures. While in practice these would first need to be estimated using one of the methods described in Ref. [15], we use the true values for simplicity.

In principle our procedure does not depend critically on the jet representation, so long as the representation can be integrated with a normalizing flow architecture. We choose to represent the jets as sets of energy flow polynomials (EFPs) [63], a complete basis of IRC-safe jet observables. An EFP can be identified with a multigraph where each node gives an energy-weighted sum over particles, and each edge denotes an opening angle between two particles. The polynomial corresponding to a particular graph G with N nodes is

$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(j,k)\in G} \theta_{i_j i_k}^{\beta}, \tag{7}$$

where M is the number of jet constituents, $z_a = p_{T,a}/p_{T,J}$ is the p_T fraction of constituent a relative to the jet and β is a parameter. EFPs are suitable for the present study since normalizing flows require a fixed-size representation. They are also strong quark/gluon discriminants. In fact a growing collection of recent work has demonstrated that, with careful selection, only a small number of EFPs are required to achieve classification on par with large neural networks based on low-level information [64–69]. As such, we opt to use a small set, namely the $\beta = \frac{1}{2}$ connected EFPs with less than or equal to 3 edges, of which there are 8 elements.[3] Training in higher dimension is also possible, but requires larger networks in general (See Ref. [70] for a dedicated study on NF scaling to high dimension as well as Refs. [26–29] for examples). That said, even in 8 dimensions, working with a histogram is impractical since just 10 bins per dimension leads to 10^8 total bins.

To improve the stability of the models' training, we preprocess the EFPs by applying a log-scaling then translating the datasets to have zero mean (across quark and gluon jets). The flows also benefit from removing linear correlations in the data, so we additionally change basis to the principal components. The mean and covariance matrix are computed from the training split of each dataset. Since each of these preprocessing steps is invertible, the generated samples can always be converted back into the original EFP basis.

B. Models

We employ continuous-time normalizing flows [71] as our generative models. Specifically, we use the `tensorflow_probability` [72] implementation of FFJORD [73]. The derivative network is a residual network with two blocks, each containing two layers of 256 nodes. We use tanh activations and minimize the loss of Eq. (6) with the

[2]This condition is not necessarily satisfied in experimental data [15,62].

[3]We also exclude the single $d = 0$ EFP under which all jets are degenerate.

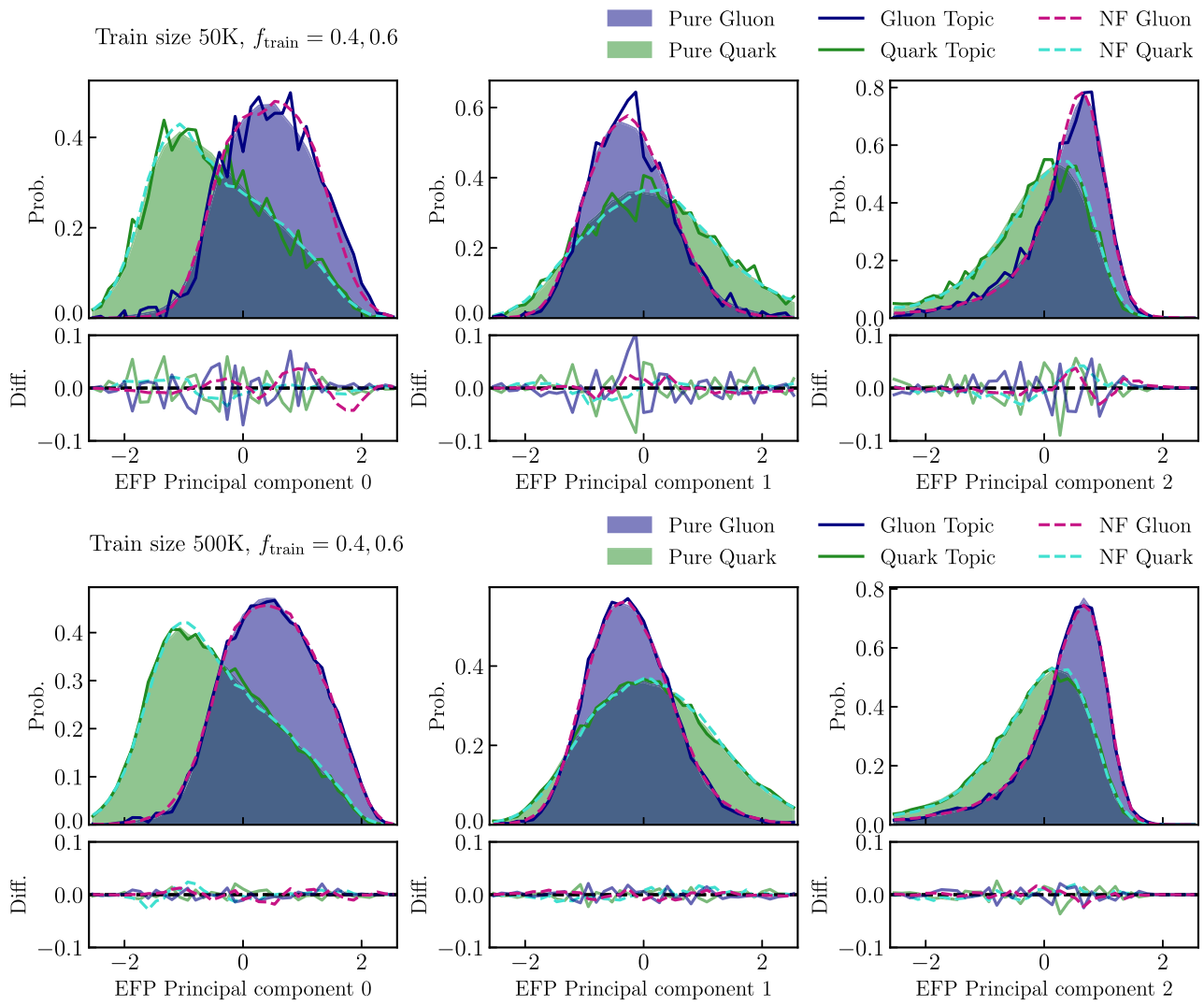

FIG. 1. Validation of EFPs (3 of 8 components shown) generated by a normalizing flow trained on 40% and 60% quark mixtures. The shaded distributions are the pure quarks and gluons from the PYTHIA test set and the solid lines are jet topic distributions constructed from the training set. The top and bottom rows are results for training set sizes of 50 K and 500 K respectively. Plots of the remaining components can be found in Appendix.

Adam optimizer. We set the batch size to either 1000 or the largest value that allows 25 batches per epoch, whichever is smaller. A callback that monitors the validation loss over each epoch of training is used to reduce the learning rate by a factor 10 if no improvement is observed after 5 epochs. We use an initial learning rate of 10^{-3} and halt training once the learning rate falls below 10^{-6}.

III. RESULTS

A. Sample quality

We demonstrate the performance of the trained models in Fig. 1, where the pure EFP distributions from the PYTHIA test set are compared with equal-sized samples generated with the flows. We show the first three components out of the eight which we use. We also compare to the jet topic distributions constructed from the training set. The plots show results for a small training set containing 50 K jets (top row) as well as a large training set containing 500 K (bottom row). The lower subpanels in each plot show the absolute difference between the topic/NF and the pure distribution derived from PYTHIA. In the smaller training set, the histogram-based topics exhibit large statistical fluctuations which are not present in the NF distribution, since (a) no subtraction is performed and (b) the models have been sampled beyond the size of the training set. The flow samples instead exhibit systematic error corresponding to the fact that the networks cannot be trained to perfect

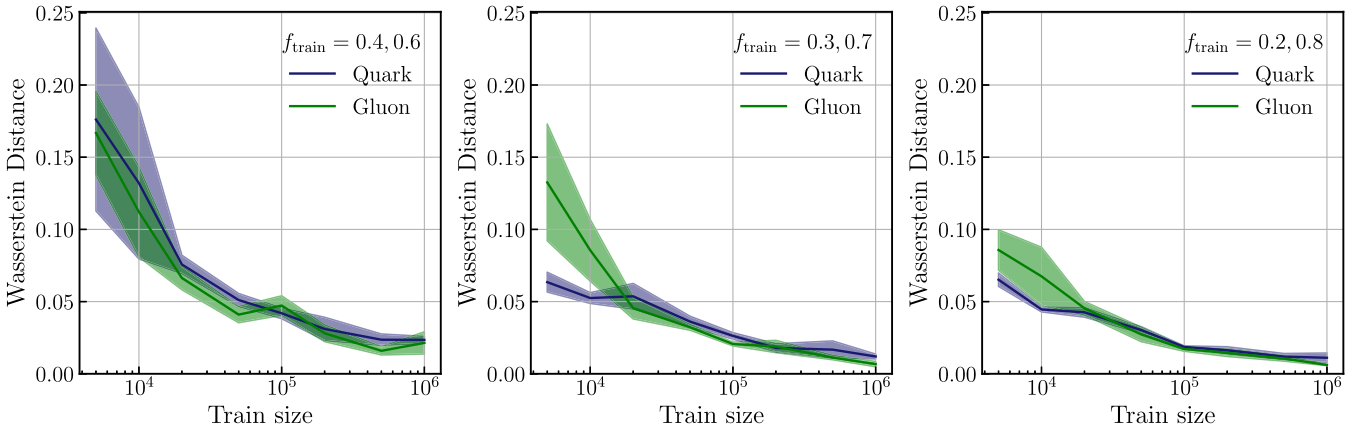

FIG. 2. Average Wasserstein distances between EFP distributions in the PYTHIA test set and those generated from networks trained on mixtures of different purity. Error bands show one standard deviation over 5 runs.

convergence. The systematic variations of the flow appear to be equal or smaller than the statistical fluctuations of the binned topics. In the larger dataset, the statistical and systematic errors are diminished and both methods agree with the pure distributions derived from the truth-labels from PYTHIA. We find that the fractional errors for the 50 K training sample are around 20% and around 5% for the 500 K training sample near the peaks of the distributions for both standard topics and normalizing flows, although these become larger in the tails where statistical uncertainties are more important.

To concretely measure the quality of a trained NF, we evaluate 1-Wasserstein distances between the testing sample and an equal-size sample generated by the flow for each EFP in the set. Figure 2 shows plots of the average of these distances against the size of the training set for mixtures with quark fractions (0.4, 0.6), (0.3, 0.7) and (0.2, 0.8).[4] As expected, the agreement between the test set and the NF samples improves with the size of the training set. For the most part, the quark and gluon distributions are reproduced equally well, however for small datasets, the flows do not perform as well for gluons. This is due to a tendency for the models to overfit in these settings, and so training was halted early. This could likely be addressed by optimizing the network size depending on the dataset at hand. One could also set the reducibility factor in Eq. (6) to zero for some warm up period during training, effectively pretraining the model on one of the available mixtures.

The figure also reveals a dependence on the purity of the training mixtures. The NFs perform better when trained on mixtures with quark fractions away from 0.5, particularly when the dataset is small. This simply reflects the fact that optimizing the objective Eq. (6) is more difficult when M_1 and M_2 are similar.

Since the NF models represent multi-dimensional probability distributions, correlations between EFPs should also be captured. To judge the pairwise correlations produced by the flow models we plot the quark and gluon distributions projected onto a selection of 2-dimensional planes in the full EFP space, shown in Fig. 3. The heat maps and solid black contours correspond to the (truth) PYTHIA distributions and the dashed contours are produced from equal-size samples of a trained NF. The nontrivial profiles of these distributions illustrate that correlations indeed remain between the components even after rotating to the principal component basis. The two sets of contours agree to a precise level, demonstrating that the generative model indeed produces the correct correlations. As a consequence, any cuts on a particular EFP in a generated sample will appropriately affect the distributions of the other EFPs without sacrificing statistics (since more samples can always be generated from the flow).

B. Classifier calibration

One of the advantages of an operational definition of quarks and gluons is that the performance of a classifier can be evaluated in a data-driven manner. As discussed in Ref. [14], this enables weakly supervised classifiers to self-calibrate using jet topics defined by their output. This proceeds by training a classifier on mixtures M_1 and M_2, then extracting reducibility factors and corresponding mixture fractions from the output x using Eq. (2). Signal and background efficiencies can then be calculated for a given threshold t on the output according to

$$\epsilon_Q(t) = \frac{p_{M_1}(x > t)(1 - f_2) - p_{M_2}(x > t)(1 - f_1)}{f_1 - f_2},$$
$$\epsilon_G(t) = \frac{p_{M_2}(x > t) f_1 - p_{M_1}(x > t) f_2}{f_1 - f_2}. \tag{8}$$

[4]We present results only for symmetric mixture purity ($f_2 = 1 - f_1$) to ensure that we have balanced training sets. This is not a limitation of the method in general, and we verified that results of equal quality are obtained using asymmetric purity.

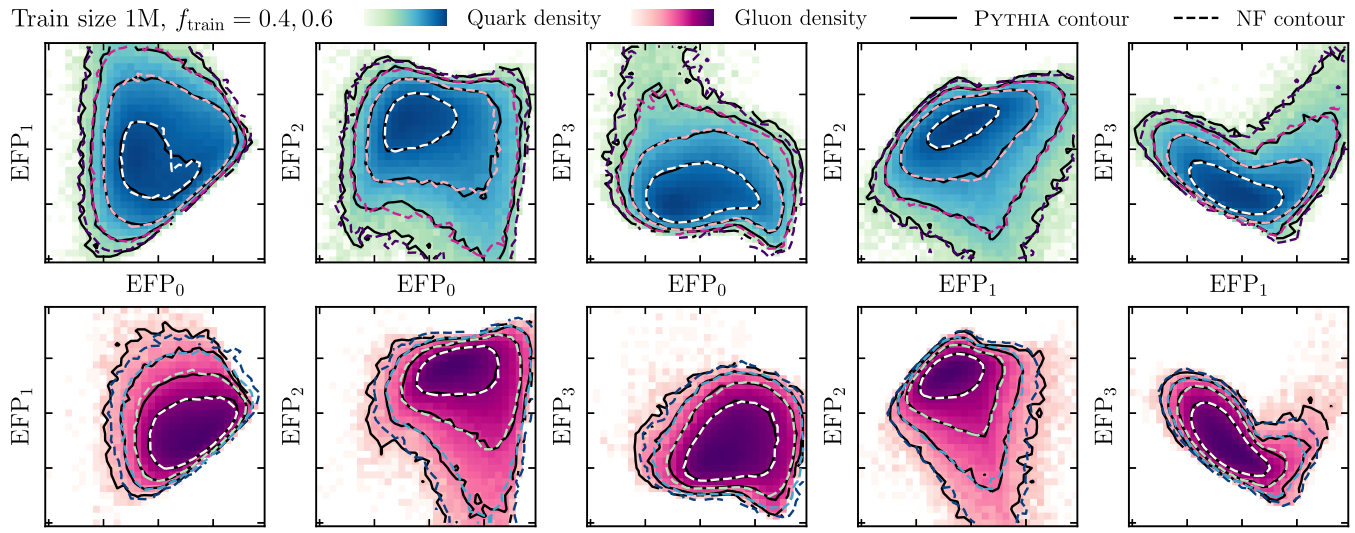

FIG. 3. Comparison between the correlations among EFPs present in the PYTHIA dataset and NF-generated samples (five of 28 pairs shown). The heat maps represent the simulated data density, with darker colors indicating larger density. The solid black lines are the contours of this density and the dashed colored contours map the density of generated samples. The top and bottom rows correspond to quarks and gluons, respectively. Plots of the remaining pairs can be found in Appendix.

These efficiencies are again subject to amplified statistical uncertainty due to the subtraction. Thus errors can be quite large, especially in regions of low efficiency. However, one is often interested in regions of low background efficiency, especially for powerful classifiers. We find that even for sizeable datasets ($\sim 10^5$ events) there can be insufficient statistics to produce smooth ROC curves throughout the entire region of interest. This is because larger datasets train stronger classifiers which increase class separation and thereby preserve sparsely occupied regions of the output space. This usually leads to a systematic underestimate of the background efficiency and can prevent reliable calibration.

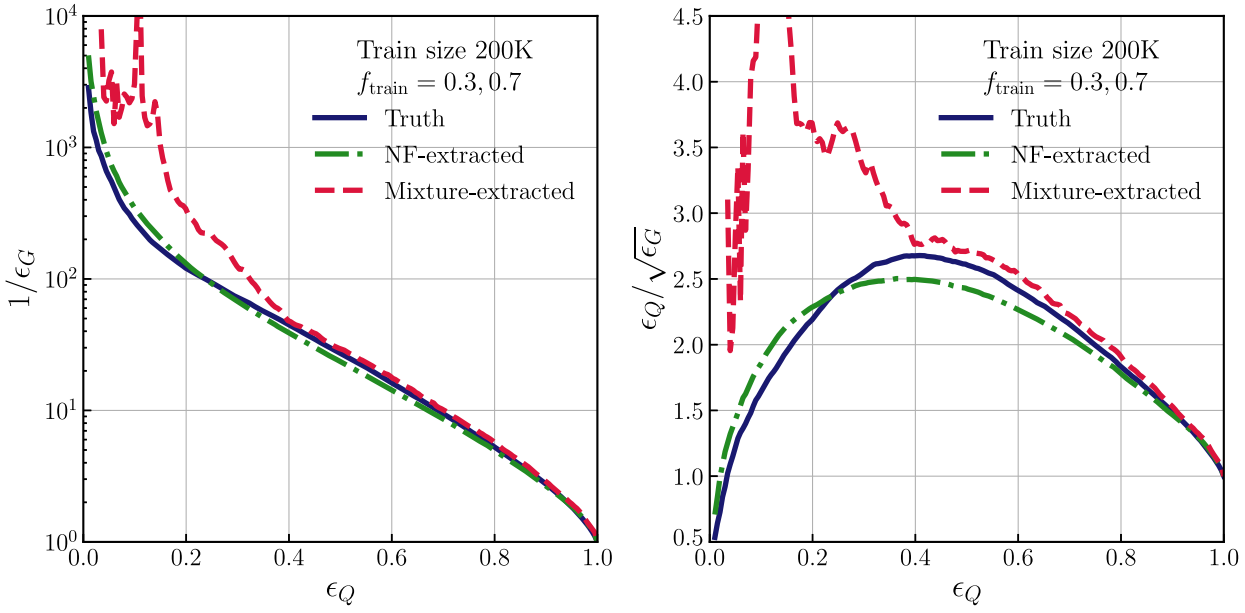

FIG. 4. ROC (left) and SI (right) curves of a CWoLa classifier evaluated using Eq. (8) applied to the training mixtures and using samples from NF-learned topic distributions.

As a concrete example, we train a CWoLa classifier [74] and plot the receiver operating characteristic (ROC) and significance improvement (SI) curves yielded by Eq. (8), comparing with the PYTHIA truth. We show this in Fig. 4 as the red dashed line. At high quark efficiencies, the mixture-extracted curves agree well with the truth. However, below an efficiency of 0.4 the curves diverge due to low statistics. Since the deviation causes the curve to reach its maximum significance improvement far from the true value, it is not an accurate means of calibration.

We can address this sensitivity to the training statistics with dedicated generative models for the pure distributions. Instead of using Eq. (8), samples can be generated from each flow and passed through the classifier yielding the "pure" class predictions.[5] These predictions can then be used to estimate efficiencies directly, without the need for any subtraction. We emphasize that oversampling classifier outputs in this way relies on full multidimensional generative models for the topic distributions since the classifier may be sensitive to high-order correlations.

The ROC and SI curves that we evaluate using the flow models are also shown in Fig. 4. Compared to the mixture-extracted curves, those obtained from the NF samples remain close to the truth for a larger range of efficiencies. Critically, the location of the maximum significance improvement agrees closely with the truth. Once again, the trade-off is between statistical and systematic error, although we generally find that it is better to use the generative model.

C. Generative classification

In addition to sampling, another advantage of unbinned models for pure quark and gluon distributions is their ability to evaluate likelihoods of points in the full data space. In particular, one can use the trained normalizing flows to construct a quark/gluon classifier via the likelihood ratio $p_Q(x)/p_G(x)$.[6] Such a model is known as a generative classifier [75]. Compared with typical discriminative classifiers, these models are often not as performant since they model the likelihood functions globally, whereas only the decision boundary is relevant for classification. Indeed, we find that likelihood ratios constructed with our normalizing flows are unable to match a discriminative CWoLa classifier. However, using TopicFlow to learn the quark/gluon ratio leads to much stronger classification than using standard normalizing flows to construct the mixture ratio defined as $p_{M_1}(x)/p_{M_2}(x)$, despite the fact that the true ratios are related monotonically. Figure 5, shows the significance improvement curves for these generative classifiers, compared to a green discriminative CWoLa classifier.[7] The performance of the NF mixture ratio can be explained by the poor out-of-distribution sensitivity of normalizing flows [76]. Specifically, we find that gluons typically have higher likelihood than quarks even under an NF trained on a quark-enriched mixture. However, the positive term in Eq. (6) explicitly penalizes this for TopicFlow, leading to better-separated likelihoods and thereby better classification.

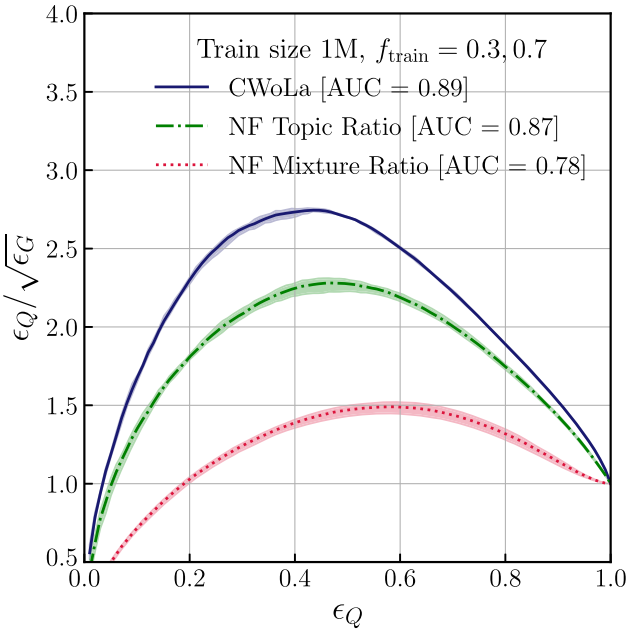

FIG. 5. SI curves for a discriminative CWoLa classifier (5×100 fully connected layers) and two NF-based generative classifiers. The topic and mixture ratios are $p_Q(x)/p_G(x)$ and $p_{M_1}(x)/p_{M_2}(x)$ respectively. Efficiencies are evaluated using PYTHIA and error bands show the range of 5 independent classifiers.

While the raw performance of the quark/gluon NF ratio does not match CWoLa, generative classifiers offer another benefit. Specifically, they provide a means of detecting anomalous events that have small likelihoods under both quark and gluon models. This sort of outlier detection is absent in discriminative models since their class predictions are normalized [77,78]. For example, a top-quark jet may have a small likelihood under both the light quark and gluon topic distributions, yet still be more compatible with the quark topic. In such a case, a discriminative classifier may confidently predict the jet to belong to the light quark distribution. However, in the generative case, one could use

[5]One could also train an NF to disentangle the outputs themselves, although this would need to be done separately for each classifier.

[6]In practice we use the log likelihood ratio.

[7]For these results, we use a normalizing flow architecture consisting of coupling blocks with rational quadratic spline transformations. Our continuous-time normalizing flows yielded extremely poor classification.

the small individual likelihoods to identify the jet as out-of-distribution.

This behavior is particularly desirable for application to real data, where the datasets are not guaranteed to contain only quark or gluon jets. By defining an anomaly score using individual likelihoods, the total significance improvement of a generative classifier may exceed the CWoLa tagger when including a cut on this score.

IV. CONCLUSIONS AND OUTLOOK

We have introduced TopicFlow, a method of modeling jet topic distributions using normalizing flows. Our approach has two advantages over the standard procedure based on histograms. First, the generative models can be oversampled to produce smooth distributions, avoiding amplified statistical uncertainty that arises when subtracting histograms. Second, the absence of binning allows the normalizing flows to be applied in high dimension and thus capture complex correlations in the data. Samples from the models are therefore suitable as input to other machine learning models. As an example, we demonstrated that our flows can enable self-calibration of a classifier that would otherwise be limited by a lack of statistics.

Normalizing flows allow the possibility of generative classification via use of the likelihood ratio. We found that using the ratio defined by TopicFlow is superior to using the ratio of mixture likelihoods with standard normalizing flows, though still not as performant a dedicated discriminative classifier based on the CWoLa method. However, classification using the individual likelihoods in this way may be more robust to out-of-distribution events in a testing sample and therefore warrants further study.

While the models in this work were trained on simulation with known quark fractions, the general framework can be extended to an unsupervised setting by first extracting mixture fractions/reducibility factors using existing methods such as [15,19]. As such a sensible direction for future work is to apply these networks to real jets in CMS Open Data.

It would also be interesting to explore particle-level jet representations such as point clouds. In this case, sampling allows for the determination of arbitrary jet observable distributions. Recent work has shown that normalizing flows can be trained effectively on such representations [28,29]. GANs are also a natural choice for point cloud data [79] and could be applied to jet topics by training with the event subtraction method of Ref. [57]. Conversely, one could use our training procedure to perform event subtraction with normalizing flows.

Other directions also include automatic extraction of the reducibility factors, uncertainty quantification with Bayesian neural networks and domain adaptation (in the jet p_T for example) with conditional generative models. We hope to explore some of these in future work.

V. DATA AVAILABILITY

Our code is publicly available and can be found at https://github.com/ayo-ore/topicflow.

ACKNOWLEDGMENTS

M. J. D. is supported by the Australian Research Council Future Fellowship FT180100324. A. O. is supported by the Australian Government Research Training Program Scholarship initiative. Computing resources were provided by the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of Linkage Infrastructure, Equipment and Facilities (LIEF) Grant No. LE170100200.

APPENDIX: ADDITIONAL PLOTS

Here we present extended versions of the plots in Sec. III A that include all EFP dimensions. Figure 6 extends Fig. 1 and Fig. 7 extends Fig. 3.

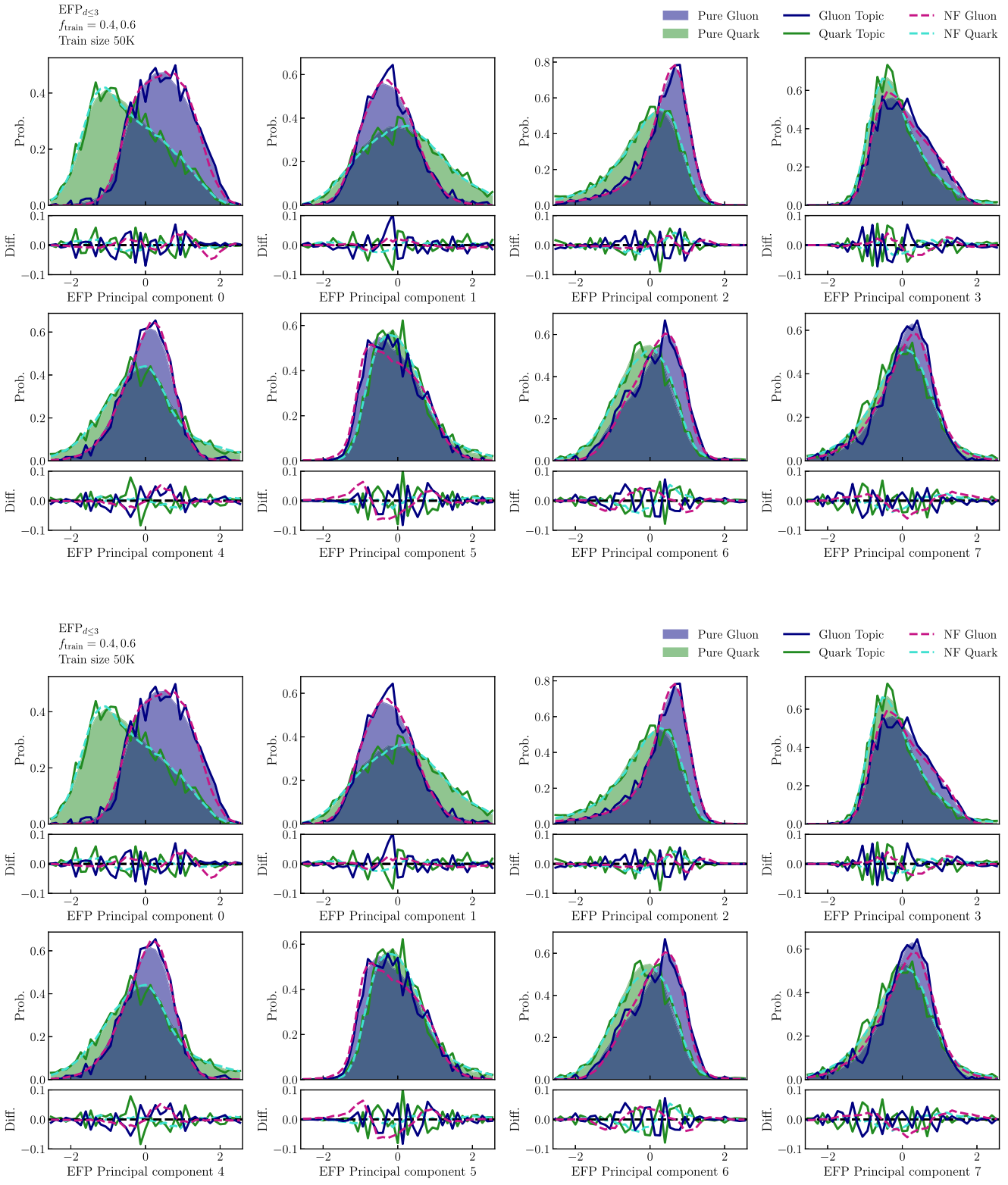

FIG. 6. Comparison of quark and gluon EFP distributions as in Fig. 1, but including all principal components.

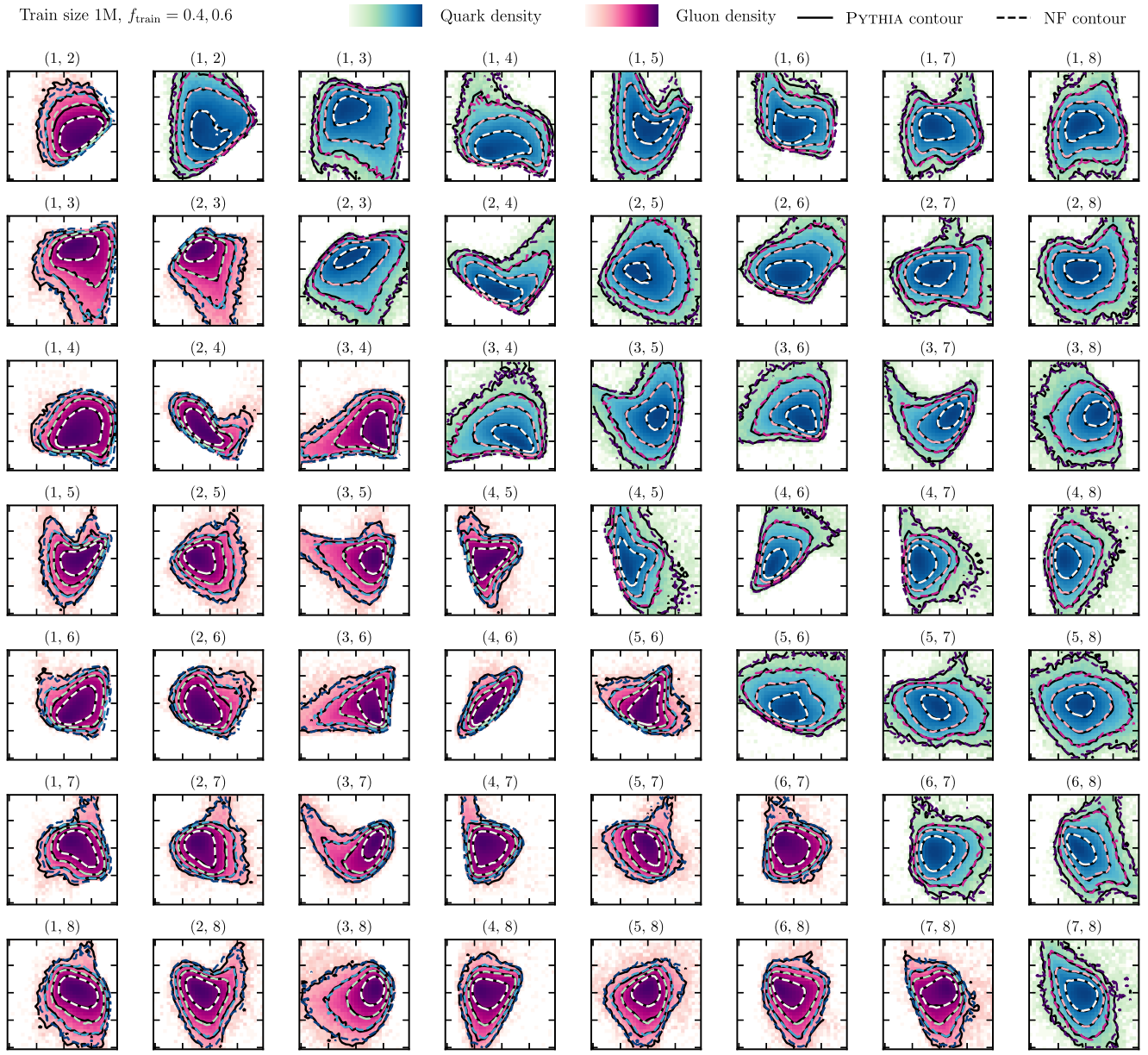

FIG. 7. Comparison of pairwise EFP correlations as in Fig. 3, but including all pairs between EFPs in the set.

[1] A. Banfi, G. P. Salam, and G. Zanderighi, Infrared safe definition of jet flavor, Eur. Phys. J. C **47**, 113 (2006).
[2] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer, A. Siódmok, P. Skands, G. Soyez, and J. Thaler, Systematics of quark/gluon tagging, J. High Energy Phys. 07 (2017) 091.
[3] J. Gallicchio and M. D. Schwartz, Quark and Gluon Tagging at the LHC, Phys. Rev. Lett. **107**, 172001 (2011).
[4] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, Deep learning in color: Towards automated quark/gluon jet discrimination, J. High Energy Phys. 01 (2017) 110.
[5] T. Cheng, Recursive neural networks in quark/gluon tagging, Comput. Software Big Sci. **2**, 3 (2018).
[6] G. Kasieczka, N. Kiefer, T. Plehn, and J. M. Thompson, Quark-gluon tagging: Machine learning vs detector, SciPost Phys. **6**, 069 (2019).

[7] J. S. H. Lee, I. Park, I. J. Watson, and S. Yang, Quark-gluon jet discrimination using convolutional neural networks, J. Korean Phys. Soc. **74**, 219 (2019).
[8] F. A. Dreyer, G. Soyez, and A. Takacs, Quarks and gluons in the Lund plane, J. High Energy Phys. 08 (2022) 177.
[9] A. Romero, D. Whiteson, M. Fenton, J. Collado, and P. Baldi, Safety of quark/gluon jet classification, arXiv:2103.09103.
[10] S. Bright-Thonney, I. Moult, B. Nachman, and S. Prestel, Systematic quark/gluon identification with ratios of likelihoods, J. High Energy Phys. 12 (2022) 021.
[11] D. Reichelt, P. Richardson, and A. Siodmok, Improving the simulation of quark and gluon jets with Herwig 7, Eur. Phys. J. C **77**, 876 (2017).
[12] J. Mo, F. J. Tackmann, and W. J. Waalewijn, A case study of quark-gluon discrimination at NNLL' in comparison to parton showers, Eur. Phys. J. C **77**, 770 (2017).
[13] E. M. Metodiev and J. Thaler, Jet Topics: Disentangling Quarks and Gluons at Colliders, Phys. Rev. Lett. **120**, 241602 (2018).
[14] P. T. Komiske, E. M. Metodiev, and J. Thaler, An operational definition of quark and gluon jets, J. High Energy Phys. 11 (2018) 059.
[15] P. T. Komiske, S. Kryhin, and J. Thaler, Disentangling quarks and gluons with CMS open data, Phys. Rev. D **106**, 094021 (2022).
[16] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, Uncovering latent jet substructure, Phys. Rev. D **100**, 056002 (2019).
[17] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc, Learning the latent structure of collider events, J. High Energy Phys. 10 (2020) 206.
[18] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc, Learning latent jet structure, Symmetry **13**, 1167 (2021).
[19] E. Alvarez, M. Spannowsky, and M. Szewc, Unsupervised quark/gluon jet tagging with Poissonian mixture models, Front. Artif. Intell. **5**, 852970 (2022).
[20] B. Stienen and R. Verheyen, Phase space sampling and inference from weighted events with autoregressive flows, SciPost Phys. **10**, 038 (2021).
[21] E. Bothmann, T. Janßen, M. Knobbe, T. Schmale, and S. Schumann, Exploring phase space with neural importance sampling, SciPost Phys. **8**, 069 (2020).
[22] C. Gao, J. Isaacson, and C. Krause, i-flow: High-dimensional integration and sampling with normalizing flows, Mach. Learn. Sci. Tech. **1**, 045023 (2020).
[23] C. Gao, S. Höche, J. Isaacson, C. Krause, and H. Schulz, Event generation with normalizing flows, Phys. Rev. D **101**, 076002 (2020).
[24] Y. Lu, J. Collado, D. Whiteson, and P. Baldi, Sparse autoregressive models for scalable generation of sparse images in particle physics, Phys. Rev. D **103**, 036012 (2021).
[25] A. Butter, T. Heimel, S. Hummerich, T. Krebs, T. Plehn, A. Rousselot, and S. Vent, Generative networks for precision enthusiasts, arXiv:2110.13632.
[26] C. Krause and D. Shih, CaloFlow: Fast and accurate generation of calorimeter showers with normalizing flows, arXiv:2106.05285.
[27] C. Krause and D. Shih, CaloFlow II: Even faster and still accurate generation of calorimeter showers with normalizing flows, arXiv:2110.11377.
[28] B. Käch, D. Krücker, I. Melzer-Pellmann, M. Scham, S. Schnake, and A. Verney-Provatas, JetFlow: Generating jets with conditioned and mass constrained normalising flows, arXiv:2211.13630.
[29] B. Käch, D. Krücker, and I. Melzer-Pellmann, Point cloud generation using transformer encoders and normalising flows, arXiv:2211.13623.
[30] R. Verheyen, Event generation and density estimation with surjective normalizing flows, SciPost Phys. **13**, 047 (2022).
[31] S. Choi, J. Lim, and H. Oh, Data-driven estimation of background distribution through neural autoregressive flows, arXiv:2008.03636.
[32] B. Nachman and D. Shih, Anomaly detection with density estimation, Phys. Rev. D **101**, 075042 (2020).
[33] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, Classifying anomalies through outer density estimation (CATHODE), Phys. Rev. D **106**, 055006 (2022).
[34] P. Jawahar, T. Aarrestad, N. Chernyavskaya, M. Pierini, K. A. Wozniak, J. Ngadiuba, J. Duarte, and S. Tsan, Improving variational autoencoders for new physics detection at the LHC with normalizing flows, Front. Big Data **5**, 803685 (2022).
[35] A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, T. Plehn, D. Shih, and R. Winterhalder, Ephemeral learning—augmenting triggers with online-trained normalizing flows, SciPost Phys. **13**, 087 (2022).
[36] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, Resonant anomaly detection without background sculpting, arXiv:2210.14924.
[37] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone, and U. Köthe, Invertible networks or partons to detector and back again, SciPost Phys. **9**, 074 (2020).
[38] S. Bieringer, A. Butter, T. Heimel, S. Höche, U. Köthe, T. Plehn, and S. T. Radev, Measuring QCD splittings with invertible networks, SciPost Phys. **10**, 126 (2021).
[39] R. Winterhalder, V. Magerya, E. Villa, S. P. Jones, M. Kerner, A. Butter, G. Heinrich, and T. Plehn, Targeting multi-loop integrals with neural networks, SciPost Phys. **12**, 129 (2022).
[40] R. Winterhalder, M. Bellagente, and B. Nachman, Latent space refinement for deep generative models, arXiv:2106.00792.
[41] A. Butter, T. Heimel, T. Martini, S. Peitzsch, and T. Plehn, Two invertible networks for the matrix element method, arXiv:2210.00019.
[42] S. Klein and T. Golling, Decorrelation with conditional normalizing flows, arXiv:2211.02486.
[43] A. Butter and T. Plehn, Generative networks for LHC events, arXiv:2008.08558.
[44] S. Badger *et al.*, Machine learning and LHC event generation, arXiv:2203.07460.
[45] J. Gallicchio and M. D. Schwartz, Pure samples of quark and gluon jets at the LHC, J. High Energy Phys. 10 (2011) 103.

[46] G. Blanchard and C. Scott, Decontamination of Mutually Contaminated Models, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 33, edited by S. Kaski and J. Corander (PMLR, Reykjavik, Iceland, 2014), pp. 1–9.
[47] A. J. Larkoski and E. M. Metodiev, A theory of quark vs. gluon discrimination, J. High Energy Phys. 10 (2019) 014.
[48] A. Takacs and K. Tywoniuk, Quenching effects in the cumulative jet spectrum, J. High Energy Phys. 10 (2021) 038.
[49] I. W. Stewart and X. Yao, Pure quark and gluon observables in collinear drop, J. High Energy Phys. 09 (2022) 120.
[50] G. Aad *et al.* (ATLAS Collaboration), Properties of jet fragmentation using charged particles measured with the ATLAS detector in pp collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. D **100**, 052011 (2019).
[51] J. Brewer, J. Thaler, and A. P. Turner, Data-driven quark and gluon jet modification in heavy-ion collisions, Phys. Rev. C **103**, L021901 (2021).
[52] Y. Ying, J. Brewer, Y. Chen, and Y.-J. Lee, Data-driven extraction of the substructure of quark and gluon jets in proton-proton and heavy-ion collisions, arXiv:2204.00641.
[53] M. LeBlanc, B. Nachman, and C. Sauer, Going off topics to demix quark and gluon jets in α_S extractions, J. High Energy Phys. 02 (2023) 150.
[54] D. J. Rezende and S. Mohamed, Variational inference with normalizing flows, in *Proceedings of the 32nd International Conference on Machine Learning* (PMLR, 2015), pp. 1530–1538, https://proceedings.mlr.press/v37/rezende15.html.
[55] I. Kobyzev, S. J. Prince, and M. A. Brubaker, Normalizing flows: An introduction and review of current methods, IEEE Trans. Pattern Anal. Mach. Intell. **43**, 3964 (2021).
[56] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, J. Mach. Learn. Res. **22**, 1 (2021), https://jmlr.org/papers/v22/19-1028.html.
[57] A. Butter, T. Plehn, and R. Winterhalder, How to GAN event subtraction arXiv:1912.08824.
[58] P. T. Komiske, E. M. Metodiev, and J. Thaler, PYTHIA8 quark and gluon jets for energy flow, Zenodo, 10.5281/zenodo.3164691 (2019).
[59] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, J. High Energy Phys. 01 (2019) 121.
[60] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015).
[61] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[62] S. Bright-Thonney and B. Nachman, Investigating the topology dependence of quark and gluon jets, J. High Energy Phys. 03 (2019) 098.
[63] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow polynomials: A complete linear basis for jet substructure, J. High Energy Phys. 04 (2018) 013.
[64] T. Faucett, J. Thaler, and D. Whiteson, Mapping machine-learned physics into a human-readable space, Phys. Rev. D **103**, 036020 (2021).
[65] J. Collado, K. Bauer, E. Witkowski, T. Faucett, D. Whiteson, and P. Baldi, Learning to isolate muons, J. High Energy Phys. 21 (2021) 200.
[66] J. Collado, J. N. Howard, T. Faucett, T. Tong, P. Baldi, and D. Whiteson, Learning to identify electrons, Phys. Rev. D **103**, 116028 (2021).
[67] T. Faucett, S.-C. Hsu, and D. Whiteson, Learning to identify semi-visible jets, J. High Energy Phys. 12 (2022) 132.
[68] P. Cal, J. Thaler, and W. J. Waalewijn, Power counting energy flow polynomials, J. High Energy Phys. 09 (2022) 021.
[69] R. Das, G. Kasieczka, and D. Shih, Feature selection with distance correlation, arXiv:2212.00046.
[70] A. Coccaro, M. Letizia, H. Reyes-Gonzalez, and R. Torre, On the curse of dimensionality for normalizing flows, arXiv:2302.12024.
[71] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, Neural ordinary differential equations, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Red Hook, New York, 2018), Vol. 31.
[72] J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. D. Hoffman, and R. A. Saurous, Tensorflow distributions, arXiv:1711.10604.
[73] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, FFJORD: free-form continuous dynamics for scalable reversible generative models, in *ICLR* (OpenReview.net, 2019).
[74] E. M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.
[75] A. Ng and M. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, in *NeurIPS*, edited by T. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, New York, 2001), Vol. 14.
[76] P. Kirichenko, P. Izmailov, and A. G. Wilson, Why normalizing flows fail to detect out-of-distribution data, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, New York, 2020), Vol. 33, pp. 20578–20589.
[77] L. Ardizzone, R. Mackowiak, C. Rother, and U. Köthe, Training normalizing flows with the information bottleneck for competitive generative classification, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, New York, 2020), Vol. 33, pp. 7828–7840.
[78] R. Mackowiak, L. Ardizzone, U. Kothe, and C. Rother, Generative classifiers as a basis for trustworthy image classification, in *CVPR* (IEEE Computer Society, Los Alamitos, CA, USA, 2021), pp. 2970–2980.
[79] R. Kansal, J. Duarte, H. Su, B. Orzari, T. Tomei, M. Pierini, M. Touranakou, J.-R. Vlimant, and D. Gunopulos, Particle cloud generation with message passing generative adversarial networks, arXiv:2106.11535.