

## Infrared-safe energy weighting does not guarantee small nonperturbative effects

Samuel Bright-Thonney<sup>1,\*</sup>, Benjamin Nachman<sup>2,3,†</sup> and Jesse Thaler<sup>4,5,‡</sup>

<sup>1</sup>*Physics Department, Cornell University, 109 Clark Hall, Ithaca, New York 14853, USA*

<sup>2</sup>*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

<sup>3</sup>*Berkeley Institute for Data Science, University of California, Berkeley, California 94720, USA*

<sup>4</sup>*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

<sup>5</sup>*The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, USA*



(Received 27 February 2024; accepted 30 May 2024; published 22 July 2024)

Infrared and collinear (IRC) safety has long been used a proxy for robustness when developing new jet substructure observables. This guiding philosophy has been carried into the deep learning era, where IRC-safe neural networks have been used for many jet studies. For graph-based neural networks, the most straightforward way to achieve IRC safety is to weight particle inputs by their energies. However, energy-weighting by itself does not guarantee that perturbative calculations of machine-learned observables will enjoy small nonperturbative corrections. In this paper, we demonstrate the sensitivity of IRC-safe networks to nonperturbative effects, by training an energy flow network (EFN) to maximize its sensitivity to hadronization. We then show how to construct Lipschitz energy flow networks ( $L$ -EFNs), which are both IRC safe and relatively insensitive to nonperturbative corrections. We demonstrate the performance of  $L$ -EFNs on generated samples of quark and gluon jets, and showcase fascinating differences between the learned latent representations of EFNs and  $L$ -EFNs.

DOI: [10.1103/PhysRevD.110.014029](https://doi.org/10.1103/PhysRevD.110.014029)

### I. INTRODUCTION

Infrared and collinear (IRC) safety has played a central role in perturbative quantum chromodynamics (QCD) [1–10], in particular for studying the properties of high-energy jets arising from the fragmentation of quarks and gluons. By being insensitive to soft and collinear splittings within a parton shower, an IRC-safe observable has a cross section that can be well-described by a fixed-order perturbation series in the strong coupling constant,  $\alpha_s$ . For this reason, IRC safety has been used as a guiding principle in developing new jet substructure observables [11–18]. While other classes of calculable observables exist—such as Sudakov-safe observables [19,20] and track-based

observables [21,22]<sup>1</sup>—IRC safety has driven the selection of collider observables for both measurements and searches involving jets.

Machine learning (ML) methods have significantly extended the sensitivity of jet tagging, in part by leveraging low-level information [15,25,26]. Such taggers are generally IRC unsafe, which means that they are reliant on the nonperturbative models in the generators used to train and test them. Since these nonperturbative models are imperfectly constrained by experimental data, this has motivated the development of ML observables that are IRC (or Sudakov) safe [27–29]. One class of IRC-safe architectures are energy flow networks (EFNs) [28], which are built on the deep sets framework [30] and are able to model any permutation-invariant IRC-safe observable. EFNs, along with extensions that incorporate graph-like structures [31–33], achieve IRC safety by linearly weighting particle inputs by their energies.

In this paper, we emphasize that energy-weighting, by itself, does not ensure trustworthy perturbative calculations for ML observables. While IRC safety does

\*[skb93@cornell.edu](mailto:skb93@cornell.edu)

†[bpnachman@lbl.gov](mailto:bpnachman@lbl.gov)

‡[jthaler@mit.edu](mailto:jthaler@mit.edu)

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP<sup>3</sup>.

<sup>1</sup>More generally, there is a broader class of observables that can be calculated after absorbing perturbative singularities into universal nonperturbative objects like fragmentation functions [23,24].

ensure the existence of a fixed-order perturbation series in  $\alpha_s$ ,<sup>2</sup> it does not guarantee that nonperturbative corrections will be small. Using EFNs as a representative example, we show how to train an IRC-safe neural network to be maximally sensitive to nonperturbative hadronization, thereby constructing an observable that is “safe but incalculable.”<sup>3</sup> As a step toward restoring calculability, we introduce Lipschitz energy flow networks ( $L$ -EFNs), whose bounded gradients ensure bounded sensitivity to nonperturbative corrections.

The fact that IRC-safe observables can have cross sections with large nonperturbative corrections is not new, even if it may not be widely appreciated. The standard (but misleading) lore is that IRC-safe observables should have nonperturbative corrections that are power-suppressed as  $(\Lambda_{\text{QCD}}/E)^n$ , where  $\Lambda_{\text{QCD}}$  is the QCD confinement scale,  $E$  is the energy scale of the process in consideration, and  $n$  is some integer power (typically 1 or 2). Already, though, it is known that jet angularities [34–37] with angular exponent  $\beta \lesssim 1$  have nonperturbative corrections with  $n = \beta$  scaling [8,19,19,38], which turns into  $\mathcal{O}(1)$  effects as  $\beta \rightarrow 0$ . Because there is no general first-principles understanding of nonperturbative QCD effects, then the cross section is essentially incalculable (or at least untrustable) if these corrections grow large.

In the context of IRC-safe ML models, we are not aware of any previous studies of the general impact of nonperturbative effects. Here, to identify ML observables with maximal nonperturbative sensitivity, we train an IRC-safe classifier to distinguish parton-level from hadron-level events. Classifiers whose cross sections have controlled nonperturbative corrections should be unable to distinguish between these samples. Instead, we find that EFNs are highly effective at parton-level versus hadron-level classification, implying large nonperturbative sensitivity. Our new  $L$ -EFN architecture reduces this sensitivity by imposing spectral normalization [39,40], which is equivalent to bounding the Lipschitz norm of the network (see related work in Refs. [41,42]). This approach is motivated by the Kantorovich-Rubinstein duality theorem [43] and the energy mover’s distance (EMD) [44], which provides a robust way to estimate the size of nonperturbative effects.

The remainder of this paper is organized as follows. In Sec. II method, we introduce  $L$ -EFNs and explain how the Lipschitz constraint enforces an EMD bound on nonperturbative corrections. We then perform a case study in Sec. III to compare the hadronization sensitivity of EFNs

and  $L$ -EFNs. We investigate the learned latent representations of ( $L$ -)EFNs in Sec. IV and conclude in Sec. V. For completeness, we perform a quark/gluon discrimination study in the Appendix.

## II. METHODOLOGIES

### A. Lipschitz energy flow networks

The  $L$ -EFN architecture we propose in this work is built on top of a standard EFN, which provides a generic framework for learning IRC-safe observables. Given a jet with constituent momenta  $p_1, p_2, \dots, p_M$ , an EFN computes a function of the form:

$$\text{EFN}(\{p_1, \dots, p_M\}) = F\left(\sum_{i=1}^M z_i \Phi(\hat{p}_i)\right), \quad (1)$$

where  $z_i = p_{T,i}/p_{T,\text{jet}}$  is the constituent momentum or energy fraction and  $\hat{p}_i$  is the particle’s angular position relative to the jet axis. The function  $\Phi: \mathbb{R}^2 \rightarrow \mathbb{R}^\ell$  maps individual particles to a latent space of dimension  $\ell$ . The function  $F: \mathbb{R}^\ell \rightarrow \mathbb{R}^{d_{\text{out}}}$  maps the latent representation to the final output. In a standard EFN, the functions  $\Phi$  and  $F$  are unconstrained and typically implemented as neural networks. The additive and energy-weighted structure of an EFN guarantees a naturally permutation-invariant and IRC safe output; see Ref. [28] for further discussion.

An  $L$ -EFN extends the EFN setup by constraining  $\Phi$  and  $F$  to be  $L$ -Lipschitz, meaning that

$$\begin{aligned} \|\Phi(\hat{p}_1) - \Phi(\hat{p}_2)\| &\leq L \|\hat{p}_1 - \hat{p}_2\|, \\ \|F(x_1) - F(x_2)\| &\leq L \|x_1 - x_2\|. \end{aligned}$$

This is effectively a bound on the gradients of these functions, though the Lipschitz constraint does not require  $\Phi$  and  $F$  to be everywhere differentiable. In principle, one could choose different  $L$  values for  $\Phi$  and  $F$ , but we keep them the same for simplicity of discussion.

If  $\Phi$  and  $F$  are neural networks with  $L$ -Lipschitz activations,<sup>4</sup> this amounts to a constraint on the *spectral norm* of their weight matrices  $W^i$  [39]:

$$\sigma(W^i) := \max_{\mathbf{h} \neq \mathbf{0}} \frac{\|W^i \mathbf{h}\|_2}{\|\mathbf{h}\|_2} \leq L. \quad (2)$$

This can be enforced during training by scaling the weight matrices as  $W^i \rightarrow LW^i/\sigma(W^i)$  using a computationally efficient estimation of  $\sigma(W^i)$  [39]. We focus on the  $L = 1$  case throughout this paper, and  $L$ -EFN should be henceforth understood as  $L = 1$ .

<sup>2</sup>Strictly speaking, this statement depends on the precise definition one uses for “IRC safety.” In the language of Ref. [10], we really mean Hölder continuous.

<sup>3</sup>The opposite case of “unsafe but calculable” observables can arise in resummed perturbation theory, where there is no order-by-order  $\alpha_s$  expansion but nevertheless nonperturbative corrections are suppressed [19,20].

<sup>4</sup>Many standard activation functions are 1-Lipschitz, such as ReLU, LeakyReLU, and Sigmoid.

For the studies in Sec. III B results we use the base architectures from the EnergyFlow package [27,28], implemented and trained using TensorFlow [45], Keras [46], and Adam [47]. To enforce the 1-Lipschitz constraint when training  $L$ -EFNs, we replace all linear Dense layers in the networks with SpectralDense layers from the DEEL-LIP package [48]. Unless otherwise specified, the  $\Phi$  ( $F$ ) networks of all ( $L$ -)EFNs are implemented as three-layer fully-connected neural networks with a width of 60 (80) and 0% (10%) dropout [49]. We use ReLU activations for internal nodes and Sigmoid for the 1D output node, both of which have a Lipschitz constant of 1. The networks are trained to learn 0/1 class labels based on the binary cross-entropy loss function.

As a reference, we also show results for a particle flow network (PFN), which is IRC unsafe by construction:

$$\text{PFN}(\{p_1, \dots, p_M\}) = F\left(\sum_{i=1}^M \Phi(p_i)\right). \quad (3)$$

This differs from the EFN in Eq. (1) by the lack of  $z_i$  weighting and the use of the unnormalized particle momentum  $p_i$  instead of just the particle direction  $\hat{p}_i$ . We also consider a Lipschitz PFN ( $L$ -PFN) which imposes a spectral norm constraint on  $F$  and  $\Phi$ .

### B. Bounds on the energy mover's distance

With the simple addition of a Lipschitz constraint, we can now make precise statements about the sensitivity of an  $L$ -EFN to nonperturbative effects. Here, we leverage a recent geometric language for analyzing collision events based on the EMD [44], which measures the “work” required to transform one jet into another. The EMD is a variant of the Earth Mover’s Distance from computer vision [50–54] and is equivalent to the 1-Wasserstein metric [55,56] in certain limits. The EMD provides a pairwise metric distance for IRC-safe energy flows, which can be used to triangulate the space of jets and define various geometric quantities.

Consider an additive IRC-safe observable  $f$  acting on a jet  $\mathcal{J}$ :

$$f(\mathcal{J}) = \sum_{i \in \mathcal{J}} E_i \Phi(\hat{p}_i), \quad (4)$$

where  $\Phi$  has Lipschitz constant  $L$ . Given a pair of jets  $\mathcal{J}_1$ ,  $\mathcal{J}_2$ , a key result from Ref. [44] places a bound on the difference between  $f(\mathcal{J}_1)$  and  $f(\mathcal{J}_2)$  based on the EMD between  $\mathcal{J}_1$  and  $\mathcal{J}_2$ :

$$\frac{1}{RL} |f(\mathcal{J}_1) - f(\mathcal{J}_2)| \leq \text{EMD}(\mathcal{J}_1, \mathcal{J}_2), \quad (5)$$

where  $R$  is the jet radius. The composition of two Lipschitz functions with constants  $L_A$  and  $L_B$  is also Lipschitz, with

constant  $L_A L_B$ . Thus, for an  $L$ -EFN, we have:

$$\frac{1}{RL^2} |L\text{-EFN}(\mathcal{J}_1) - L\text{-EFN}(\mathcal{J}_2)| \leq \text{EMD}(\mathcal{J}_1, \mathcal{J}_2). \quad (6)$$

This bound holds for any  $\mathcal{J}_1$  and  $\mathcal{J}_2$ , and we again emphasize that we focus on  $L = 1$  in this paper.

Now, consider  $\mathcal{J}_1$  and  $\mathcal{J}_2$  to be the *same jet* at parton- and hadron-level, respectively. Of course, parton-level information cannot be accessed at a real collider, but it can be done in the context of a parton shower event generator. Then, Eq. (6) gives an upper bound on modifications of the  $L$ -EFN observable due to hadronization. This prevents an  $L$ -EFN from becoming overly sensitive to nonperturbative physics, and makes it naturally robust to any unphysical mismodeling artifacts present in Monte Carlo simulations.

By contrast, a plain EFN does not have this constraint, which explain why it can be arbitrarily sensitive to non-perturbative effects, as we will see in the following case study. Note that this sensitive potentially affects all past studies based on EFNs. In the Appendix, we show that quark/gluon discrimination power degrades substantially after imposing a Lipschitz constraint, implying that some of the discrimination power seen in previous EFN studies [28] can be attributed to nonperturbative effects.

## III. CASE STUDY WITH QUARK/GLUON JETS

### A. Generated dataset

To demonstrate the impact of nonperturbative physics on ML algorithms, we perform a case study involving quark and gluon jets. Following Ref. [57], we generate samples of  $e^+e^- \rightarrow H \rightarrow q\bar{q}$  (quark jets) and  $gg$  (gluon jets) using PYTHIA 8.307. We take a center-of-mass collision energy of  $\sqrt{s} = m_H = 1$  TeV, and we fix  $q = u$  (i.e. up quarks only) and  $m_u = 0$  for simplicity.

In each event, we independently cluster the partons (postshower) and final-state hadrons into anti- $k_T$  jets with  $R = 1.0$  [58,59] and the winner-take-all (WTA) axis. We match the leading (highest  $p_T$ ) hadron-level jet to the nearest parton-level jet within  $\Delta R < 0.1$ ,<sup>5</sup> and save the constituent kinematic information for both jets. We generate a total of 200,000  $e^+e^- \rightarrow q\bar{q}$  and 200,000  $e^+e^- \rightarrow gg$  events. Each dataset is split into training, validation, and test sets with fractions 50%, 25%, and 25%, respectively.

In Fig. 1, we show the spectrum of three jet substructure observables: the jet mass, the number of jet constituents, and  $p_T^D$  [60]. The latter observable is the sum of the squares of the momentum fractions of constituents within the jet. Of these observables, only the jet mass is IRC safe. Not surprisingly, the jet mass is the observable with the smallest

<sup>5</sup>If no parton-level jet is found within this  $\Delta R$  cone, the event is skipped.

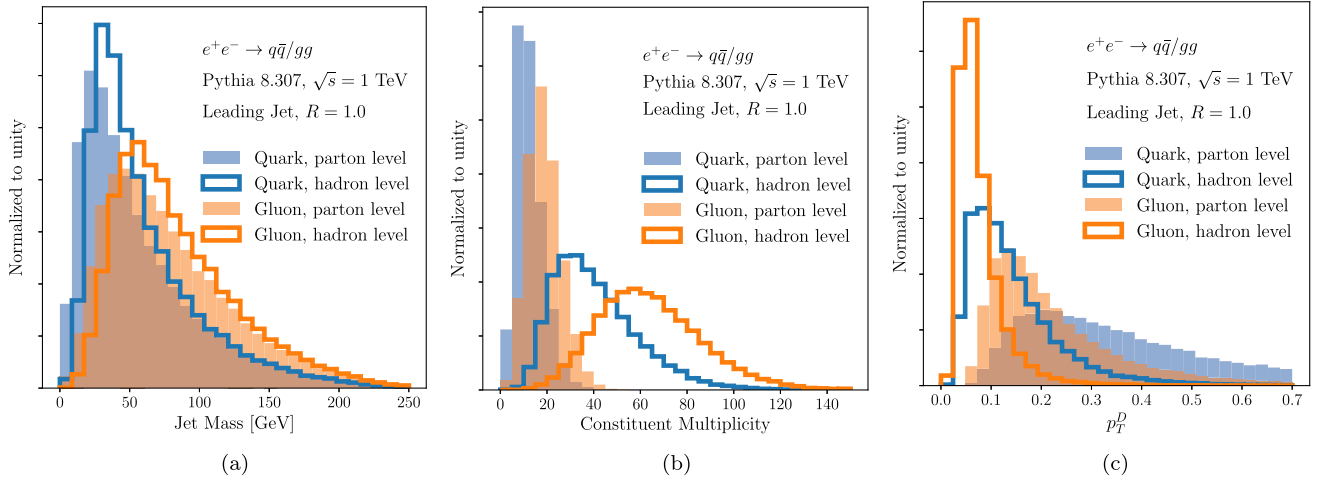


FIG. 1. The spectrum of (a) jet mass, (b) jet constituent multiplicity, and (c)  $p_T^D$ . Shown are the normalized distributions for quarks (blue) and gluons (orange) at parton-level (filled) versus hadron-level (solid).

difference between the parton-level and hadron-level distributions. For the number of constituents and  $p_T^D$ , nonperturbative corrections are an  $\mathcal{O}(1)$  effect. Nevertheless, these two observables are highly effective for quark versus gluon jet tagging [61], which is why one has to be careful about nonperturbative sensitivity in the ML context.

## B. Classification results

To investigate the impact of Lipschitz constraints on nonperturbative sensitivity, we train EFNs and  $L$ -EFNs to distinguish hadron-level from parton-level jets. For this analysis, we treat the quark and gluon datasets separately. This allows us to focus solely on an ( $L$ -)EFN's sensitivity to parton-level versus hadron-level differences rather than

any of the features distinguishing quark and gluon jets. Specifically, gluon jets tend to be less collimated and have larger constituent multiplicities due to the larger color factors for gluons ( $C_A = 3$ ) compared to quarks ( $C_F = 4/3$ ), but hadronization also tends to broaden jets and increase their number of constituents. (See the Appendix for a quark/gluon discrimination study.)

To verify the stability of our trainings and estimate uncertainties on performance metrics, we train ten versions of each model with a different random initialization. In testing, we found that  $L$ -EFNs train more stably with a significantly larger batch size than EFNs (10,000 vs 128) and a smaller learning rate ( $10^{-4}$  vs  $10^{-3}$ ). This is likely due to the spectral normalization of the weight matrices at each

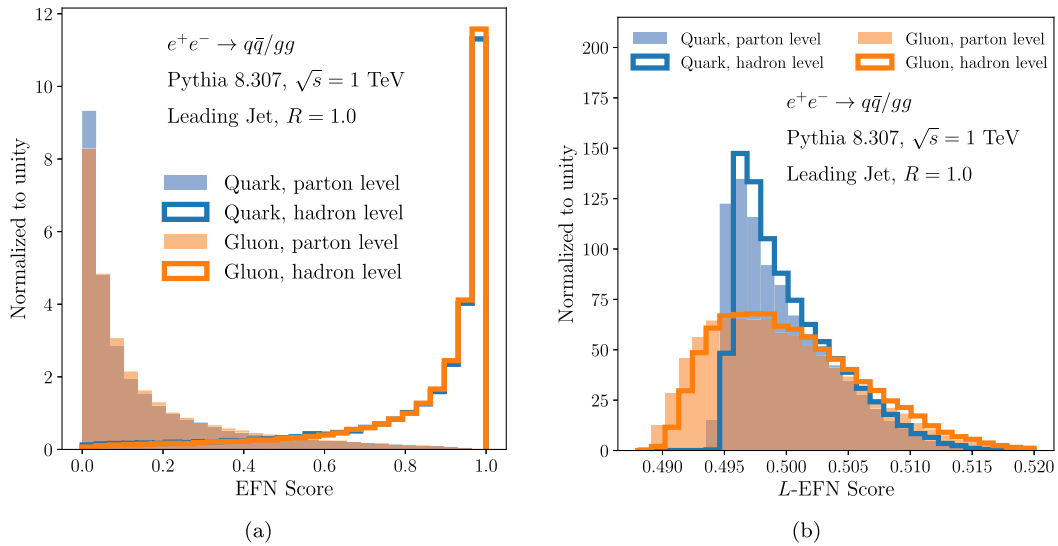


FIG. 2. Score distributions for the (a) EFN and (b)  $L$ -EFN architectures, when trained to distinguish parton-level from hadron-level jets. The color scheme is the same as Fig. 1. Despite being IRC safe, the EFN achieves excellent parton/hadron separation power. After imposing the Lipschitz constraint, the  $L$ -EFN exhibits the desired insensitivity to nonperturbative effects.

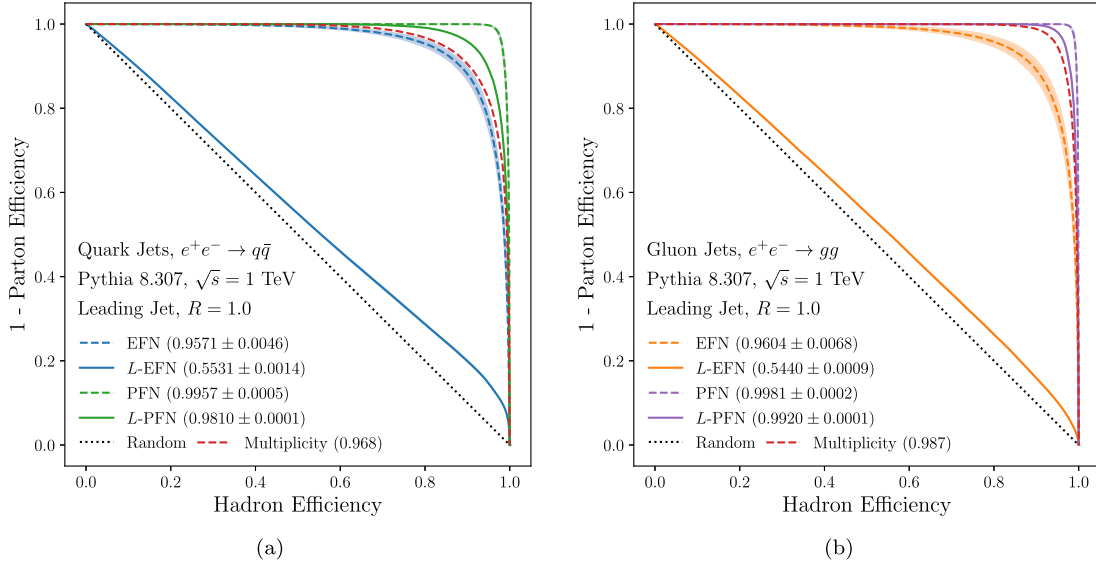


FIG. 3. ROC curves for parton- vs hadron-level discrimination for (a) quark and (b) gluon jets using an EFN/PFN (solid) and an  $L$ -EFN/ $L$ -PFN (dashed). The bands represent the spread from starting with 10 different random initializations of the neural networks. The results for constituent multiplicity is shown for reference. Only the  $L$ -EFN shows the desired lack of sensitivity to nonperturbative effects.

step, and the inherent difficulty of distinguishing parton-level from hadron-level jets given the constraint of Eq. (6). With these modifications, however, all  $L$ -EFN trainings converged easily.

Distributions of the learned EFN and  $L$ -EFN scores are shown in Fig. 2, considering all four combinations of quark versus gluon and parton-level versus hadron-level jets. Despite being IRC safe, the EFN is highly effective at distinguishing parton-level from hadron-level jets. This implies that the nonperturbative corrections to the EFN observable are large. Assuming that the EFN architecture and training procedure are sufficiently flexible,<sup>6</sup> then this is an example of an IRC-safe observable with maximal nonperturbative sensitivity. By contrast, the  $L$ -EFN is not very effective at distinguishing parton-level from hadron-level jet, so according to the hadronization model in PYTHIA, the nonperturbative corrections are small.

To better assess the learned separation power, we plot receiver operating characteristic (ROC) curves in Fig. 3, obtained by placing a sliding cut over the distributions in Fig. 2. We also indicate their area under the curve (AUC) scores in the legend. The ROC curves underscore and quantify the trends observed in Fig. 2. The EFN is an extremely efficient parton-versus-hadron classifier, whereas the  $L$ -EFN performs barely better than random chance. We therefore conclude that the 1-Lipschitz modification is effective at suppressing the  $L$ -EFN's sensitivity to nonperturbative effects.

<sup>6</sup>We did not perform an extensive hyperparameter scan, but we found little gains from small variations on our setup.

As a point of comparison, Fig. 3 also shows ROCs and AUCs for multiplicity, PFNs and  $L$ -PFNs—all explicitly IRC-unsafe observables. Unsurprisingly, the PFN is able to nearly perfectly distinguish parton from hadron jets, and the  $L$ -PFN performs only slightly worse. This indicates that the 1-Lipschitz constraint does not severely limit the expressiveness of the model, and that the large gap between the EFN and  $L$ -EFN is a genuine result of nonperturbative effects.

### C. Confirming the EMD bound

As a cross-check of our analysis, we verify that the learned  $L$ -EFN satisfies the EMD bound in Eq. (6). The plots in Fig. 4 compare the difference between observables at parton level versus hadron level to the EMD between parton- and hadron-level events.

The top row of Fig. 4 reproduces Fig. 2 of Ref. [44] for the  $\beta = 1$  angularity:

$$\lambda^{(1)} = \sum_{i \in \text{jet}} z_i \Delta R_{i,\text{jet}}, \quad (7)$$

where  $z_i = p_{T,i}/p_{T,\text{jet}}$  and  $\Delta R_{i,\text{jet}}^2 = \Delta \eta_{i,\text{jet}}^2 + \Delta \phi_{i,\text{jet}}^2$ . This function is 1-Lipschitz and the bound is respected and nearly saturated.

The middle column of Fig. 4 shows the nonperturbative modification to the EFN score, which clearly violates the theoretical bound. This is expected, since the EFN is not required to be a Lipschitz function. Indeed, there is no finite slope (i.e. Lipschitz constant) that would bound the EFN score relative to the EMD.

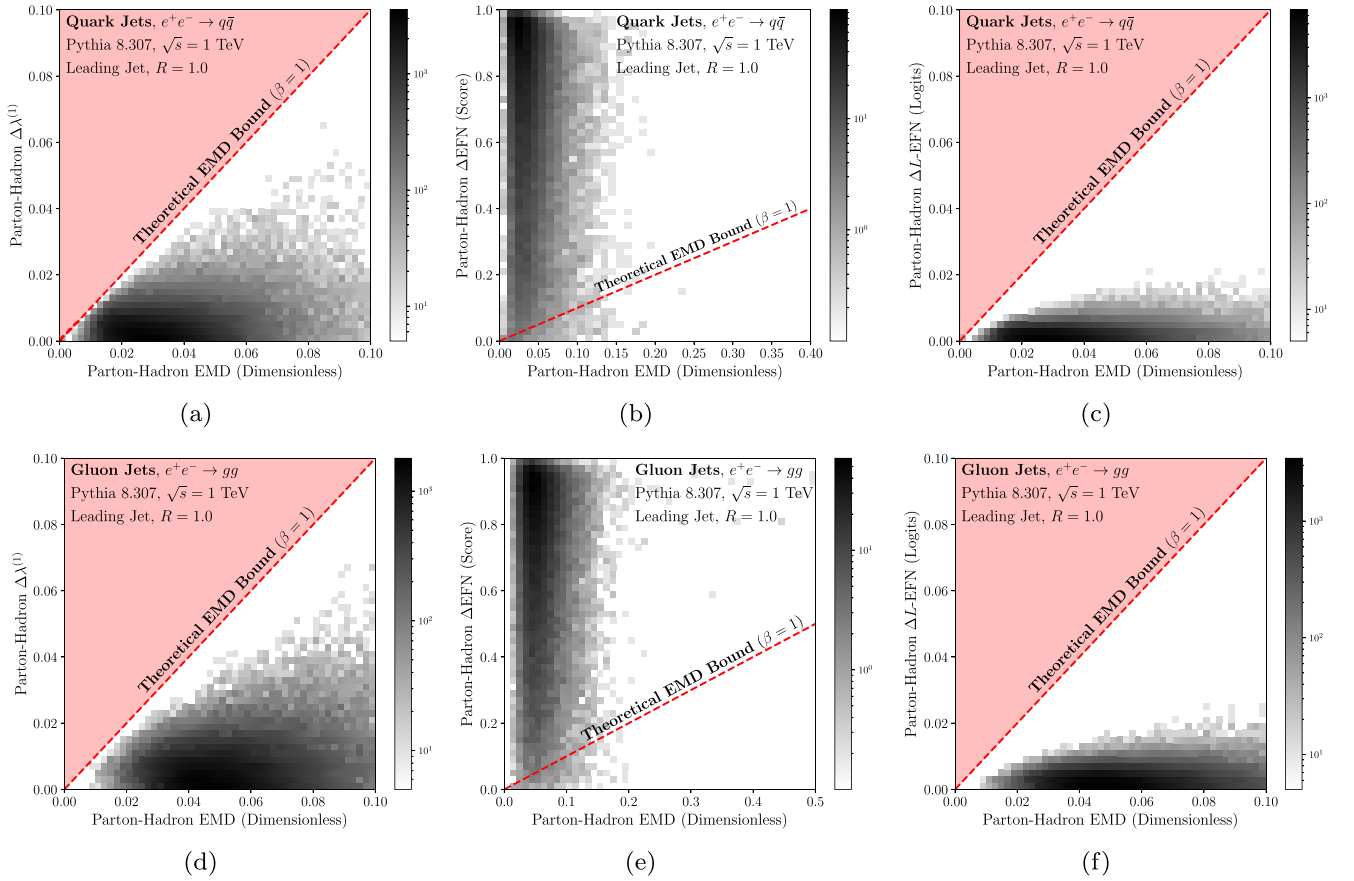


FIG. 4. Two dimensional distributions of the parton-hadron EMD versus the difference in observables at parton- and hadron-level. Shown are results for (a, d) the  $\beta = 1$  angularity  $\lambda^{(1)}$ , (b, e) an EFN, (c, f) an  $L$ -EFN, separately for (a–c) quark jets and (d–f) gluon jets. The EMD is computed using momentum fractions  $z_i$  in keeping with the  $\lambda^{(1)}$ /EFN/ $L$ -EFN calculations, and is thus dimensionless.

By contrast, the bottom row of Fig. 4 demonstrates that the  $L$ -EFN does not respect the EMD bound. For illustrative purposes, we show the  $L$ -EFN logits (i.e. pre-activations on the final layer), since the actual output scores all lie along the bottom of the plot. Even after performing the logit transformation, the  $L$ -EFN does not appear to saturate the bound, unlike what was seen for the IRC-safe angularities.

#### IV. VISUALIZING NONPERTURBATIVE EFFECTS

The above case study demonstrates that IRC safety is insufficient to protect against large nonperturbative modifications, but it gives no immediate insight into which observables are most susceptible. In this section, we take a step toward answering this question by training EFNs with a minimal latent dimension size of  $\ell = 1$ . This corresponds to a single learned filter  $\Phi(\eta, \phi)$  which can be easily visualized, and a simple one-dimensional classification function  $F$ .

As in Sec. III B, we train  $\ell = 1$  EFNs and  $L$ -EFNs to discriminate parton- and hadron-level jets, treating the quark and gluon samples separately. Perhaps surprisingly, the  $\ell = 1$  EFNs are still able to achieve AUCs of 0.90

(quarks) and 0.91 (gluons), compared to 0.96 for the full  $\ell = 60$  models in Sec. III B. The  $\ell = 1$   $L$ -EFNs perform similarly well, with AUCs of 0.54 and 0.53, compared to 0.55 and 0.54 previously. We therefore conclude that an  $\ell = 1$  analysis is sufficient to capture the leading sources of nonperturbative sensitivity.

In Fig. 5, we plot the filter profiles  $\Phi(\eta, \phi)$  for the  $\ell = 1$  ( $L$ -)EFNs (left and middle columns) alongside their angle-averaged radial profiles (right column). All filters exhibit cylindrical symmetry, as expected since the generated samples are for unpolarized jets with no preferred angular orientation. Interestingly, the EFN filters appear to be maximally sensitive to *highly collinear* radiation. Despite the jet radius being  $R = 1.0$ , the EFN filters peak within a radius of  $\sim 0.02$ , with a plateau going out to larger scales. The EFN filters also vanish at the origin, reflecting the consistent presence of a particle at  $R = 0$  aligned with the WTA jet axis.<sup>7</sup>

<sup>7</sup>If angles are instead measured with respect to the standard jet momentum axis, then the  $\ell = 1$  EFN filters often exhibit anisotropic features.

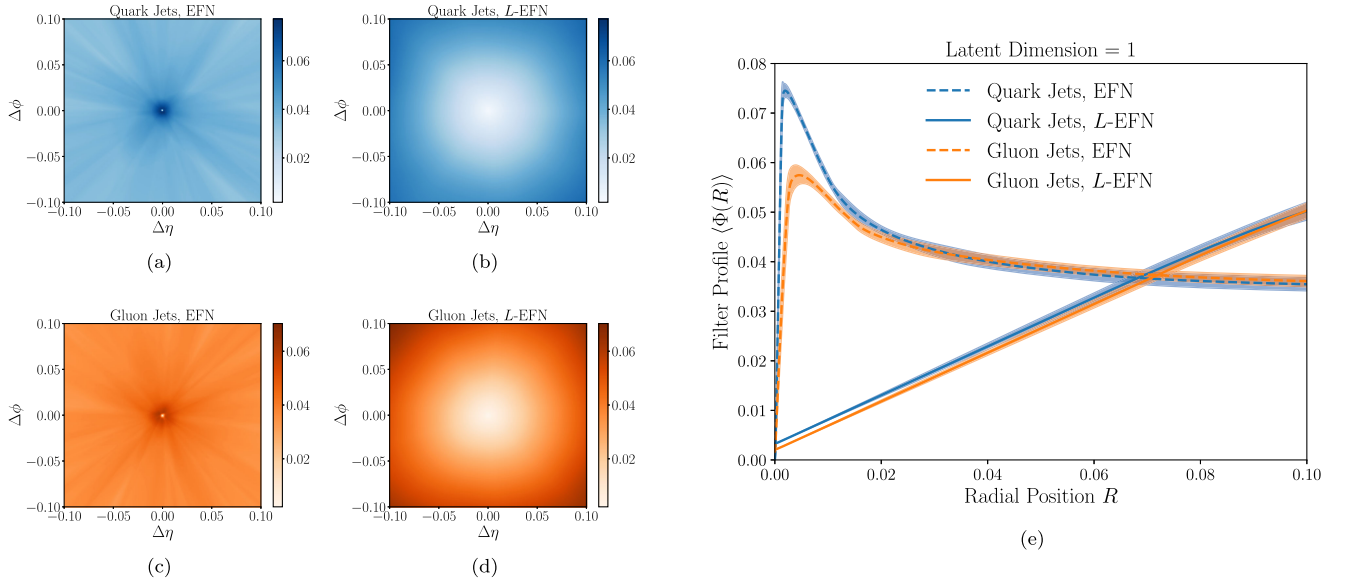


FIG. 5. Angular filters  $\Phi(\eta, \phi)$  for parton-level versus hadron-level discrimination with a latent space of dimension  $\ell = 1$ . The EFN filters are shown for (a) quark and (c) gluon jets, and similarly for the  $L$ -EFN filters for (b) quark and (d) gluon jets. The radial profiles  $\Phi(R = \sqrt{\eta^2 + \phi^2})$  are shown in (e), demonstrating the EFN sensitivity to highly collinear emissions ( $R \lesssim 0.02$ ). While the jet radius is  $R = 1.0$ , we have zoomed in to show the collinear structure of the filters.

The  $L$ -EFN filters, on the other hand, cannot vary so rapidly near the origin due to the 1-Lipschitz constraint. Thus, the  $L$ -EFN cannot be as sensitive to collinear radiation, while simultaneously ignoring the central particle.

This differing behavior of EFNs and  $L$ -EFN is consistent with theoretical expectations about the nonperturbative sensitivity of the angularities:

$$\lambda^{(\beta)} = \sum_{i \in \text{jet}} z_i \Delta R_{i, \text{jet}}^\beta, \quad (8)$$

which has Lipschitz constant  $L = \beta$ . When  $\beta \gtrsim 1$ , nonperturbative corrections are suppressed by  $\Lambda_{\text{QCD}}/E_{\text{jet}}$ , so they are relatively small and independent of  $\beta$  for high energy jets, in agreement with the  $L$ -EFN behavior.

For  $\beta \lesssim 1$ , though, nonperturbative corrections scale as  $(\Lambda_{\text{QCD}}/E_{\text{jet}})^\beta$ . This is unsuppressed as  $\beta \rightarrow 0$ , which yields large differences between parton-level and hadron-level distributions. Note that in the  $\beta \rightarrow 0$  limit, the angularities are dominated by collinear emissions. This explains why

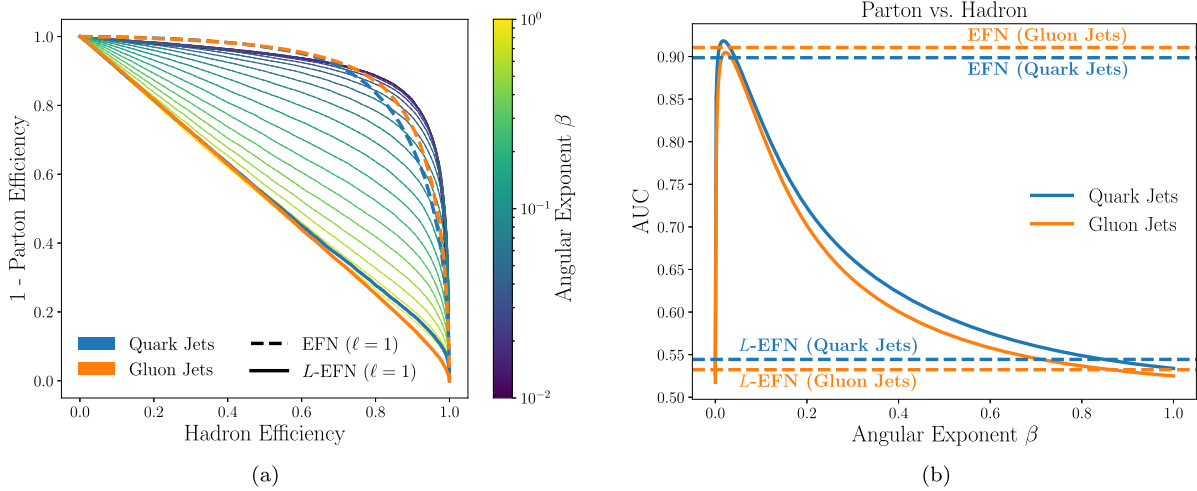


FIG. 6. Parton versus hadron discriminating power for the angularities  $\lambda^{(\beta)}$  from Eq. (8) as a function  $\beta$ . Shown are (a) ROC curves and (b) AUC scores. We also show ROC curves for the  $\ell = 1$  EFNs and  $L$ -EFNs, which agree well with the  $\beta \ll 1$  and  $\beta \approx 1$  limits. For simplicity, in (a) we only show quark ROC curves for the  $\beta$  scan, though the gluon results are qualitatively similar.

the IRC-safe EFN with no Lipschitz constraint can have large nonperturbative sensitivity coming from small-angle physics.

In Fig. 6, we quantify the parton versus hadron discrimination power of the angularities as a function of  $\beta$ . ROC curves are shown for the EFN and  $L$ -EFNs for comparison, and we see a rough interpolation where  $\beta \rightarrow 0$  maps to the EFN, while  $\beta \approx 1$  maps to the  $L$ -EFN. The AUCs monotonically increase as  $\beta$  decreases, peaking around  $\beta = 0.02$  before losing sensitivity as  $\beta \rightarrow 0$ .

## V. CONCLUSIONS AND OUTLOOK

In this paper, we demonstrated the susceptibility of IRC-safe neural networks to large, incalculable nonperturbative corrections. We showed how to make the network maximally sensitive to these corrections by training it to distinguish between parton-level and hadron-level jets. In general, there is nothing constraining this effect in a typical application. Although IRC safety guarantees that the network's output is perturbatively calculable *in principle*, this is insufficient for robustness to large and potentially mismodeled nonperturbative corrections.

To mitigate this issue, we introduced a new neural network architecture, the  $L$ -EFN, which has significantly reduced sensitivity to nonperturbative physics. We achieve this by constraining the networks to be  $L$ -Lipschitz functions, which bounds potential nonperturbative modifications according to the EMD between parton- and hadron-level jets. Using PYTHIA's hadronization model as a proxy for nonperturbative effects, we demonstrated that an  $L$ -EFN is barely able to distinguish between the same jet at parton- and hadron-level. In the future, it would be interesting to explore different hadronization models and other nonperturbative effects like multiparton interactions (i.e. underlying event). It would also be interesting to study the robustness of  $L$ -EFNs to perturbative uncertainties in parton showers.

The ideas behind  $L$ -EFNs could be extended to other architectures, such as graph neural networks and transformers, though it may be more complicated to derive theoretical bounds similar to Eq. (6). It would also be interesting to benchmark  $L$ -EFNs and  $L$ -EFN-inspired architectures on standard jet tagging tasks and assess the tradeoff between performance and robustness to hadronization. In the case of  $L$ -EFNs, the Lipschitz constant provides an easily tunable network constraint that balances these considerations. The non-Lipschitz networks trained to distinguish parton and hadron jets may also be useful to explore the impact of hadronization where it is predicted to be the largest. For non-IRC-safe networks, our studies did not reveal dramatic

differences between PFNs and  $L$ -PFNs, but the flexibility to adjust  $L$  may be beneficial for ablation studies.

With a growing trend to build robust and interpretable neural networks that are structured with particle physics in mind, it is important to be guided by both formal and practical considerations. Finding a balance between neat theoretical constraints and performance demands is inevitably delicate, but  $L$ -EFNs present a concrete step toward uniting the power of deep learning with a reasoned caution against learning spurious or unphysical details of a simulation. We look forward to continued investigation and development in future work, and hope to see the ongoing development of new machine learning tools alongside theoretical tools to better understand them.

## ACKNOWLEDGMENTS

We thank Rikab Gambhir for feedback on the manuscript and discussions of the  $\beta \rightarrow 0$  limit of angularities. We thank Eric Metodiev and Andrew Larkoski for additional helpful comments. S. B. T. is supported by the National Science Foundation under Award No. PHY-2209443. B. N. is supported by the U.S. Department of Energy (DOE), Office of Science under Contract DE-AC02-05CH11231. J. T. is supported by the DOE Office of High Energy Physics under Grant Contract No. DE-SC0012567, by the National Science Foundation under Cooperative Agreement No. PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, [62]), and by the Simons Foundation through Investigator Grant No. 929241. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award HEP-ERCAP0021099.

## APPENDIX: QUARK VS GLUON JET DISCRIMINATION WITH $L$ -EFNS

In Secs. III and IV, we contrasted EFNs and  $L$ -EFNs for the relatively contrived scenario of distinguishing parton- from hadron-level jets. While this demonstrated our point about the large potential impact of nonperturbative physics on EFN observables, it does not reflect a realistic use case of ( $L$ -)EFNs. In this appendix, we compare EFNs and  $L$ -EFNs on a canonical collider physics task: discriminating quark and gluon jets.

Using the datasets described in Sec. III A, we train different ( $L$ -)EFNs and ( $L$ -)PFNs to discriminate quark jets from gluon jets. We perform separate trainings on hadron-level and parton-level inputs, but always evaluate the performance on

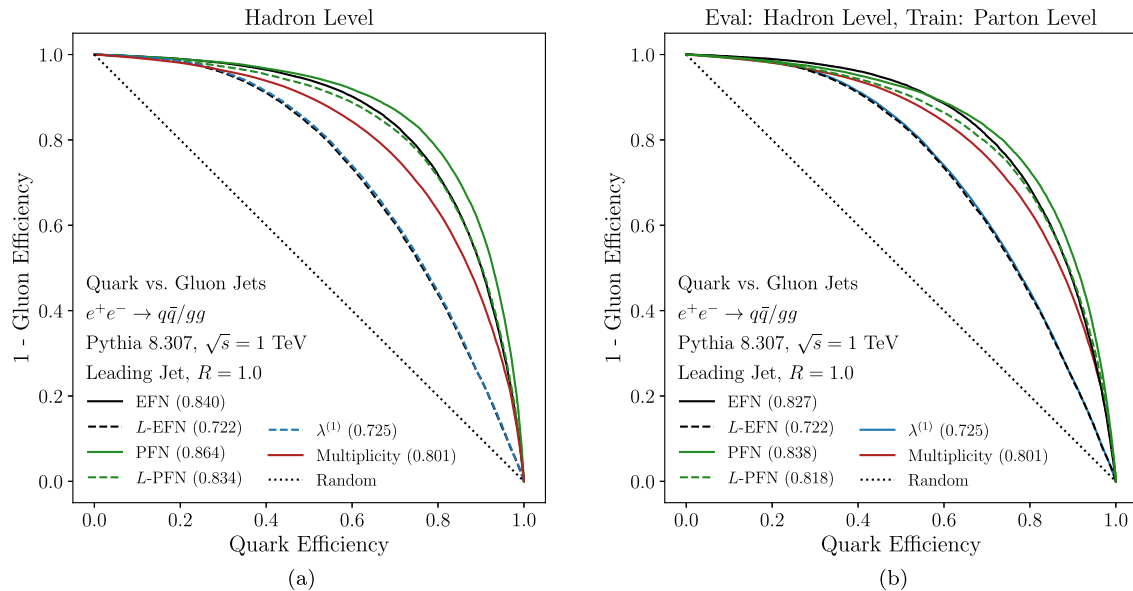


FIG. 7. ROC curves for hadron-level quark/gluon jet discrimination using an EFN/PFN (solid) and an  $L$ -EFN/ $L$ -PFN (dashed), trained using (a) hadron-level and (b) parton-level inputs. The results for constituent multiplicity and the angularity  $\lambda^{(1)}$  are shown for reference. While the  $L$ -EFN underperforms with respect to the EFN, it is minimally affected by training at parton-level and evaluating at hadron-level. The EFN loses a small amount of discriminating power, and the PFN/ $L$ -PFN lose even more.

hadron-level jets. In Fig. 7, we show ROC curves for quark/gluon jet discrimination using these ( $L$ )-EFNs/PFNs. As expected, the default  $L$ -EFN with  $L = 1$  performs exactly as well as the angularity  $\lambda^{(1)}$ , both of which underperform

relative to the unconstrained EFN and both PFNs. Notably, the  $L$ -EFN performance is nearly identical whether it is trained on hadron- or parton-level inputs, highlighting the expected robustness to nonperturbative effects.

- 
- [1] T. Kinoshita, Mass singularities of Feynman amplitudes, *J. Math. Phys. (N.Y.)* **3**, 650 (1962).
  - [2] T. D. Lee and M. Nauenberg, Degenerate systems and mass singularities, *Phys. Rev.* **133**, B1549 (1964).
  - [3] G. F. Sterman and S. Weinberg, Jets from quantum chromodynamics, *Phys. Rev. Lett.* **39**, 1436 (1977).
  - [4] G. F. Sterman, Mass divergences in annihilation processes. 1. Origin and nature of divergences in cut vacuum polarization diagrams, *Phys. Rev. D* **17**, 2773 (1978).
  - [5] G. F. Sterman, Mass divergences in annihilation processes. 2. Cancellation of divergences in cut vacuum polarization diagrams, *Phys. Rev. D* **17**, 2789 (1978).
  - [6] R. Brock *et al.* (CTEQ Collaboration), Handbook of perturbative QCD: Version 1.0, *Rev. Mod. Phys.* **67**, 157 (1995).
  - [7] S. Weinberg, *The Quantum Theory of Fields. Vol. 1: Foundations* (Cambridge University Press, Cambridge, England, 2005).
  - [8] A. Banfi, G. P. Salam, and G. Zanderighi, Principles of general final-state resummation and automated implementation, *J. High Energy Phys.* **03** (2005) 073.
  - [9] G. F. Sterman, Review of theoretical status: The long and short of high energy jets, eConf C **0601121**, R001 (2006).
  - [10] P. T. Komiske, E. M. Metodiev, and J. Thaler, The hidden geometry of particle collisions, *J. High Energy Phys.* **07** (2020) 006.
  - [11] A. Abdesselam *et al.*, Boosted objects: A probe of beyond the standard model physics, *Eur. Phys. J. C* **71**, 1661 (2011).
  - [12] A. Altheimer *et al.*, Jet substructure at the Tevatron and LHC: New results, new tools, new benchmarks, *J. Phys. G* **39**, 063001 (2012).
  - [13] A. Altheimer *et al.*, Boosted objects and jet substructure at the LHC. Report of BOOST2012, held at IFIC Valencia, 23rd–27th of July 2012, *Eur. Phys. J. C* **74**, 2792 (2014).
  - [14] D. Adams *et al.*, Towards an understanding of the correlations in jet substructure, *Eur. Phys. J. C* **75**, 409 (2015).
  - [15] A. J. Larkoski, I. Moult, and B. Nachman, Jet substructure at the Large Hadron Collider: A review of recent advances in theory and machine learning, *Phys. Rep.* **841**, 1 (2020).
  - [16] R. Kogler *et al.*, Jet substructure at the Large Hadron Collider: Experimental review, *Rev. Mod. Phys.* **91**, 045003 (2019).
  - [17] S. Marzani, G. Soyez, and M. Spannowsky, *Looking Inside Jets: An Introduction to Jet Substructure and Boosted-Object Phenomenology*, Lecture Notes in Physics Vol. 958 (Springer, Cham, 2019).

- [18] R. Kogler, *Advances in Jet Substructure at the LHC: Algorithms, Measurements and Searches for New Physical Phenomena*, Springer Tracts in Modern Physics Vol. 284 (Springer, Cham, 2021).
- [19] A. J. Larkoski and J. Thaler, Unsafe but calculable: Ratios of angularities in perturbative QCD, *J. High Energy Phys.* **09** (2013) 137.
- [20] A. J. Larkoski, S. Marzani, and J. Thaler, Sudakov safety in perturbative QCD, *Phys. Rev. D* **91**, 111501 (2015).
- [21] H.-M. Chang, M. Procura, J. Thaler, and W. J. Waalewijn, Calculating track-based observables for the LHC, *Phys. Rev. Lett.* **111**, 102002 (2013).
- [22] Y. Li, I. Moulst, S. S. van Velzen, W. J. Waalewijn, and H. X. Zhu, Extending precision perturbative QCD with track functions, *Phys. Rev. Lett.* **128**, 182001 (2022).
- [23] G. Altarelli, Partons in quantum chromodynamics, *Phys. Rep.* **81**, 1 (1982).
- [24] B. T. Elder, M. Procura, J. Thaler, W. J. Waalewijn, and K. Zhou, Generalized fragmentation functions for fractal jet observables, *J. High Energy Phys.* **06** (2017) 085.
- [25] A. Butter *et al.*, The machine learning landscape of top taggers, *SciPost Phys.* **7**, 014 (2019).
- [26] M. Feickert and B. Nachman, A living review of machine learning for particle physics, [arXiv:2102.02770](https://arxiv.org/abs/2102.02770).
- [27] 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.
- [28] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, *J. High Energy Phys.* **01** (2019) 121.
- [29] D. Athanasakos, A. J. Larkoski, J. Mulligan, M. Płoskoń, and F. Ringer, Is infrared-collinear safe information all you need for jet classification?, [arXiv:2305.08979](https://arxiv.org/abs/2305.08979).
- [30] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. R. Salakhutdinov, and A. J. Smola, Deep sets, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30.
- [31] P. Konar, V. S. Ngairangbam, and M. Spannowsky, Energy-weighted message passing: An infra-red and collinear safe graph neural network algorithm, *J. High Energy Phys.* **02** (2022) 060.
- [32] W. Shen, D. Wang, and J. M. Yang, Hierarchical high-point energy flow network for jet tagging, *J. High Energy Phys.* **09** (2023) 135.
- [33] P. Konar, V. S. Ngairangbam, and M. Spannowsky, Hypergraphs in LHC phenomenology—The next frontier of IRC-safe feature extraction, *J. High Energy Phys.* **01** (2024) 113.
- [34] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, and C. Lee, Jet shapes and jet algorithms in SCET, *J. High Energy Phys.* **11** (2010) 101.
- [35] A. J. Larkoski, D. Neill, and J. Thaler, Jet shapes with the broadening axis, *J. High Energy Phys.* **04** (2014) 017.
- [36] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung, and J. Virzi, Substructure of high- $p_T$  jets at the LHC, *Phys. Rev. D* **79**, 074017 (2009).
- [37] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, Gaining (mutual) information about quark/gluon discrimination, *J. High Energy Phys.* **11** (2014) 129.
- [38] A. V. Manohar and M. B. Wise, Power suppressed corrections to hadronic event shapes, *Phys. Lett. B* **344**, 407 (1995).
- [39] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, Spectral normalization for generative adversarial networks, [arXiv:1802.05957](https://arxiv.org/abs/1802.05957).
- [40] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree, Regularisation of neural networks by enforcing Lipschitz continuity, [arXiv:1804.04368](https://arxiv.org/abs/1804.04368).
- [41] C. Anil, J. Lucas, and R. Grosse, Sorting out Lipschitz function approximation, [arXiv:1811.05381](https://arxiv.org/abs/1811.05381).
- [42] O. Kitouni, N. Nolte, and M. Williams, Robust and provably monotonic networks, *Mach. Learn. Sci. Tech.* **4**, 035020 (2023).
- [43] L. Kantorovich and G. S. Rubinstein, On a space of totally additive functions, *Vestn. Leningr. Univ. Fiz. Khim.* **13**, 52 (1958).
- [44] P. T. Komiske, E. M. Metodiev, and J. Thaler, Metric space of collider events, *Phys. Rev. Lett.* **123**, 041801 (2019).
- [45] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, TensorFlow: A system for large-scale machine learning., in *OSDI (USENIX, 2016)*, Vol. 16, pp. 265–283.
- [46] F. Chollet, Keras, <https://github.com/fchollet/keras> (2015).
- [47] D. Kingma and J. Ba, Adam: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [48] M. Serrurier, F. Mamalet, A. González-Sanz, T. Boissin, J.-M. Loubes, and E. del Barrio, Achieving robustness in classification using optimal transport with hinge regularization, [arXiv:2006.06520](https://arxiv.org/abs/2006.06520).
- [49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* **15**, 1929 (2014).
- [50] S. Peleg, M. Werman, and H. Rom, A unified approach to the change of resolution: Space and gray-level, *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 739 (1989).
- [51] Y. Rubner, C. Tomasi, and L. J. Guibas, A metric for distributions with applications to image databases, in *Proceedings of the Sixth International Conference on Computer Vision, ICCV '98* (IEEE Computer Society, Washington, DC, USA, 1998), pp. 59–66.
- [52] Y. Rubner, C. Tomasi, and L. J. Guibas, The Earth mover's distance as a metric for image retrieval, *Int. J. Comput. Vis.* **40**, 99 (2000).
- [53] O. Pele and M. Werman, A linear time histogram metric for improved SIFT matching, in *Computer Vision—ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, 2008, Proceedings, Part III*, edited by D. A. Forsyth, P. H. S. Torr, and A. Zisserman, Lecture Notes in Computer Science Vol. 5304 (Springer, New York, 2008), pp. 495–508.
- [54] O. Pele and B. Taskar, The tangent Earth mover's distance, in *Geometric Science of Information—First International Conference, GSI 2013, Paris, France, 2013. Proceedings*, edited by F. Nielsen and F. Barbaresco, Lecture Notes in Computer Science Vol. 8085 (Springer, New York, 2013), pp. 397–404.

- [55] L. N. Wasserstein, Markov processes over denumerable products of spaces describing large systems of automata, *Probl. Inf. Transm.* **5**, 47 (1969).
- [56] R. L. Dobrushin, Prescribing a system of random variables by conditional distributions, *Theory Probab. Appl.* **15**, 458 (1970).
- [57] S. Bright-Thonney, I. Mout, B. Nachman, and S. Prestel, Systematic quark/gluon identification with ratios of likelihoods, *J. High Energy Phys.* **12** (2022) 021.
- [58] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet* user manual, *Eur. Phys. J. C* **72**, 1896 (2012).
- [59] M. Cacciari, G. P. Salam, and G. Soyez, The anti- $k_r$  jet clustering algorithm, *J. High Energy Phys.* **04** (2008) 063.
- [60] S. Chatrchyan *et al.* (CMS Collaboration), Search for a Higgs boson in the decay channel  $H \rightarrow ZZ^* \rightarrow q\bar{q}\ell^-\ell^+$  in  $pp$  collisions at  $\sqrt{s} = 7$  TeV, *J. High Energy Phys.* **04** (2012) 036.
- [61] 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.
- [62] <http://iaifi.org/>.