Published for SISSA by Springer

Received: *November 5, 2018*
Revised: *January 3, 2019*
Accepted: *January 7, 2019*
Published: *January 15, 2019*

Energy flow networks: deep sets for particle jets

JHEP01(2019)121

Patrick T. Komiske, Eric M. Metodiev and Jesse Thaler

Center for Theoretical Physics, Massachusetts Institute of Technology,
77 Massachusetts Avenue, Cambridge, MA 02139, U.S.A.
Department of Physics, Harvard University,
17 Oxford Street, Cambridge, MA 02138, U.S.A.

E-mail: pkomiske@mit.edu, metodiev@mit.edu, jthaler@mit.edu

Abstract: A key question for machine learning approaches in particle physics is how to best represent and learn from collider events. As an event is intrinsically a variable-length unordered set of particles, we build upon recent machine learning efforts to learn directly from sets of features or "point clouds". Adapting and specializing the "Deep Sets" framework to particle physics, we introduce Energy Flow Networks, which respect infrared and collinear safety by construction. We also develop Particle Flow Networks, which allow for general energy dependence and the inclusion of additional particle-level information such as charge and flavor. These networks feature a per-particle internal (latent) representation, and summing over all particles yields an overall event-level latent representation. We show how this latent space decomposition unifies existing event representations based on detector images and radiation moments. To demonstrate the power and simplicity of this set-based approach, we apply these networks to the collider task of discriminating quark jets from gluon jets, finding similar or improved performance compared to existing methods. We also show how the learned event representation can be directly visualized, providing insight into the inner workings of the model. These architectures lend themselves to efficiently processing and analyzing events for a wide variety of tasks at the Large Hadron Collider. Implementations and examples of our architectures are available online in our EnergyFlow package.

Keywords: Jets, QCD Phenomenology

ArXiv ePrint: 1810.05165

https://doi.org/10.1007/JHEP01(2019)121

Contents

1 Introduction

Collisions at accelerators like the Large Hadron Collider (LHC) produce multitudes of particles. Particles are the fundamental objects of interest in collider physics and provide an interface between theoretical calculations and experimental measurements, often reconstructed experimentally via "particle flow" algorithms [1–3]. Analyses of collider data rely on observables to distill these complicated multiparticle events and capture essential aspects of the underlying physics. Because each collision event consists of a variable-length list of particles with no intrinsic ordering, collider observables must be sensibly defined as functions of *sets* of particles. In this paper, we develop a novel architecture for processing and learning from collider events in their natural set-based representation.

Recently, modern machine learning techniques have been used to achieve excellent performance on a variety of collider tasks by learning specalized functions of the events, which can be viewed as observables in their own right. For instance, hadronic jet classification has been thoroughly studied using low-level [4–23] and high-level [24–29] input observables. Additional tasks include the removal of pileup [30], model-independent new physics searches [31–36], constraining effective field theories [37–39], probabilistic and generative modeling of physics processes [40–44], and enhancing existing physics analyses [45–53]. See refs. [54–58] for more detailed reviews of machine learning in high-energy physics.

Two key choices must be made when using machine learning for a collider task: how to represent the event and how to analyze that representation. These choices are often made together, with examples from collider physics including calorimeter images paired with convolutional neural networks (CNNs) [5–14], particle lists paired with recurrent/recursive neural networks (RNNs) [15–19], collections of ordered inputs paired with dense neural networks (DNNs) [20–27], and Energy Flow Polynomials (EFPs) paired with linear methods [29]. One lesson that emerges from this body of work is that any two sufficiently general models, given access to complete enough information, achieve similar performance. In light of this, criteria such as understandability of the model and closeness to theoretical and experimental constructs are of central importance.

Given that events are fundamentally sets of particles, particle-level inputs such as those used in refs. [15–22] are a desirable way of representing an event for use in a model. That said, RNNs and DNNs, the two architectures typically used with particle-level inputs, each fail to be fully satisfactory methods for processing events: DNNs because they require a fixed-size input and RNNs because they are explicitly dependent on the ordering of the inputs. Though ad hoc workarounds for these problems exist, such as zero padding for DNNs or ordering particles by their transverse momentum (p_T) or clustering history for RNNs, an ideal architecture would manifestly respect the permutation symmetry of the problem. Such an architecture would be able to handle variable-length inputs while being inherently symmetric with respect to the ordering of the input particles.

The machine learning community has recently developed (and continues to explore) technology which is ideally suited for representing sets of objects for a model [59–68]. One context where this appears is learning from *point clouds*, sets of data points in space. For instance, the output of spatial sensors such as lidar, relevant for self-driving car technologies, is often in the form of a point cloud. As point clouds share the variable-length and permutation-symmetric properties with collider events, it is worthwhile to understand and expand upon point cloud techniques for particle physics applications.

The Deep Sets framework for point clouds, recently developed in ref. [63], demonstrates how permutation-invariant functions of variable-length inputs can be parametrized in a fully general way. In ref. [63], the method was applied to a wide variety of problems including red-shift estimation of galaxy clusters, finding terms associated with a set of words, and detecting anomalous faces in a set of images. The key observation is that *summation*, which is clearly symmetric with respect to the order of the arguments, is general enough to encapsulate all symmetric functions if one is allowed a large enough internal (*latent*) space.

In the context of a physics observable $\mathcal{O}$ that is a symmetric function of an arbitrary number of particles each with d features, the result from ref. [63] can be stated as:

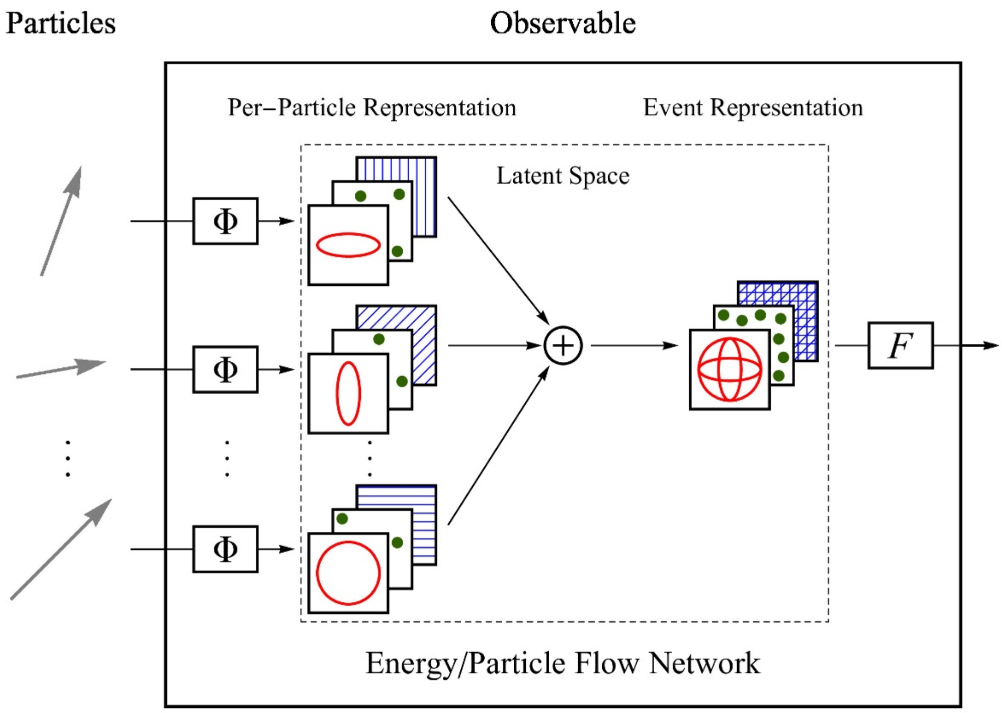


Figure 1. A visualization of the decomposition of an observable via eq. (1.1). Each particle in the event is mapped by Φ to an internal (latent) particle representation, shown here as three abstract illustrations for a latent space of dimension three. The latent representation is then summed over all particles to arrive at a latent event representation, which is mapped by F to the value of the observable. For the IRC-safe case of eq. (1.2), Φ takes in the angular information of the particle and the sum is weighted by the particle energies or transverse momenta.

Observable Decomposition. *An observable $\mathcal{O}$ can be approximated arbitrarily well as:*

$$\mathcal{O}(\{p_1, \dots, p_M\}) = F\left(\sum_{i=1}^{M} \Phi(p_i)\right), \tag{1.1}$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}^\ell$ is a per-particle mapping and $F : \mathbb{R}^\ell \to \mathbb{R}$ is a continuous function.

A schematic representation of eq. (1.1) is shown in figure 1. Inherent in the decomposition of eq. (1.1) is a latent space of dimension ℓ that serves to embed the particles such that an overall latent event representation is obtained when the sum is carried out. One should think of the d features for each particle as possibly being kinematic information, such as the particle's p_T, rapidity y, and azimuthal angle ϕ, or other quantum numbers such as the particle's charge or flavor. Section 2 contains additional mathematical details regarding this decomposition.

With a suitable modification of eq. (1.1), we can restrict the decomposition to infrared- and collinear-safe (IRC-safe) observables:

IRC-Safe Observable Decomposition. *An IRC-safe observable* $\mathcal{O}$ *can be approximated arbitrarily well as:*

$$\mathcal{O}(\{p_1,\dots,p_M\}) = F\left(\sum_{i=1}^{M} z_i \Phi(\hat{p}_i)\right), \tag{1.2}$$

where z_i *is the energy (or* p_T*) and* $\hat{p}_i$ *the angular information of particle* i*.*

The energy-weighting factors z_i as well as the energy-independent $\hat{p}_i$ in eq. (1.2) ensure that the event representation in the latent space is IRC-safe.

In this paper, we show that many common observables are naturally encompassed by simple choices of Φ and F from eqs. (1.1) and (1.2). Furthermore, we can parametrize Φ and F by neural network layers, capable of learning essentially any function, in order to explore more complicated observables. In keeping with the naming convention of ref. [29] for methods involving IRC-safe observables, we term a network architecture implementing eq. (1.2) an *Energy Flow Network* (EFN). By contrast, we refer to the more general case of an architecture that implements eq. (1.1) as a *Particle Flow Network* (PFN). These two network architectures can be mathematically summarized as:

$$\text{EFN:}\quad F\left(\sum_{i=1}^{M} z_i \Phi(\hat{p}_i)\right), \qquad\qquad \text{PFN:}\quad F\left(\sum_{i=1}^{M} \Phi(p_i)\right). \tag{1.3}$$

Our framework manifestly respects the variable length and permutation invariance of particle sets, achieves performance competitive with existing techniques on key collider tasks, and provides a platform for visualizing the information learned by the model. Beyond this, we demonstrate how our framework unifies the existing event representations of calorimeter images and radiation moments, and we showcase the extraction of novel analytic observables from the trained model.

One ever-present collider phenomenon that involves complicated multiparticle final states is the formation and observation of *jets*, sprays of color-neutral hadrons resulting from the fragmentation of high-energy quarks and gluons in quantum chromodynamics (QCD). Numerous individual observables have been proposed to study jets including the jet mass, constituent multiplicity, image activity [69], N-subjettiness [70, 71], track-based observables [72, 73], generalized angularities [74], (generalized) energy correlation functions [75, 76], soft drop multiplicity [77, 78], and many more (see refs. [54, 79–83] for reviews). Machine learning methods have found tremendous applicability to jet classification tasks, greatly outperforming individual standard observables. Jet classification provides an ideal case study for the Deep Sets method in a collider setting since jets, like events, are fundamentally variably sized and invariant under reorderings of their constituents.

Many existing collider observables ranging from e^+e^- event shapes to jet substructure observables naturally fit into the decomposition of eq. (1.1). Observables that are defined directly in terms of the particles themselves (i.e. not algorithmically) can often be exactly encompassed. Several examples of such observables are summarized in table 1, with the associated functions Φ and F listed for each observable. The fact that the decomposition holds exactly in these familiar cases indicates that the Observable Decomposition indeed captures an essential aspect of particle-level collider observables.

JHEP01(2019)121

Observable $\mathcal{O}$		Map Φ	Function F
Mass	m	p^μ	$F(x^\mu) = \sqrt{x^\mu x_\mu}$
Multiplicity	M	1	$F(x) = x$
Track Mass	m_{track}	$p^\mu \mathbb{I}_{\text{track}}$	$F(x^\mu) = \sqrt{x^\mu x_\mu}$
Track Multiplicity	M_{track}	$\mathbb{I}_{\text{track}}$	$F(x) = x$
Jet Charge [72]	$\mathcal{Q}_\kappa$	$(p_T, Q\, p_T^\kappa)$	$F(x, y) = y/x^\kappa$
Eventropy [74]	$z \ln z$	$(p_T, p_T \ln p_T)$	$F(x, y) = y/x - \ln x$
Momentum Dispersion [93]	p_T^D	(p_T, p_T^2)	$F(x, y) = \sqrt{y/x^2}$
C parameter [94]	C	$(\lvert\vec{p}\rvert, \vec{p} \otimes \vec{p}/\lvert\vec{p}\rvert)$	$F(x, Y) = \frac{3}{2x^2}[(\operatorname{Tr} Y)^2 - \operatorname{Tr} Y^2]$

Table 1. A variety of common collider observables decomposed into per-particle maps Φ and functions F according to eq. (1.1). Here $\mathbb{I}_{\text{track}}$ is an indicator function over charged tracks. In the last column, the arguments of F are placeholders for the summed output of Φ.

To showcase the efficacy of EFNs and PFNs, we apply them to the task of distinguishing light-quark jets from gluon jets [84–87], finding that they achieve excellent classification performance. In general, the PFN model outperforms the EFN model, indicating that IRC-unsafe information is helpful for discriminating quark and gluon jets. Additionally, including particle identification information improves the classification performance of the PFN. It would be interesting to apply all of these methods in a fully-data driven way [88–90] to test these conclusions beyond the limited scope of parton shower generators.

One fascinating aspect of EFNs is that they enable a natural visualization of the learned latent space, providing insights as to what exactly the machine is learning. In particular, since the function Φ of an EFN typically takes the two-dimensional angular information of a particle as input, this two-dimensional space is easily visualized. In the context of quark/gluon discrimination, we observe that the EFN learns a latent representation that "pixelates" the rapidity-azimuth plane, dynamically sizing the pixels to be smaller near the core of the jet and larger farther out. We also find qualitative and quantitative evidence that the EFN has in a sense "understood" the collinear singularity structure of QCD.

The rest of this paper is organized as follows. Section 2 provides a detailed mathematical discussion of the observable decompositions and explores eqs. (1.1) and (1.2) in the context of specific observables and event representations. Section 3 discusses the implementation details of our EFN and PFN architectures, with other models discussed in appendix A. Section 4 contains the case study discriminating quark- and gluon-initiated jets and demonstrates our new techniques for visualizing and analyzing the learned information. Conclusions are presented in section 5. A supplementary top jet tagging study is presented in appendix B, and additional visualizations of the models are provided in appendix C. The EFN and PFN architectures are available online as part of our `EnergyFlow` package [91] along with example code.

2 A general framework for observables

Events consist of variable numbers of particles with no intrinsic ordering, so observables are described mathematically as functions of sets of particles. Such a mathematical formulation allows for a direct exploration of the space of observables. For instance, ref. [29] exploited IRC safety to construct a linear approximating basis of all IRC-safe observables. Here, we treat the entire space of observables (both with and without IRC safety), using their mathematical structure to arrive at a general decomposition relevant for theoretically categorizing observables as well as developing machine learning techniques.

2.1 Observables as functions of sets of particles

The key mathematical fact that we exploit, due to ref. [63], is that a generic function of a set of objects can be decomposed to arbitrarily good approximation in a practical and intuitive way. We state this result explicitly below:

Deep Sets Theorem [63]. *Let $\mathfrak{X} \subset \mathbb{R}^d$ be compact, $X \subset 2^{\mathfrak{X}}$ be the space of sets with bounded cardinality of elements in $\mathfrak{X}$, and $Y \subset \mathbb{R}$ be a bounded interval. Consider a continuous function $f : X \to Y$ that is invariant under permutations of its inputs, i.e. $f(x_1, \dots, x_M) = f(x_{\pi(1)}, \dots, x_{\pi(M)})$ for all $x_i \in \mathfrak{X}$ and $\pi \in S_M$. Then there exists a sufficiently large integer ℓ and continuous functions $\Phi : \mathfrak{X} \to \mathbb{R}^\ell$, $F : \mathbb{R}^\ell \to Y$ such that the following holds to an arbitrarily good approximation:*[1]

$$f(\{x_1, \dots, x_M\}) = F\left(\sum_{i=1}^{M} \Phi(x_i)\right). \tag{2.1}$$

We only rely on the Deep Sets Theorem to justify the generality of eq. (2.1), which can otherwise be regarded as an interesting, manifestly permutation-invariant parameterization.

The Deep Sets Theorem can be immediately applied to the collider physics context where observables are viewed as functions of sets of particles. We denote an event with M particles as $\{p_i\}_{i=1}^M$, where $p_i \in \mathbb{R}^d$ contains the relevant attributes of particle i (momentum, charge, flavor, etc.). Phrased in the collider physics language, it states that an observable $\mathcal{O}$ can be approximated arbitrarily well as:

$$\mathcal{O}(\{p_1, \dots, p_M\}) = F\left(\sum_{i=1}^{M} \Phi(p_i)\right), \tag{2.2}$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}^\ell$ is a per-particle mapping and $F : \mathbb{R}^\ell \to Y$ is a continuous function. This provides a mathematical justification for the Observable Decomposition stated in eq. (1.1).

The content of the Observable Decomposition is that any observable can be viewed as linearly summing over the particles in some internal space and then mapping the result

[1]It is formally necessary to restrict the domains and ranges of the functions to be compact because the proof of the Deep Sets Theorem, given fully in ref. [63], makes use of the Stone-Weierstrass polynomial approximation theorem [92], which applies for compact spaces. After the expansion in polynomials of the features, the result follows by careful application of the fundamental theorem of symmetric polynomials.

JHEP01(2019)121

to an output space. We refer to $\mathbb{R}^\ell$ as the *latent space* and each component of the per-particle mapping $\Phi(p_i)$ as a *filter*. The latent space could be, for example, the pixel values of a detector image or a moment decomposition of the radiation pattern. Summing $\Phi(p_i)$ over the particles induces a latent description of the entire event, which is mapped by the function F to the value of the observable.

2.2 Enforcing infrared and collinear safety

We can formulate the Observable Decomposition specifically for a class of observables of particular theoretical interest, namely IRC-safe observables [95–98]. IRC safety corresponds to robustness of the observable under collinear splittings of a particle or additions of soft particles, which makes the observable tractable in perturbative quantum field theory as well as robust to experimental resolution effects.

Remarkably, building IRC safety into the latent representation simply corresponds to energy-weighting the contributions of each particle and restricting Φ to only depend on the particle geometry $\hat{p}_i$. The energy-weighting z_i and geometry $\hat{p}_i$ for particle i depends on the collider context. At an e^+e^- collider, it is natural to take $z_i = E_i$ and $\hat{p}_i = p_i^\mu/E_i$, where E_i is the energy and p_i^μ the four-momentum. At a hadron collider, it is natural to take $z_i = p_{T,i}$ and $\hat{p}_i = (y_i, \phi_i)$, where $p_{T,i}$ is the transverse momentum, y_i is the rapidity, and ϕ_i the azimuthal angle.[2] In practice, we typically focus on dimensionless observables and use the appropriate normalized weights: $z_i = E_i/\sum_j E_j$ or $z_i = p_{T,i}/\sum_j p_{T,j}$.

Any IRC-safe observable $\mathcal{O}$ can be approximated arbitrarily well by the decomposition:

$$\mathcal{O}\left(\{p_i\}_{i=1}^M\right) = F\left(\sum_{i=1}^M z_i\, \Phi(\hat{p}_i)\right), \tag{2.3}$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}^\ell$ is a per-particle angular mapping and $F : \mathbb{R}^\ell \to \mathbb{R}$ is continuous. All observables of the form in eq. (2.3) are manifestly IRC safe due to the energy-weighted linear sum structure, the dependence of Φ on purely geometric inputs $\hat{p}_i$, and the fact that continuous functions of IRC-safe observables are IRC safe.[3]

The fact that the energy-weighted decomposition in eq. (2.3) suffices to approximate all IRC-safe observables is intuitive from the fact that a continuous function of a sufficiently high-resolution calorimeter image can be used to approximate an IRC-safe observable arbitrarily well [101–103]. As discussed in section 2.3, an image of the calorimeter deposits is exactly encompassed by the energy-weighted observable decomposition.

Here, we provide a direct argument to arrive at eq. (2.3), building off the Deep Sets Theorem and following similar logic as ref. [29]. Given the decomposition of an IRC-safe observable $\mathcal{O}$ into F and Φ via eq. (2.2), the IRC safety of the observable $\mathcal{O}$ corresponds

[2] As discussed in ref. [29], another sensible choice for the angular measure is $\hat{p}_i = p_i^\mu/p_{T,i}$. Particle mass information, if present, can be passed to a PFN via flavor information.

[3] Ratios of IRC-safe observables are not necessarily IRC safe [99, 100] since division is discontinuous at zero.

to the following statements:

$$\text{IR safety:}\quad F\left(\sum_{i=1}^{M}\Phi(z_i,\hat{p}_i)\right)=F\left(\Phi(0,\hat{p}_0)+\sum_{i=1}^{M}\Phi(z_i,\hat{p}_i)\right), \tag{2.4}$$

$$\text{C safety:}\quad F\left(\sum_{i=1}^{M}\Phi(z_i,\hat{p}_i)\right)=F\left(\Phi(\lambda z_1,\hat{p}_1)+\Phi((1-\lambda)z_1,\hat{p}_1)+\sum_{i=2}^{M}\Phi(z_i,\hat{p}_i)\right), \tag{2.5}$$

where eq. (2.4) holds for all directions $\hat{p}_0$ that a soft particle could be emitted and eq. (2.5) holds for all energy fractions $\lambda \in [0,1]$ of the collinear splitting. In eq. (2.5), we have selected particle 1 to undergo the collinear splitting but the statement holds for any of the particles by permutation symmetry. The equations here only hold to a specified accuracy of approximation in the Observable Decomposition, which we leave implicit since it does not alter the structure of our argument.

We now make the following suggestive redefinition of Φ to ensure that the latent representation of a particle vanishes if the particle has zero energy:

$$\Phi(z,\hat{p}) \to \Phi(z,\hat{p}) - \Phi(0,\hat{p}). \tag{2.6}$$

Infrared safety via eq. (2.4) ensures that the value of the observable is unchanged under this redefinition, so without loss of generality we may take Φ to vanish on arbitrarily soft particles.

Making another convenient redefinition of Φ, we choose a $\lambda \in [0,1]$ and let:

$$\Phi(z,\hat{p}) \to \Phi(\lambda z,\hat{p}) + \Phi((1-\lambda)z,\hat{p}). \tag{2.7}$$

Collinear safety via eq. (2.5) ensures that the value of the observable is unchanged under such a redefinition, which holds for any $\lambda \in [0,1]$.

We now show that the freedom to redefine the mapping Φ using eqs. (2.6) and (2.7) for an IRC-safe observable leads to the IRC-safe Observable Decomposition in eq. (2.3). To see this, consider approximating Φ in the energy argument z via the Stone-Weierstrass theorem. Calling the angular coefficients of each term $C_n(\hat{p})$ yields:

$$\Phi(z,\hat{p}) = \sum_{n=0}^{\mathcal{N}} z^n\, C_n(\hat{p}) = C_0(\hat{p}) + z\, C_1(\hat{p}) + \sum_{n=2}^{\mathcal{N}} z^n\, C_n(\hat{p}), \tag{2.8}$$

for some large but finite $\mathcal{N}$. How large $\mathcal{N}$ must be depends on the specified precision that we have been leaving implicit.

Invoking the soft redefinition in eq. (2.6), Φ may be taken to vanish on arbitrarily soft particles, which allows $C_0(\hat{p})$ to be set to zero without changing the value of the observable. Implementing the collinear redefinition in eq. (2.7) after the expansion in energy, we obtain:

$$\Phi(z,\hat{p}) = z\, C_1(\hat{p}) + \sum_{n=2}^{\mathcal{N}} (\lambda^n + (1-\lambda)^n) z^n C_n(\hat{p}). \tag{2.9}$$

From this equation, we seek to argue that $C_n(\hat{p})$ for $n \geq 2$ may be taken to vanish. For $\lambda \in (0,1)$, this redefinition decreases the higher-order coefficients $C_n(\hat{p})$ by a factor of

$\lambda^n + (1-\lambda)^n < 1$ without changing the corresponding observable. Iterated application of this fact allows the higher-order coefficients to be removed while keeping the term linear in the energy. Thus, to an arbitrarily good approximation, we can take $\Phi(z,\hat{p}) = z\,C_1(\hat{p})$ for some angular function $C_1(\hat{p})$, which we subsequently rename to $\Phi(\hat{p})$.

To summarize, the Deep Sets Theorem, combined with IRC safety, shows that the map Φ can be taken to be linear in energy without loss of generality. Collinear safety was critical in arguing that Φ could be taken to be affine linear in the energy and infrared safety was critical in arguing that the constant piece could be set to zero without loss of generality. This is exactly the result needed to justify the IRC-safe Observable Decomposition in eq. (2.3), thereby completing the argument. Beyond potential applications for building IRC safety directly into models, such an observable decomposition for IRC-safe observables may be useful for shedding light on the structure of these important observables more broadly.

2.3 Encompassing image and moment representations

Beyond the single observables tabulated in table 1, entire event representations can be encompassed in the Observable Decomposition framework as well.

One common event representation is to view events as images by treating the energy deposits in a calorimeter as pixel intensities [4–9]. Since typical pixelizations for jet classification are $33 \times 33 \simeq 1000$, the images are quite sparse, with an order of magnitude more pixels than particles. Treating the detector as a camera and events as images allows for modern image recognition technology to be applied to collider physics problems. These images are typically fed into a convolutional neural network, which is trained to learn a function of the images optimized for a specific task.

The image-based event representation of a jet as a collection of pixels fits naturally into the Observable Decomposition. The energy (or transverse momentum) deposited in each pixel is simply a sum over the energies z_i of the particles hitting that pixel. Letting $\mathbb{I}_{j,k}(y,\phi)$ be an indicator function of pixel (j,k) in the rapidity-azimuth plane, we have that the intensity $P_{j,k}$ of pixel (j,k) is:

$$P_{j,k} = \sum_i z_i\,\mathbb{I}_{j,k}(y_i,\phi_i). \tag{2.10}$$

Thus, having Φ be an indicator function for the location of the pixel directly allows the latent representation of the IRC-safe Observable Decomposition to be a detector image. We illustrate this in figure 2 for the rapidity-azimuth plane relevant for a hadron collider. Here, the filters are a collection of localized square bumps evenly spaced throughout the rapidity-azimuth plane.

Another way to represent an event or jet is as a collection of moments of its radiation pattern. Moments (or tensors) have been considered for analyzing hadronic energy flow patterns both for e^+e^- and hadron colliders [104–106]. A moment-based representation has yet to be directly exploited for machine learning applications in collider physics, though is closely related to the EFPs [29].[4] Here we restrict to the collimated case of jets, but a

[4]There is a rich connection between the moments of the event radiation pattern and multiparticle energy correlators, a detailed discussion of which we leave to future work. See footnote 8 of ref. [29].

JHEP01(2019)121

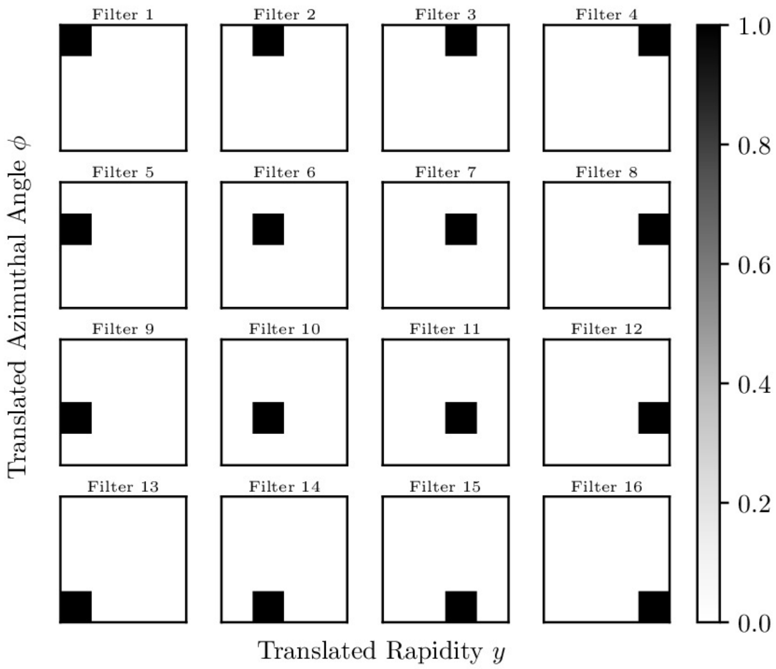


Figure 2. The calorimeter image representation decomposed into a collection of $\Phi(y, \phi)$ filters according to the IRC-safe Observable Decomposition, shown here for the illustrative case of a 4×4 image. The energy deposits in each pixel can be decomposed via eq. (1.2) into an indicator function $\Phi(y, \phi)$ determining whether a particle in position (y, ϕ) hits the pixel.

similar discussion holds at the event level. The moments $I_{m,n}$ of the radiation pattern in the rapidity-azimuth plane are:

$$I_{m,n} = \sum_i z_i \, y_i^m \phi_i^n. \tag{2.11}$$

This can be manifestly decomposed according to the IRC-safe Observable Decomposition by simply making each filter $\Phi(y, \phi) = y^m \phi^n$, as illustrated in figure 3. Here, the filters are a collection of non-localized functions which weight different parts of the event differently.

More generally, we can visualize $\Phi(y, \phi)$ for learned IRC-safe latent spaces, where the model itself learns its own event representation. In interpreting these visualizations, it is worth keeping in mind that localized filters like figure 2 correspond to an image-like representation, while global filters like figure 3 correspond to a moment-like representation. The flexibility of the IRC-safe Observable Decomposition allows for more complicated filters as well. As we will see in section 4.4, visualizing the latent space is extremely useful in understanding the behavior of EFNs. Moreover, similar (albeit higher-dimensional) visualizations can be performed in the general PFN case of $\Phi(p)$ and have been explored in the point cloud context [60].

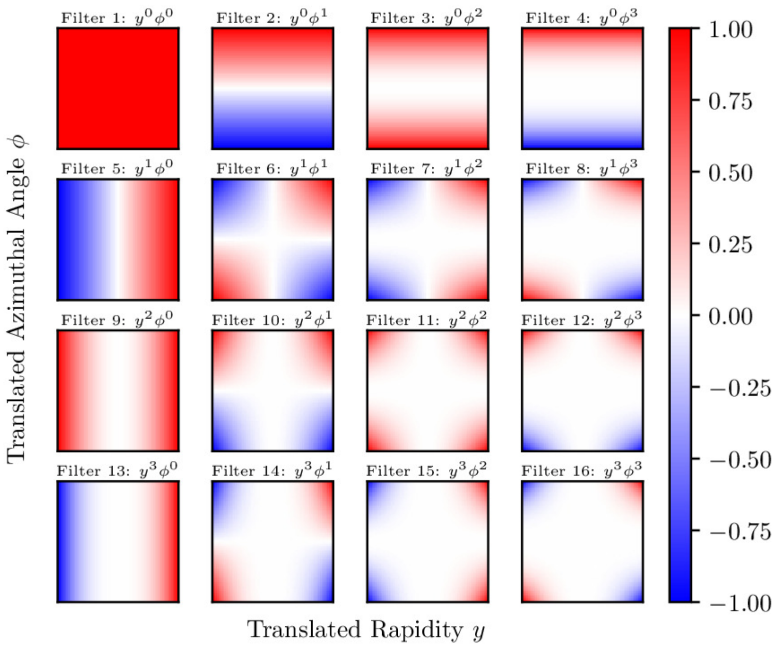


Figure 3. The radiation moment representation decomposed into a collection of $\Phi(y, \phi)$ filters according to the IRC-safe Observable Decomposition. The (m, n) moment of the energy distribution in the rapidity-azimuth plane can be decomposed via eq. (1.2) into $\Phi(y, \phi) = y^m \phi^n$, shown here with increasing m downward and increasing n to the right.

3 Network implementation

In this section, we describe our implementation and adaptation of the Deep Sets decomposition for use in a particle physics context. In light of the quark versus gluon jet case study presented in section 4, we focus here on inputting individual jets to the model, though we emphasize that the method is broadly applicable at the event level.

3.1 Preprocessing

The goal of preprocessing inputs is to assist the model in its effort to solve an optimization problem. Typically, preprocessing steps are optional, but are applied in order to improve the numerical convergence of the model, given the practical limitations of finite dataset and model size, as well as the particular choice of parameter initialization. The preprocessing described in this section was found to be helpful, and sometimes necessary, for achieving a well-trained EFN or PFN model for the applications considered in section 4. It is likely that for further applications of EFNs or PFNs, such as event-level studies, the appropriate preprocessing steps may change.

For the models we construct, kinematic information — transverse momentum p_T, rapidity y, and azimuthal angle ϕ — are always given for each particle. We preprocess these features as follows: the transverse momenta are divided by the total scalar sum p_T and the rapidities and azimuthal angles are centered based on the rapidity and azimuthal angle of the jet, using the E-scheme jet axis. In terms of the four-momentum of each particle, this preprocessing step can be cast into the following suggestive form:

$$p_{T,i} \to \frac{p_{T,i}}{\sum_j p_{T,j}}, \quad y_i \to y_i - \left(\sum_j p_{T,j}\,\hat{p}_j\right)_y, \quad \phi_i \to \phi_i - \left(\sum_j p_{T,j}\,\hat{p}_j\right)_\phi, \tag{3.1}$$

with $\hat{p}_i = p_i^\mu/p_{T,i}$, where the subscripts indicate the rapidity and azimuth of the jet four-vector. This notation makes clear that the per-particle preprocessing of eq. (3.1) solely relies on the scalar sum p_T, rapidity, and azimuth of the jet, which itself can be written in terms of an IRC-safe Observable Decomposition with $\Phi(\hat{p}) = (1, \hat{p})$. Alternative jet centerings, such as those based on the p_T-weighted centroid, also fit nicely into this framework.[5]

Optionally, the inputs may also include particle identification (ID) information. Though typically encoded using the Particle Data Group (PDG) particle numbering scheme [107], the large and irregular integer values employed therein are not ideal inputs to a model expecting inputs roughly in the numerical range $[-1, 1]$. Therefore, a mapping from PDG IDs to small floats is performed for each particle (the details of which are provided below). While this approach, which only uses a single feature to encode the particle ID information, should be sufficient to input this information to the model, alternative approaches using multiple categorical features may be easier for the model to interpret, since particle ID is inherently discrete rather than continuous. For instance, using two additional features per particle, one feature could indicate the charge of the particle $\{-1, 0, +1\}$ and the other one could indicate $\{h, \gamma, e, \mu\}$ (where h corresponds to a hadron, one of π, K, n, p), covering an experimentally realistic particle ID scheme. One-hot encoding of the particle ID is another option.

In order to explore how particle identification is helpful to a model,[6] we use it in four different ways, each with a PFN architecture. We describe each of the different models and levels of information used throughout section 4 below:

- **PFN-ID**: PFN, adding in the full particle ID information. For the case study in section 4, particles are indicated as being one of γ, π^+, π^-, K^+, K^-, K_L, n, $\bar{n}$, p, $\bar{p}$, e^-, e^+, μ^-, μ^+, which are represented to the model as a single float value starting at 0 and increasing by 0.1 for each distinct type, respectively.[7]

- **PFN-Ex**: PFN, adding in experimentally realistic particle ID information. For the case study in section 4, particles are indicated as being one of γ, h^+, h^-, h^0, e^-, e^+, μ^-, μ^+, which are represented to the model analogously to the PFN-ID case.[8]

[5]These observations motivate an iterative local-global architecture which learns an event representation, applies it per-particle, and repeats. Such an architecture could explicitly or learnably fold in this preprocessing as a first step. We leave further developments in this direction to future work.

[6]We perform this comparison at particle level without detector simulation. Detector effects may change or degrade the information available in the different particle types. Doing such an exploration with detector simulation (or in data) is an interesting avenue for additional exploration.

[7]Note that π^0 is absent since we include its decay, usually into two photons.

[8]These categories are based on particle flow reconstruction algorithms at ATLAS and CMS [1–3], where $h^\pm = \pi^\pm/K^\pm/p/\bar{p}$ and $h^0 = K_L/n/\bar{n}$. Additional experimental information, such as $\pi/K/p$ separation, feasible at ALICE and LHCb (or at ATLAS and CMS at low p_T), can carry added information, as could

- **PFN-Ch**: PFN, adding in the electric charge of the particles as an additional feature.
- **PFN**: the particle flow network using only three-momentum information via eq. (1.1).
- **EFN**: the energy flow network using only IRC-safe latent space information via eq. (1.2).

3.2 Network architecture

So far, there has not yet been any machine learning in our effort to apply the decompositions in eqs. (1.1) and (1.2) to collider data. The machine learning enters by choosing to approximate the functions Φ and F with neural networks.[9] Neural networks are a natural choice to use because sufficiently large neural networks can approximate any well-behaved function.

To parametrize the functions Φ and F in a sufficiently general way, we use several dense neural network layers as universal approximators, as shown in figure 4. For Φ, we employ three dense layers with 100, 100, and ℓ nodes, respectively, where ℓ is the latent dimension that will be varied in powers of 2 up to 256. For F, we use three dense layers, each with 100 nodes. We confirmed that several network architectures with more or fewer layers and nodes achieved similar performance. Each dense layer uses the ReLU activation function [108] and He-uniform parameter initialization [109]. A two-unit layer with a softmax activation function is used as the output layer of the classifier. See appendix A for additional details regarding the implementations of the EFN, PFN, and other networks. The `EnergyFlow` Python package [91] contains implementations and examples of EFN and PFN architectures.

4 Discriminating quark and gluon jets

To demonstrate the EFN architecture in a realistic setting, we implement and train an EFN and several PFN variants to discriminate light-quark from gluon initiated jets [84–87], a problem relevant for new physics searches as well as precision measurements. See appendix B for a similar study on classifying top jets from QCD jets using samples based on ref. [22].

4.1 Event generation

The samples used for this study were $Z(\to \nu\bar{\nu})+g$ and $Z(\to \nu\bar{\nu})+(u,d,s)$ events generated with Pythia 8.226 [110, 111] at $\sqrt{s} = 14\,\text{TeV}$ using the `WeakBosonAndParton:qqbar2gmZg` and `WeakBosonAndParton:qg2gmZq` processes, ignoring the photon contribution and requiring the Z to decay invisibly to neutrinos. Hadronization and multiple parton interactions (i.e. underlying event) were turned on and the default tunings and shower parameters were

exclusive hadron reconstruction. Particle ID information is typically captured in likelihood ratios for different particle hypotheses, which fits naturally into a categorical encoding scheme where there is a feature for each particle-type likelihood ratio.

[9] Ref. [63] describes two types of architectures in the Deep Sets framework, termed *invariant* and *equivariant*. Equivariance corresponds to producing per-particle outputs that respect permutation symmetry. For this paper, our interest is in the invariant case, but we leave for future work an exploration of the potential particle physics applications of an equivariant architecture.

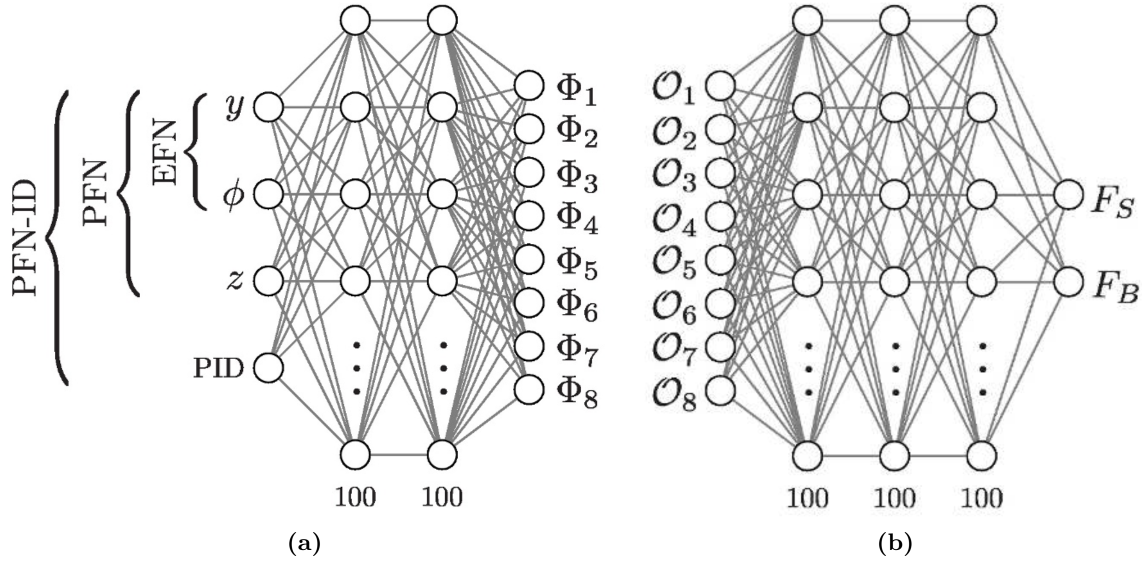


Figure 4. The particular dense networks used here to parametrize (a) the per-particle mapping Φ and (b) the function F, shown for the case of a latent space of dimension $\ell = 8$. For the EFN, the latent observable is $\mathcal{O}_a = \sum_i z_i \, \Phi_a(y_i, \phi_i)$. For the PFN family, the latent observable is $\mathcal{O}_a = \sum_i \Phi_a(y_i, \phi_i, z_i, \text{PID}_i)$, with different levels of particle-ID (PID) information. The output of F is a softmaxed signal (S) versus background (B) discriminant.

used. Final state non-neutrino particles were clustered into $R = 0.4$ anti-k_T jets [112] using FastJet 3.3.0 [113]. Jets with $p_T \in [500, 550]$ GeV and $|y| < 2.0$ were kept. No detector simulation was performed.[10] While labeling these jets using quark/gluon parton labels is manifestly unphysical, applications of these techniques at colliders could rely on an operational jet flavor definition [90] and weak supervision techniques for training directly on data [88, 89] (see also refs. [114–117]).

4.2 Classification performance

A standard tool to analyze a classifier is the receiver operating characteristic (ROC) curve, obtained from the true positive ε_s and false positive ε_b rates as the decision threshold is varied. This may also be plotted as a Significance Improvement (SI) curve [85], namely $\varepsilon_s/\sqrt{\varepsilon_b}$ as a function of ε_s. To condense the performance of a classifier into a single quantity, the area under the ROC curve (AUC) is commonly used, which is also the probability that the classifier correctly sorts randomly drawn signal (quark jet) and background (gluon jet) samples. An AUC of 0.5 corresponds to a random classifier and an AUC of 1.0 corresponds to a perfect classifier. We also report the background rejection at 50% signal efficiency ($1/\varepsilon_b$ at $\varepsilon_s = 50\%$) as an alternative performance metric.

For each of the models, we sweep the latent dimension ℓ of the internal representation from 2 to 256 in powers of 2. As discussed in section 3.1, four PFN models were trained

[10] In the context of experimental applications, it is worth noting that the different resolutions of different particle types can be naturally accomodated in our framework.

JHEP01(2019)121

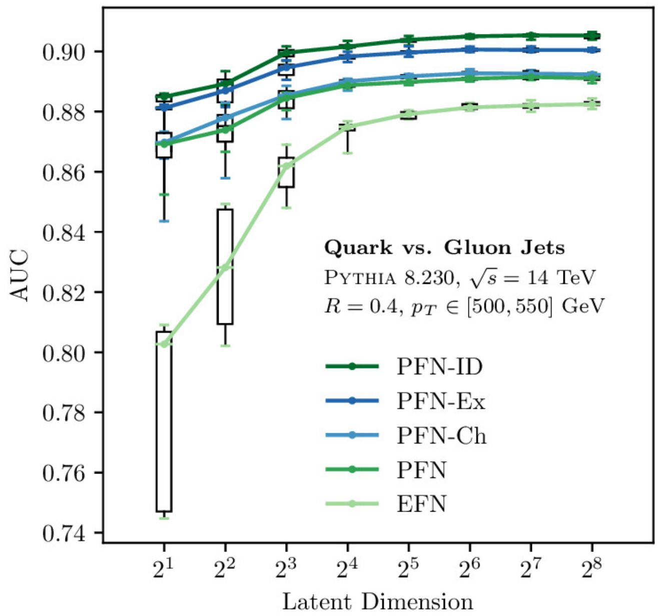


Figure 5. The AUC performance of the EFN and PFN models as a function of the latent dimension of the model, which is varied from 2 to 256 in powers of 2. The spread in values is due to training the model ten times with different initializations. The performance generally increases with larger latent dimensions, with saturation observed by latent dimension 256. The best model is PFN-ID, which uses full particle-type information, followed closely by PFN-Ex, which uses experimentally realistic particle-type information. The PFN without any extra information performs roughly the same as the PFN-Ch, which uses charge information. The fact that the EFN is lowest on this plot indicates that there is discrimination power to be found in IRC-unsafe information.

each with different particle-type information. Models are trained ten times each to give a sense of the variation and stability of the training. The resulting model performances as quantified by the AUC are shown in figure 5. As anticipated, the performance of each model increases as the latent dimension increases, with good performance achieved by $\ell = 16$. The higher variance at low latent dimensions arises because some of the filters fail to train to non-zero values in those cases. The performance of the models appears to saturate by the larger latent dimensions, which justifies our use of $\ell = 256$ as our benchmark latent dimension size for additional explorations.

In figure 6, we show the full ROC and SI curves of these models with latent dimension 256. The best model performance of all tested techniques and models was the PFN-ID with full particle ID, followed closely by the PFN-Ex with experimentally realistic particle ID. Figures 5 and 6 show a well-defined hierarchy of model performances at all latent dimension sizes based on the information available to the model. The fact that the PFNs outperform the EFN indicates that IRC-unsafe information is helpful for discrimination, which is not surprising in light of the fact that the constituent multiplicity is IRC unsafe and is known to be a good quark/gluon discriminant [85]. Though IRC-unsafe information is helpful, it is instructive to test both EFNs and PFNs to probe how different kinds of information are

JHEP01(2019)121

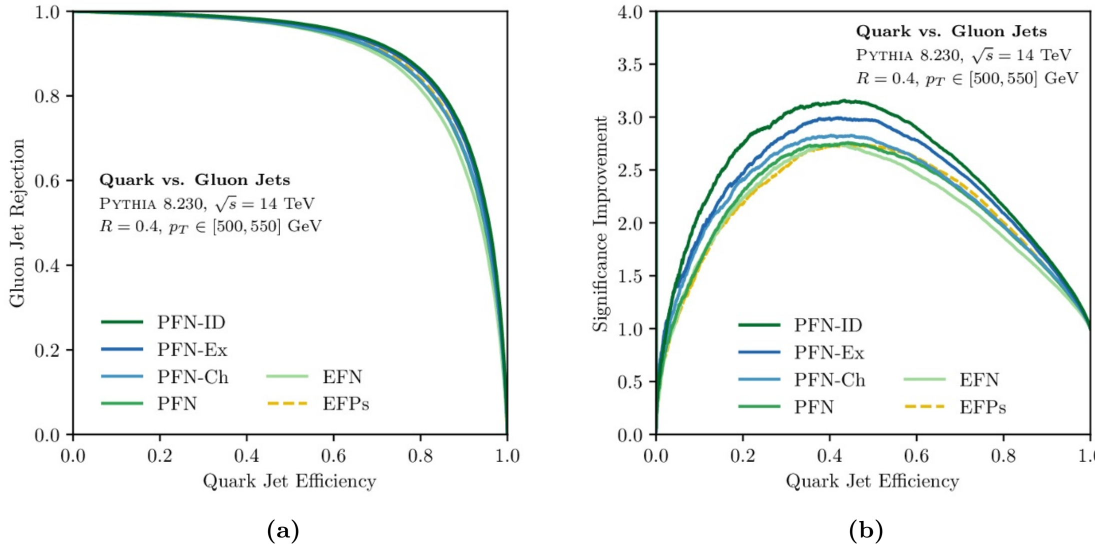


Figure 6. The (a) ROC and (b) SI curves of the median (selected by AUC) EFN and PFN models with latent dimensions of 256. The linear EFP model is shown for comparison. The PFN-ID with full particle ID yields the best performance of all models, followed by the PFN-Ex using experimentally realistic particle ID. The EFN and EFP models perform comparably in terms of maximum SI, indicating that the available IRC-safe information is being captured consistently by these very different architectures.

used by the classifier. Furthermore, sometimes an IRC-safe model is desired as it may be more robust to detector effects or mismodeling of infrared physics such as hadronization in simulated training data.

4.3 Comparison to other architectures

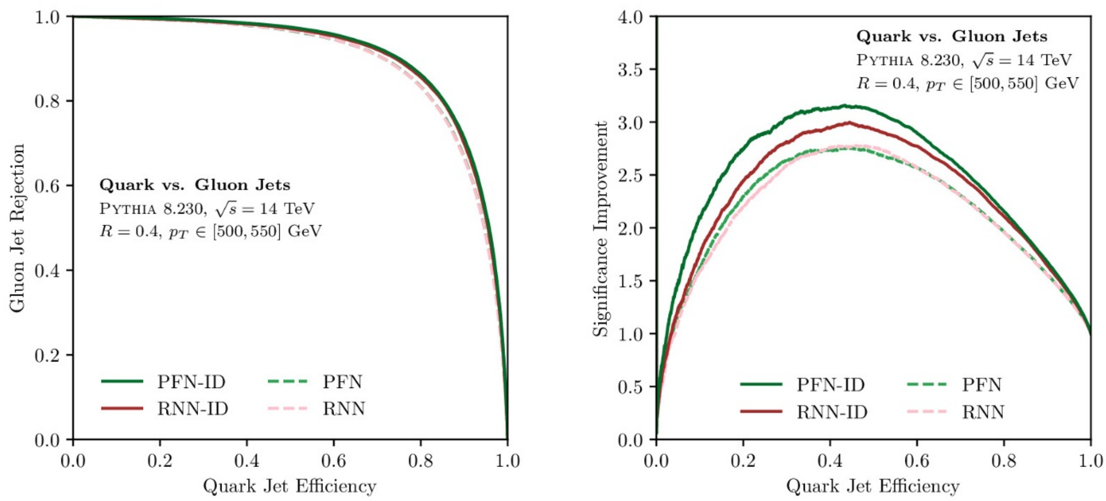
Besides comparing the EFN and PFN architectures to each other, we also compare the $\ell = 256$ models to a variety of other classifiers, summarized in table 2 and described in more detail in appendix A.

Of particular interest are the RNN-ID and RNN models, which also take particles as input (with and without full particle ID, respectively), but process them in a way which is dependent on the order the particles were fed into the network (decreasing p_T ordering was used). In figure 7, ROC and SI curves are shown for the RNN-ID and RNN architectures, as well as their natural counterparts, PFN-ID and PFN. We see that PFN-ID slightly outperforms RNN-ID whereas the PFN and RNN are comparable, though we emphasize that making broad conclusions based on this one result is difficult given the variety of different RNN architectures we could have chosen. Since PFNs are less expressive than RNNs, which can learn order-dependent functions, it is satisfying that both the PFN and RNN architectures achieve comparable classification performance with similar information.

The other machine learning architectures we compare to are a DNN trained on the N-subjettiness basis [24, 70, 71], a CNN trained on jet images [4, 5, 8], and a linear

Symbol	Name	Short Description
PFN-ID	Particle Flow Network w. ID	PFN with full particle ID
PFN-Ex	Particle Flow Network w. PF ID	PFN with realistic particle ID
PFN-Ch	Particle Flow Network w. charge	PFN with charge information
PFN	Particle Flow Network	Using three-momentum information
EFN	Energy Flow Network	Using IRC-safe information
RNN-ID	Recurrent Neural Network w. ID	RNN with full particle ID
RNN	Recurrent Neural Network	Using three-momentum information
EFP	Energy Flow Polynomials	A linear basis for IRC-safe information
DNN	Dense Neural Network	Trained on an N-subjettiness basis
CNN	Convolutional Neural Network	Trained on 33×33 grayscale jet images
M	Constituent Multiplicity	Number of particles in the jet
n_{SD}	Soft Drop Multiplicity	Probes number of perturbative emissions
m	Jet Mass	Mass of the jet

Table 2. The (top) PFN/EFN architectures, (middle) other machine learning models, and (bottom) jet substructure observables used in comparisons of quark/gluon discrimination performance, along with their corresponding symbols and short descriptions. A detailed discussion of model implementation and observable computation is given in appendix A.

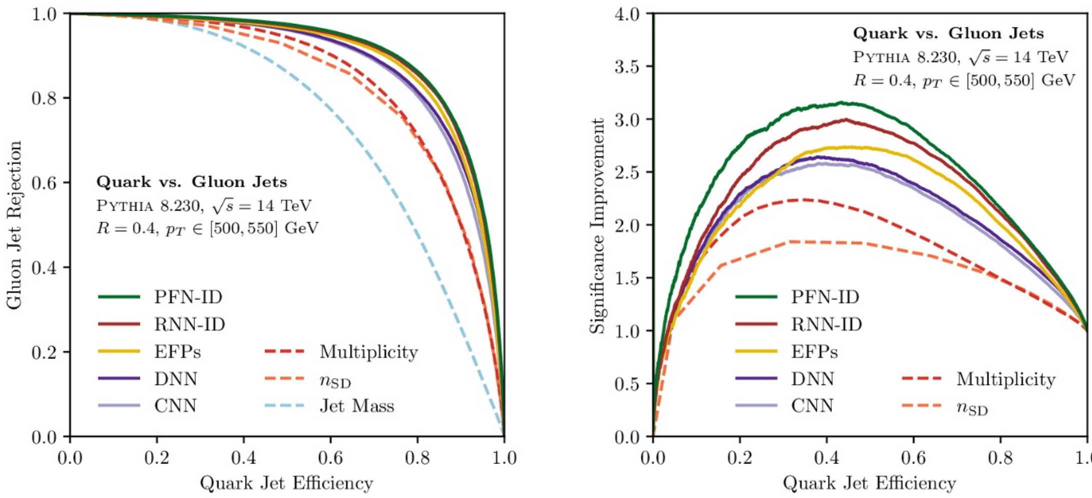


(a) **(b)**

Figure 7. The (a) ROC and (b) SI curve classification performances of PFN and RNN models both with and without full particle ID information. From the SI curve, it appears that the PFN-ID model is doing better than the RNN-ID model, whereas the PFN and RNN models perform roughly equally.

Model	AUC	$1/\varepsilon_g$ at $\varepsilon_q = 50\%$
PFN-ID	$\mathbf{0.9052} \pm 0.0007$	$\mathbf{37.4} \pm 0.7$
PFN-Ex	0.9005 ± 0.0003	34.7 ± 0.4
PFN-Ch	0.8924 ± 0.0001	31.2 ± 0.3
PFN	0.8911 ± 0.0008	30.8 ± 0.4
EFN	0.8824 ± 0.0005	28.6 ± 0.3
RNN-ID	0.9010	34.4
RNN	0.8899	30.5
EFP	0.8919	29.7
DNN	0.8849	26.4
CNN	0.8781	25.5
M	0.8401	19.0
n_{SD}	0.8297	14.2
m	0.7401	7.2

Table 3. The classification performances, quantified by the AUC and background rejection at 50% signal efficiency, for each of the models and observables in table 2. Reported uncertainties on the EFN and PFN family of models are half of the interquartile range over ten trainings. Performance uniformly improves with the inclusion of more particle-type information.



(a) (b)

Figure 8. The (a) ROC and (b) SI curve classification performances of several different models and observables. The PFN-ID and RNN-ID curves are shown in order to facilitate comparison with figures 6 and 7. The PFN-ID architecture compares well to existing techniques, often notably outperforming them.

classifier trained on the energy flow basis [29]. Their performance, as given by their AUC and background rejection at 50% signal efficiency, is summarized in table 3. Classification improves with the addition of IRC-unsafe information, as seen in the gain that the various PFN and RNN models have over the EFN and EFP models. There is also a boost in performance from providing the model with ever-more specific particle-type information.

Figure 8 shows ROC and SI curves for all of these models, as well as some common jet substructure observables. The best model is PFN-ID, followed by RNN-ID, and then (as shown in this figure) linear EFPs, which, somewhat remarkably, is the best architecture by AUC and SI curve height that does not take particles as direct inputs. We note that for the CNN, one can in principle include particle ID information via additional channels, though training a 14-channel CNN is computationally challenging as each channel comes with $O(1000)$ additional numbers, most of which are zero. Similar to RNNs and CNNs, the EFN and PFN architectures endeavor to be efficient by reducing the number of trainable parameters using weight sharing by applying the same Φ network to each particle. Adding particle-type information to the EFPs or the N-subjettiness DNNs might be possible through a suitable generalization of jet charge [72], though we know of no concrete implementation of this in the literature. The fact that PFNs naturally incorporate particle ID information is a important aspect of this architecture.

4.4 Visualizing the singularity structure of QCD

Beyond their excellent classification performance, the EFN and PFN architectures have the additional benefit that the learned function Φ can be directly explored. As discussed in section 2.3, this is particularly true of the EFNs, where $\Phi(\hat{p})$ is a two-dimensional function of the angular information and thus can be directly visualized in the rapidity-azimuth plane.

We take the learned $\Phi : \mathbb{R}^2 \to \mathbb{R}^\ell$ network from the best EFN model, as determined by the AUC, and evaluate it at many rapidity-azimuth points (y, ϕ) in the range $y, \phi \in [-R, R]$ to form a set of ℓ filters representing the learned latent space. We show several of these filters from the $\ell = 256$ EFN models in figure 9. These can be directly compared with the corresponding filters for the detector image representation in figure 2 and for the radiation moment representation in figure 3. Like the image representation, we see that the learned filters are localized bumps in the rapidity-azimuth plane, and thus we say that the model appears to have learned a "pixelization" of the rapidity-azimuth plane.[11] Unlike the image representation, the "pixels" learned by the model are smaller near the core of the jet and larger near the edge of the jet.

Beyond showing individual filters, it is informative to attempt to visualize an entire EFN latent space at once. We achieve this by finding the boundary of each of the learned pixels (corresponding to one component of Φ) and showing these boundaries together. Plotting the boundary contours simultaneously allows for a direct visualization of the latent space representation learned by the model on a single figure. In this way, we arrive at a proxy for the "image" that the model projects each jet into, which empirically emerges as a dynamically-sized calorimeter image. Larger latent space dimensions correspond roughly

[11]Note that the ReLU activation function that we used in the model forces the filter values to be positive and allows the model to easily turn off regions of the inputs. Different activation functions may result in different learned latent representations.

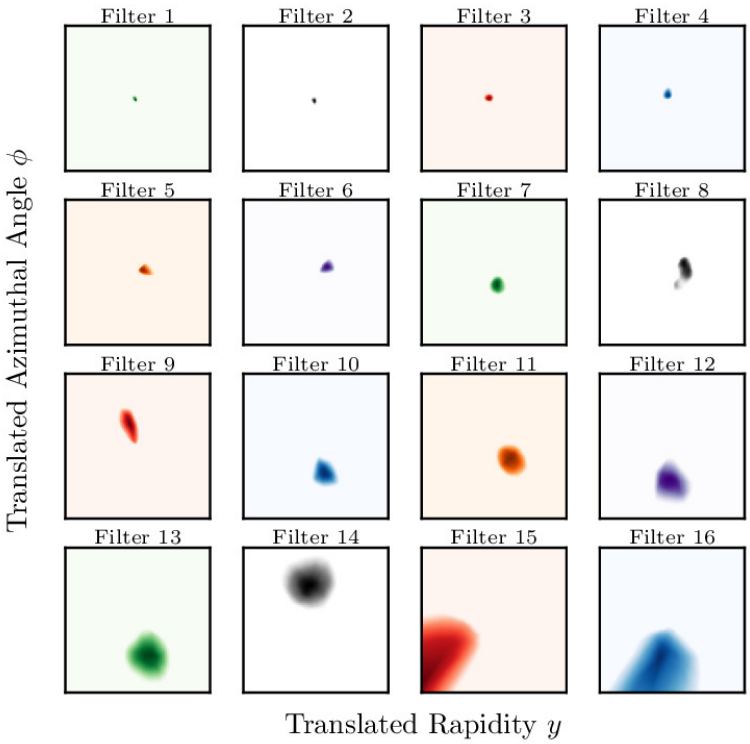


Figure 9. Visualizations of 16 of the 256 filters learned by the $\ell = 256$ EFN, with the filters sorted by their activated area. The domain is the rapidity-azimuth plane from $-R$ to R in both y and ϕ, since the jets have been preprocessed by centering them at (0,0). The localized nature of the filters leads to our interpretation that the model has learned an image-like "pixelization" of the rapidity-azimuth plane, albeit one that is not square as in figure 2.

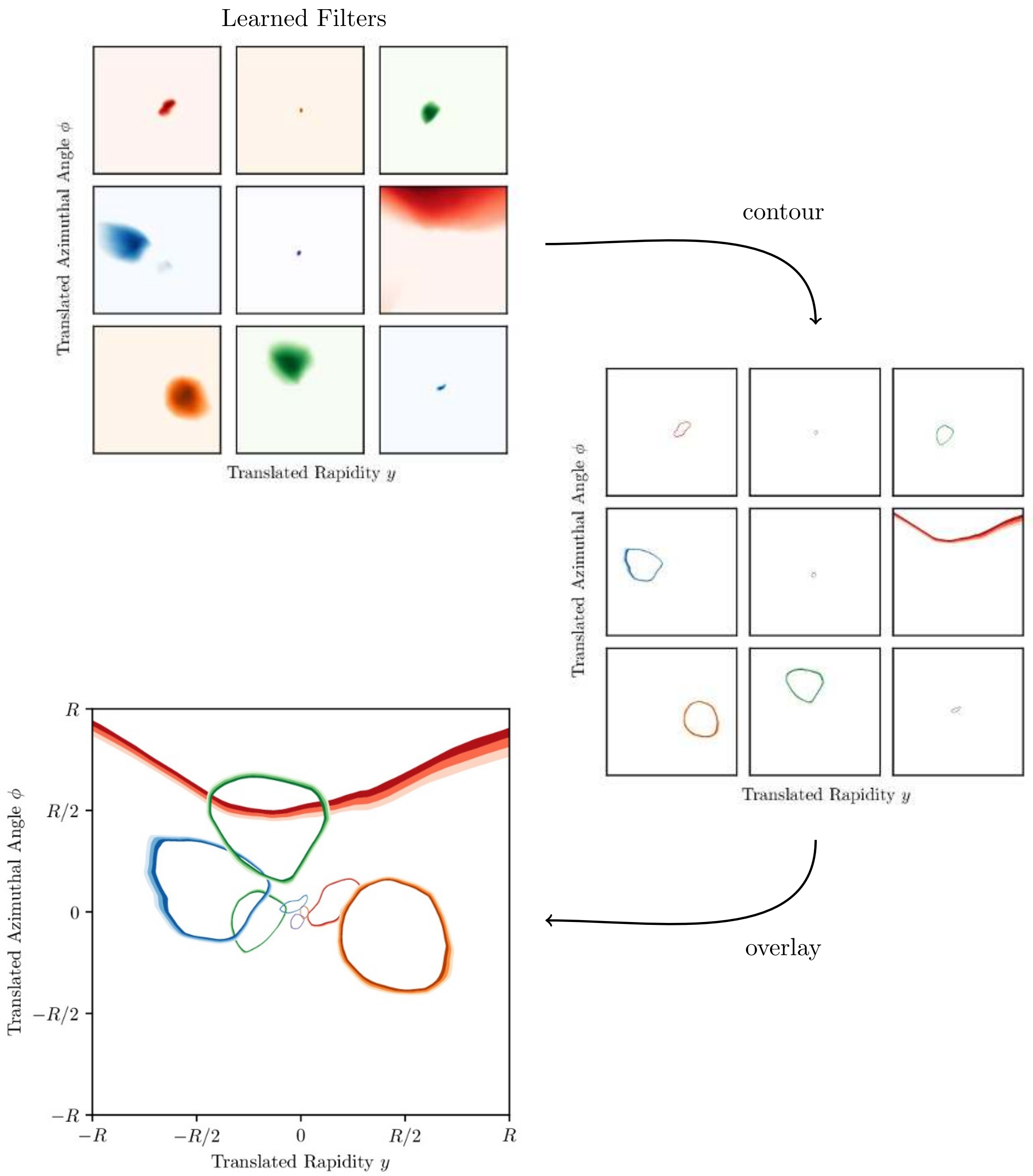


Figure 10. An illustration of our simultaneous visualization procedure for example EFN filters. Contours of each filter are shown from 45% to 55% of its maximum value. These contours are then overlaid on the same figure with different colors. The resulting contour plot shows the dynamical pixelization of the plane determined by the model.

to higher resolution images. This strategy is illustrated in figure 10, where each filter is contoured around its 50% value and the contours are overlaid.

In figure 11, we show this visualization for EFN models with latent dimension varying from 8 to 256 in powers of 2. Some of the filters are zero in the region of interest, perhaps as a result of dying ReLUs, so these are not shown. It is evident from the simultaneous overlay of the filters that their sizes are correlated with their distance from the origin, which is es-

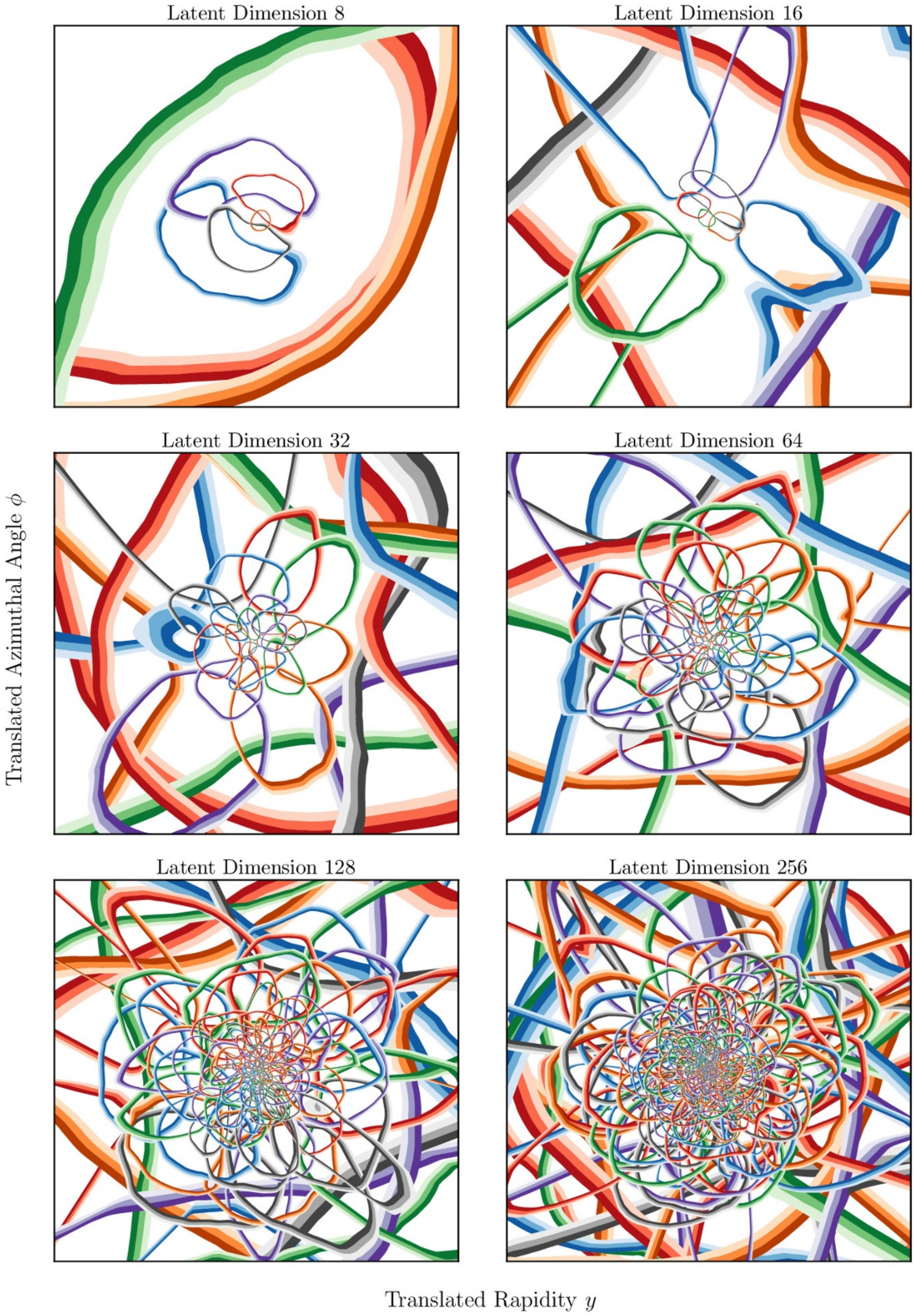


Figure 11. The learned EFN pixelization of the rapidity-azimuth plane around the jet center with latent dimensions between 8 and 256 in powers of 2. The learned filters are dynamically sized, with smaller filters probing the core of the jet and larger filters in the periphery. A large version of the last panel is shown in figure 21.

JHEP01(2019)121

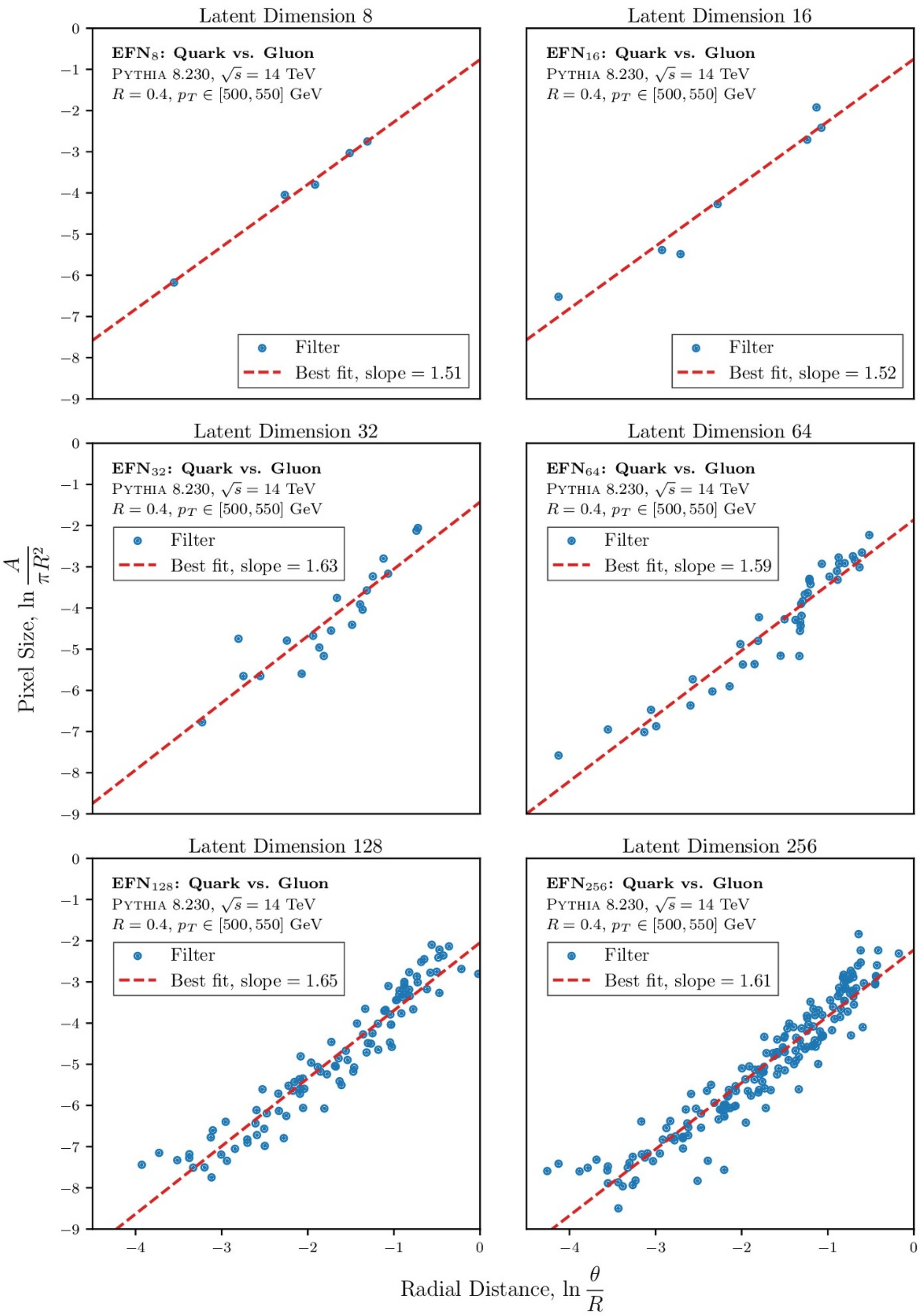


Figure 12. The size of the EFN filters as a function of their distance from the origin. The tendency of small filters to be located near the core of the jet and larger ones to be farther out is clearly visible. The best fit slope is around 2, which is the scale-invariant expectation from eq. (4.2).

pecially clear for the larger latent dimensions. As quark and gluon jets are (approximately) fractal objects with radiation singularly enhanced near the core of the jet as a result of the collinear singularity of QCD, the dynamically-sized pixelization learned by the EFN suggests that the model, in a sense, has understood this fact and adjusted itself accordingly.

To quantify the tendency of the filters to change size as they approach the center of the jet, we plot the area of each filter as a function of its distance from the origin. To define the area of a filter, A, we integrate the region of the rapidity-azimuth for which the filter is greater than or equal to half of its maximum. To capture a notion of distance from the origin, θ, we take the distance from the origin to the maximum value of the filter. We exclude filters that have centers outside of the jet radius. The resulting plots of the filters in this space are shown in figure 12 for the models with latent space dimension from 8 to 256 in powers of 2. There is a clear linear relationship between the (log) pixel size and the (log) distance to the jet core. In particular, the slope between $\ln A$ and $\ln\theta$ is around 1.6 in the cases studied.

We can attempt to understand why the slopes in figure 12 are around 2 by considering a uniform pixelization in $(\ln\frac{R}{\theta}, \varphi)$, where θ is the distance from the jet axis and φ is the azimuthal angle around the jet axis (not to be confused with ϕ). As discussed in ref. [14], this is the natural emission space of the jet. Translating an area element from this natural emission space to the rapidity-azimuth (y, ϕ) plane yields:

$$\left|\mathrm{d}\ln\frac{R}{\theta}\,\mathrm{d}\varphi\right| = \frac{\mathrm{d}\theta}{\theta}\,\mathrm{d}\varphi = \frac{\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi}{\theta^2} = \theta^{-2}\,\mathrm{d}y\,\mathrm{d}\phi. \tag{4.1}$$

Thus, a uniform pixelization in $(\ln\frac{R}{\theta}, \varphi)$ yields the following relationship between the area element (or pixel) size in the rapidity-azimuth plane and its distance from the origin:

$$\ln\frac{A}{\pi R^2} = 2\ln\frac{\theta}{R} + \text{const}, \tag{4.2}$$

explaining the slopes around 2 observed empirically in figure 12. This emergent behavior suggests an interesting connection with recent work on machine learning directly in the emission space of the jet [14]. Deviations from the scale-invariant expectation of 2 are largest near the core of the jet, where non-perturbative physics or axis-recoil effects [118] become important. The emission plane is visualized directly in appendix C, where the pixelization is indeed seen to be highly uniform and regular in that space.

4.5 Extracting new observables from the model

Given that we are able to examine Φ for a trained EFN by visualizing its components, we can attempt to go further and obtain a quantitative description of both Φ and F as closed-form observables. Obtaining novel jet substructure observables from machine learning methods has been approached previously by parameterizing an observable and learning the optimal parameters for a particular task [28]. Here, we go in a different direction and look directly at the latent observables learned by an EFN. This represents a first, concrete step towards gaining a full analytic understanding what is being learned by the model.

To make this tractable, we focus on the simple case of a two-dimensional latent space. A trained $\ell = 2$ EFN has two learned filters, $\Phi_1(y, \phi)$ and $\Phi_2(y, \phi)$, and a learned function

JHEP01(2019)121

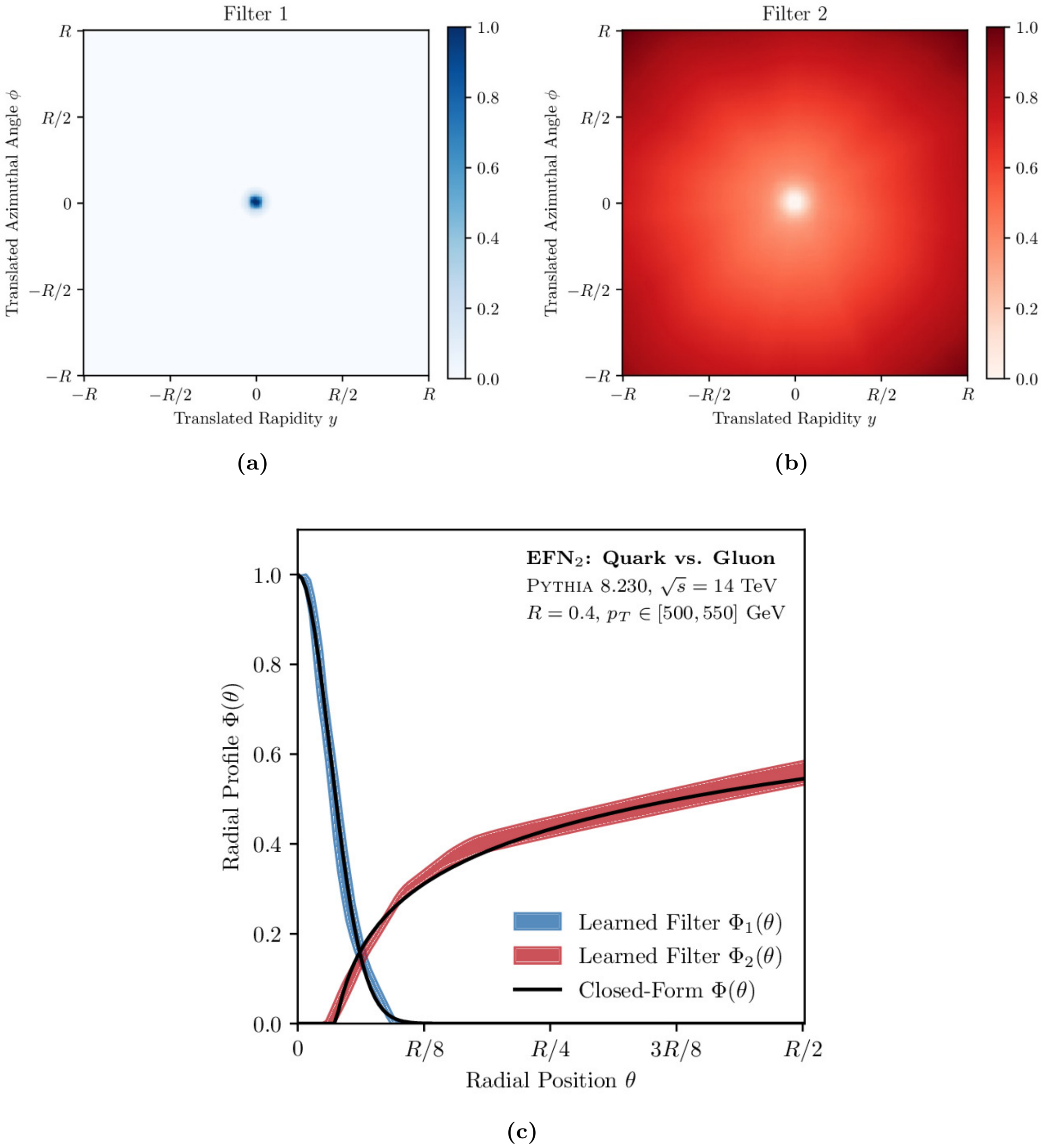
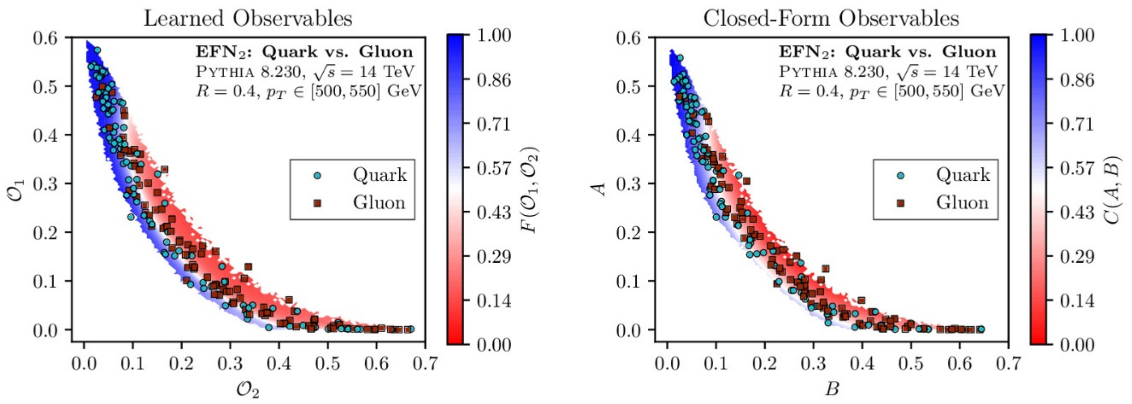


Figure 13. (a, b) The two filters learned by an $\ell = 2$ EFN, normalized to have a maximum value of 1. The rotational symmetry of the filters about the jet axis is evident, with one filter probing radiation near the core of the jet and the other probing wide-angle radiation. (c) Radial behavior of the two filters, from the center of the jet along the vertical and horizontal directions. The analytic forms of eq. (4.4) are shown as black lines, with $B_{r_1,\beta}$ scaled by 0.15.

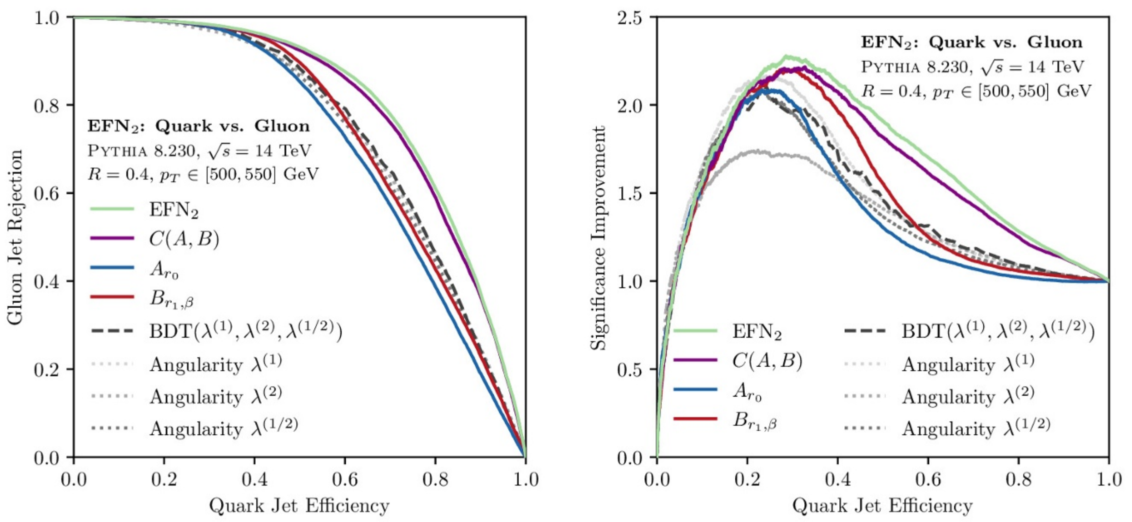
$F(\mathcal{O}_1, \mathcal{O}_2)$. The filters can be visualized in the rapidity-azimuth plane and the function F can be viewed in the $(\mathcal{O}_1, \mathcal{O}_2)$ phase space. By studying these visualizations and noting their emergent properties, we can construct observables that reproduce the behavior and predictive power of the trained EFN.

JHEP01(2019)121



(a) **(b)**

Figure 14. (a) The EFN model output $F(\mathcal{O}_1, \mathcal{O}_2)$ in the plane spanned by the learned latent space observables $\mathcal{O}_1$ and $\mathcal{O}_2$. (b) The closed-form function $C(A, B)$ in the plane of the analytic observables A_{r_0} and $B_{r_1,\beta}$. One hundred quark jets (light blue circles) and gluon jets (dark red squares) are indicated to highlight the separation power. The distribution of the closed-form observables and output value faithfully reproduce those of the trained EFN.



(a) **(b)**

Figure 15. The (a) ROC and (b) SI curves for the two closed-form observables A_{r_0} and $B_{r_1,\beta}$ as well as their combination $C(A, B)$, compared to the trained $\ell = 2$ EFN model. Three angularities are also shown for comparison, along with their corresponding performance when combined with a BDT. While the two learned observables perform similarly to the angularities on an individual basis, they are evidently more informative than the angularities when combined. The output of the trained EFN model and the closed-form estimate achieve similar performance.

In figures 13a and 13b, we show the learned filters Φ_1 and Φ_2 of a trained $\ell = 2$ EFN. It is evident that the filters exhibit approximate radial symmetry, with one of the filters concentrated at the center of the jet and the other activated at larger angular distances. Thus, we can restrict our attention to functional forms which depend only on the rapidity-azimuth distance θ from the origin. In particular, due to its built-in IRC-safety, the EFN model has learned filters that correspond to observables of the following approximate form:

$$\mathcal{O}_1 = \sum_{i=1}^{M} z_i\, \Phi_1(\theta_i), \qquad \mathcal{O}_2 = \sum_{i=1}^{M} z_i\, \Phi_2(\theta_i). \tag{4.3}$$

These are of the general form of IRC-safe angularities [74] with a generic radially-symmetric angular weighting function [85].[12] To quantify the filters further, in figure 13c we plot the value of the learned filters as a function of the radial distance, taking an envelope over several radial slices. The complementary central and wide-angle nature of the two filters are clearly evident.

By observing the properties of the curves in figure 13c, we fit two IRC-safe observables to the learned profiles of the following forms:

$$A_{r_0} = \sum_i z_i\, e^{-\theta_i^2/r_0^2}, \qquad B_{r_1,\beta} = \sum_i z_i \ln(1 + \beta(\theta_i - r_1))\, \Theta(\theta_i - r_1), \tag{4.4}$$

with values of $r_0 = 0.018$, $\beta = 200$, and $r_1 = 0.015$ found to be suitable. The observables in eq. (4.4) are then multiplied by overall factors of 0.60 and 0.18, respectively, to match the arbitrary normalization of the learned filters. While the precise values and shapes of the observable profiles changed from training to training, these general forms emerged for several of the best-performing models.

The observables A_{r_0} and $B_{r_1,\beta}$ in eq. (4.4) are IRC-safe angularities with a linear energy dependence and interesting angular weighting functions. A_{r_0} probes the collinear radiation near the core of the jet at angles $\theta \lesssim r_0$, and $B_{r_1,\beta}$ probes wide-angle radiation away from the core of the jet at angles $\theta > r_1$. The separate treatment of collinear and wide-angle particles is unlike the behavior of the traditional angularities, which have explicit contributions from both collinear and wide-angle regions of phase space. Though, as will be shown, each is individually a comparable quark/gluon jet classifier to the traditional angularities, the model is able to combine them in such a way as to achieve a significantly better performance. It would be interesting to perform a first-principles QCD study to understand in what sense the separation of collinear and wide-angle behavior is beneficial for discrimination.

We now proceed to obtain a closed-form estimate of the learned function $F(\mathcal{O}_1, \mathcal{O}_2)$. In figure 14a, we populate the $(\mathcal{O}_1, \mathcal{O}_2)$ phase space by quark and gluon jets and color most of the populated space (with a mild threshold against outliers) according to the value of F at that phase space point. We also indicate the truth labels of one hundred quark and gluon jets to guide the eye. Based on the transition from red to blue, we can see that the model selects a curved slice through the populated phase space region to obtain

[12]Following ref. [76], this section could alternatively be titled "New angles on IRC-safe angularities."

its predictions.[13] Correspondingly, we choose a closed-form estimate based on the simple (squared) Euclidean distance in phase space from a reference point (a_0, b_0), namely:

$$C(A, B) = (A - a_0)^2 + (B - b_0)^2, \tag{4.5}$$

with $a_0 = 0.445$ and $b_0 = 0.545$, where, for ease of visualization, a sigmoid is also applied to $50\,(C(A,B) - 0.460^2)$ to monotonically rescale the predictions between 0 and 1 and approximate the value of the learned function. We visualize $C(A, B)$ in figure 14b, finding a satisfying correspondence between the closed-form observables and the learned case.

Finally, we study the classification performance of these closed-form observables, compared to the trained $\ell = 2$ EFN model and several IRC-safe angularities. We consider three IRC-safe angularities $\lambda^{(\beta)} = \sum_i z_i \theta_i^\beta$ with $\beta = 2$, $\beta = 1$, and $\beta = 1/2$, which are the jet mass, jet width, and Les Houches angularity, respectively [87]. ROC curves for the various observables are shown in figure 15. As single observables, the closed-form observables A_{r_0} and $B_{r_1,\beta}$ of eq. (4.4) perform similarly to the individual angularities. When combined via $C(A, B)$ of eq. (4.5), they outperform all of the single angularities and approach the trained $\ell = 2$ EFN performance. By contrast, when the three angularities are combined with a Boosted Decision Tree (BDT), there is not a significant improvement in the classification performance.[14] The fact that A_{r_0} and $B_{r_1,\beta}$ can be combined to significantly increase the discrimination power indicates that, unlike the considered angularities, the new closed-form observables probe complementary information.

5 Conclusions

In this paper, we presented a new technique to learn from collider events in their natural representation as sets of particles. Our approach relies on the Deep Sets theorem [63], which guarantees that a generic symmetric function can be represented by an additive latent space. In the context of particle-level collider observables, each particle is mapped to a latent representation and then summed over, and observables are then functions on that latent space. This decomposition encompassed a wide variety of existing event- and jet-level collider observables and representations, including image-based and moment-based methods.

While these observable decompositions are interesting in their own right, parameterizing them with neural networks yields a machine learning framework ideally suited for learning from variable-length unordered lists of particles. We proposed two fundamental network variants. The IRC-safe EFNs treat each latent space observable as an energy-weighted function of geometry, ensuring IRC safety in the latent space by construction. The fully-general PFNs are able to incorporate additional particle-level information such as charge and flavor, maximizing the information available to achieve collider tasks. We showcased the

[13] A similar strategy of fitting analytic functions to learned decision boundaries was carried out in ref. [119].

[14] The BDT classifier was implemented using scikit-learn [120] using 100 estimators, trained and tested on 100,000 quark and gluon jets with a 50-50 train/test split. We also tested combining A_{r_0} and $B_{r_1,\beta}$ with a BDT, which resulted in similar performance to the closed-form and learned cases, verifying that the information is being adequately captured by $C(A, B)$.

JHEP01(2019)121

efficacy of these method in a quark/gluon discrimination case study, achieving favorable performance compared with many existing machine learning techniques for particle physics.

A fascinating aspect of the EFNs is that one can directly visualize the per-particle mapping into the latent space. This allowed us to peer inside the network and discover that, in the context of quark/gluon discrimination, it had learned a dynamically-sized pixelization of the rapidity-azimuth plane. This pixelization was reminiscent of a jet image but did not adhere to the strict rectilinear grid imposed by traditional jet images. We also found some compelling evidence that the model had "understood" a key property of QCD, specifically its famous collinear singularity structure, since filters closer to the center of the jet were more finely resolved. The relationship between the size and position of the filters could be roughly understood as a uniform pixelization in the natural angular emission plane.

We also presented a simple example of directly learning the physics used by the trained model and obtaining new closed-form observables. The IRC-safe Observable Decomposition, combined with the visualizability of the EFN filters, provides a general way to access what the model has learned. In the case of an $\ell = 2$ EFN, the model learned to separately probe collinear and wide-angle radiation and then use a phase space distance to classify quarks and gluons. While the quark/gluon classification performance of A_{r_0} and $B_{r_1,\beta}$ is not yet competitive with other simple observables such as the constituent multiplicity, together they notably outperform other IRC-safe angularities. Even though our numerical choices of parameters are specific to these quark/gluon jet samples, it may be interesting to explore the A_{r_0} and $B_{r_1,\beta}$ classes of observables further in the context of theoretical efforts to jointly explore and understand the correlations between two angularities [74, 121, 122].

We conclude by discussing possible extensions and additional applications of these methods. Pileup or underlying event mitigation could be an interesting avenue to explore with this architecture, similar to ref. [30] but also to PUPPI [123] in that a per-particle correction factor could be applied. Such an effort would need to make use of the equivariant designs of ref. [63], which have a separate output for each input. One may also consider adding high-level information — such as the jet four-momentum, individual substructure observables, or clustering histories — directly to Φ or F in order to improve the network performance. Further, while we used jet-level examples for our case study, the problem of learning functions of sets applies at the event level more broadly. At the event level, new challenges arise that the simple preprocessing performed here for the jet case studies may not completely solve. An iterative deep sets technique, where multiple latent spaces are constructed, may help to capture the local/global intuition that collider physicists regularly employ to study the hierarchical relationships between events, jets, and particles.

Acknowledgments

We are grateful to Kyle Cranmer, Frédéric Dreyer, Felice Frankel, Philip Harris, Michael Kagan, Gregor Kasieczka, Markus Klute, Andrew Larkoski, Sung Hak Lim, Benjamin Nachman, Siddharth Narayanan, Daniel Roberts, Daniel Whiteson, Michael Williams, and Manzil Zaheer for helpful conversations. We are also grateful to Gregor Kasieczka for providing us with the top samples used in appendix B. We thank the Harvard Center for the Fundamental Laws of Nature for hospitality while this work was completed. This work was supported by the Office of Nuclear Physics of the U.S. Department of Energy (DOE) under grant DE-SC-0011090 and by the DOE Office of High Energy Physics under grant DE-SC-0012567. JT is supported by the Simons Foundation through a Simons Fellowship in Theoretical Physics. Cloud computing resources were provided through a Microsoft Azure for Research award.

A Details of models and observables

In this appendix, we describe the details of the machine learning models and observables. All of the models employed for comparison with the EFN/PFN models were applied to our quark/gluon samples in section 4, but only the linear EFP model will be applied to the top samples in appendix B. For the substructure observables, the jet mass m and constituent multiplicity n_{const} are easily obtained from the jet in FastJet. Using FastJet contrib 1.033 [124], the RecursiveTools 2.0.0-beta1 module was used to calculate the soft drop multiplicity n_{SD} [78] with parameters $\beta = -1$, $z_{\text{cut}} = 0.005$, and $\theta_{\text{cut}} = 0$. The Nsubjettiness 2.2.4 module was used to calculate all N-subjettiness values $\tau_N^{(\beta)}$ [70, 71]. All neural networks were implemented in Keras [125] with the TensorFlow [126] backend and trained on NVIDIA Tesla K80 GPUs on Microsoft Azure.

Here, we provide details for the machine learning models listed in table 2:

- **EFN**, **PFN**: both the EFN and PFN architectures are contained as part of our `EnergyFlow` Python package [91], which contains examples demonstrating their training, evaluation, and use. Keras requires contiguous NumPy [127] arrays as input, so the events are padded with all-zero particles to ensure they have equal length. This zero-padding is a technical trick, not a conceptual limitation. It is distinct from the zero-padding necessary to use variable-length inputs in a DNN. This operation can be done on a per-batch level, but to avoid repeating this procedure at each epoch, we zero-pad all the events once at the beginning. The Keras `Masking` layer is used to ignore all-zero particles during the training and evaluation of the network. The Keras `TimeDistributed` layer is used to apply the function Φ to each particle, which in the relevant language is a "time step". To carry out the summation in the latent space, we use the Keras `Dot` layer for the EFN architecture, to product and sum the transverse momentum fractions with the latent observables, and the `Lambda` layer for the PFN architecture, to sum over the particles. It may be interesting to consider alternative "pooling" functions to summation, such as max-pooling or average-pooling as in refs. [60, 64].

- **RNN-ID**, **RNN**: recurrent Neural Networks provide an alternative way to process variable-length inputs, albeit one that is explicitly not agnostic to the order of the inputs. We choose to order the particles by their transverse momenta and train two variants: one with only kinematic information (RNN) and one with kinematic as well as full particle ID information (RNN-ID). The former should be compared with the PFN model and the latter to the PFN-ID model. We did not consider alternative particle orderings for the RNN, though other investigations have found that performance is robust to choices in particle ordering [16, 44]. Our RNN architecture consists of an `LSTM` layer with 64 nodes (the performance was insensitive to changing this value to 128 or 256) followed by three fully-connected layers each with 100 nodes and a fully-connected output layer with 2 nodes. Due to the `Masking` layer employed, the significantly faster `CuDNNLSTM` layer was not used and the batch size was taken to be 2000 to help speed up training. An architecture using two or three `SimpleRNN` layers was also tried but the performance was not as good as with an `LSTM` layer. The long training time of the RNN models was prohibitive in exploring additional hyperparameters.

- **EFPs**: the energy flow basis [29] is a linear basis for the space of IRC-safe jet substructure observables, allowing linear methods to be applied to a set of Energy Flow Polynomials with good performance. The `EnergyFlow` 0.8.2 Python package [91] was used to compute all EFPs with degree $d \leq 7$ and complexity $\chi \leq 3$, using the normalized default hadronic measure with $\beta = 0.5$. The same EFPs with $\beta = 1$ were also tested and found to perform slightly worse. These 996 EFPs, including the trivial constant EFP, were used to train a Fisher's Linear Discriminant model from the scikit-learn package [120].

- **DNN**: the N-subjettiness basis [24] is a K-body phase space basis consisting of the following $3K - 4$ observables:

$$\left\{\tau_1^{(1/2)},\, \tau_1^{(1)},\, \tau_1^{(2)}, \tau_2^{(1/2)},\, \tau_2^{(1)},\, \tau_2^{(2)},\, \ldots,\, \tau_{K-2}^{(1/2)},\, \tau_{K-2}^{(1)},\, \tau_{K-2}^{(2)}, \tau_{K-1}^{(1/2)},\, \tau_{K-1}^{(1)}\right\}. \qquad \text{(A.1)}$$

  We use $K = 25$ and, following ref. [24], compute the N-subjettiness values with respect to k_T axes. A Dense Neural Network consisting of three fully-connected layers each with 100 nodes and one fully-connected output layer with 2 nodes was trained on this set of N-subjettiness values.

- **CNN**: following ref. [8], we compute 33×33 one-channel (grayscale) jet images in a $2R \times 2R$ patch of the rapidity-azimuth plane. Images were preprocessed as in refs. [8, 90] by subtracting the mean image of the training set and dividing by the per-pixel standard deviation. A Convolutional Neural Network architecture was trained on the jet images consisting of three convolutional layers with 48, 32, and 32 filters of size 8×8, 4×4, and 4×4, respectively. These layers were followed by two fully-connected layers each with 50 nodes and one fully-connected output layer with 2 nodes. Maxpooling of size 2×2 and dropout with a rate of 0.1 was implemented after each convolutional layer.

For all quark/gluon models, one million jets were used for training, 200k for validation (except for the EFPs, which did not use validation), and 200k for testing. Common properties to all neural networks include the use of ReLU [108] activation functions for all non-output layers, a softmax activation function on the 2 node output layer, He-uniform

JHEP01(2019)121

Model	Time per Epoch (sec)	Number of Epochs
PFN-ID	73	47
PFN-Ex	67	37
PFN-Ch	64	51
PFN	71	34
EFN	74	54
RNN-ID	192	61
RNN	188	96
DNN	9	39
CNN	63	22

Table 4. Median times per epoch and numbers of epochs during training of the considered machine learning models. The linear EFP model is not shown because it is trained by a different technique. Note the substantially longer training times for our RNN implementations compared to the EFN and PFN models.

weight initialization [109], the categorical crossentropy loss function, the Adam optimization algorithm [128], and a learning rate of 0.001. All non-RNN networks had a batch size of 500; all non-CNN networks had a patience parameter of 8 with the CNN having a patience parameter of 5. Each non-EFN/PFN model was trained twice and the model with the higher AUC on the test set is reported (except when ε_g at $\varepsilon_q = 50\%$ values are reported, in which case the model with the value of this statistic is reported). In table 4, we report the typical training times for each of the models.

B Tagging top jets

In this appendix, we train EFN and PFN models to discriminate top jets from the QCD jet background to provide an additional demonstration of the excellent performance of these models. Since top jets do not have any singularities associated with radiation about their center, training EFN models on tops provides an important cross check about our conclusions in section 4.4 for how the model is learning to pixelize the rapidity-azimuth plane.

The top and QCD jets used in this study are based on the dataset used in ref. [22], which were provided to us by those authors. The jets are Pythia-generated, anti-k_T, $R = 0.8$ jets at $\sqrt{s} = 14\,\text{TeV}$ with no underlying event or pileup using a Delphes [129] detector fast-simulation of the ATLAS detector. Jets are kept if they have transverse momentum $p_T \in [550, 650]\,\text{GeV}$ and pseudorapidity $|\eta| < 2$ and if they can be matched to a top parton to within a rapidity-azimuth distance of $\Delta R = 0.8$. The jets are required to be fully merged, with the three quarks from the top decay contained within the jet radius. See ref. [22] for additional information about the samples and their generation details, as well as information about other top-tagging algorithms. There are 1.2 million training events, with 400k validation events and 400k test events. The jet samples do not contain particle-level

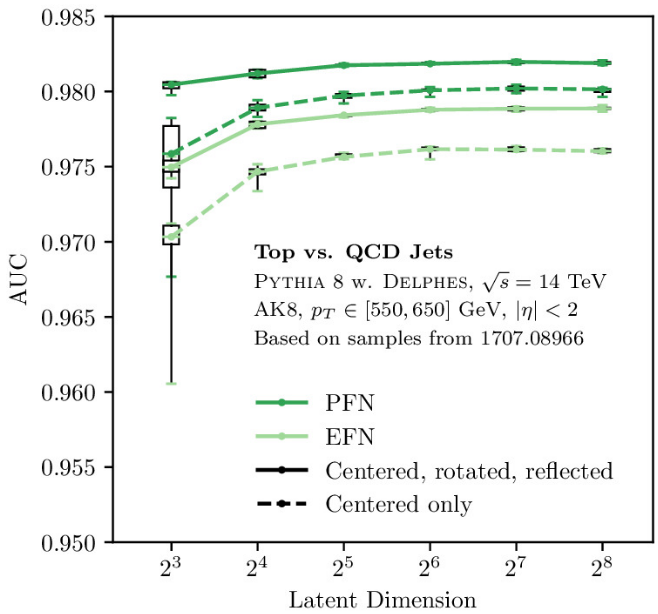
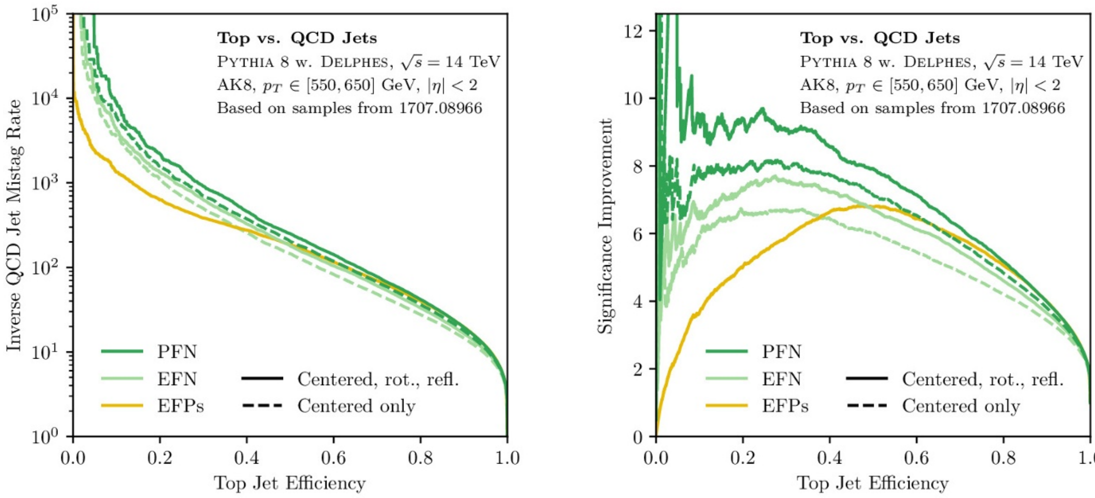


Figure 16. The AUC performance of the EFN/PFN top tagging models as a function of the latent dimension, which is varied from 8 to 256 in powers of 2. Including rotation and reflection in the rapidity-azimuth plane as preprocessing steps (solid curves) improves the model performance significantly compared to only centering the jets (dashed curves). The spread in values is due to retraining the models ten times.

flavor or charge information, and thus we can only train EFN and PFN models that make use of kinematic information alone. A linear EFP model is also trained for comparison.

Given the different topology typical of top jets compared to quark or gluon jets, we implement some additional preprocessing steps designed to help the model train more efficiently. As with the quark and gluon jets of section 4, we center all of the jets in the rapidity-azimuth plane based on the four-momentum of the jet, and we normalize the transverse momenta of the particles to sum to one. Models were trained with just this minimal preprocessing, as well as with additional rotation and reflection (r.r.) operations. For the EFN-r.r. and PFN-r.r. models, rotations were performed to align the leading eigenvector of the two-dimensional moment of the radiation pattern along the vertical axis, and reflections were performed about the horizontal and vertical axes to place the highest-p_T quadrant in a consistent quadrant.

The results of training EFN and PFN models for top tagging are shown in figure 16 with the latent dimension of the models varying from 8 to 256 in powers of 2. ROC and SI curves of the trained models are shown in figure 17 and also compared to the linear EFP model. Performance competitive with the results in ref. [22] is achieved, particularly for the PFN models. The preprocessing step of rotating and reflecting can be seen to notably improve both the EFN and PFN classification performance. In order to facilitate future comparisons with other models trained on these samples, in table 5 we report the AUC and gluon background rejection factor at both 0.3 and 0.5 quark efficiency.



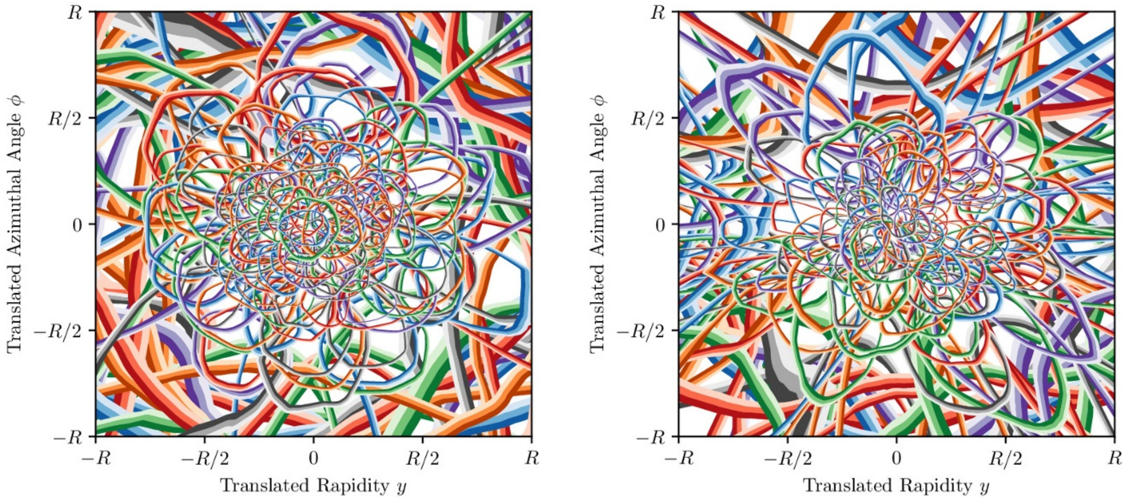
(a) **(b)**

Figure 17. The (a) inverse ROC and (b) SI curves for each of the EFN/PFN top tagging models with latent dimension 256, compared to the EFP linear model. The curve with the median AUC out of ten independent trainings is shown. As expected, the PFN yields better performance than the EFN, with the additional rotation/reflection preprocessing steps providing a further boost in performance.

Model	**AUC**	$\mathbf{1/\varepsilon_b}$ **at** $\boldsymbol{\varepsilon_s = 50\%}$	$\mathbf{1/\varepsilon_b}$ **at** $\boldsymbol{\varepsilon_s = 30\%}$
PFN-r.r.	$\mathbf{0.9819} \pm 0.0001$	$\mathbf{247} \pm 3$	$\mathbf{888} \pm 17$
PFN	0.9801 ± 0.0001	203 ± 1	732 ± 21
EFN-r.r.	0.9789 ± 0.0001	181 ± 2	619 ± 23
EFN	0.9760 ± 0.0001	143 ± 2	481 ± 12
EFPs	0.9803	184	384

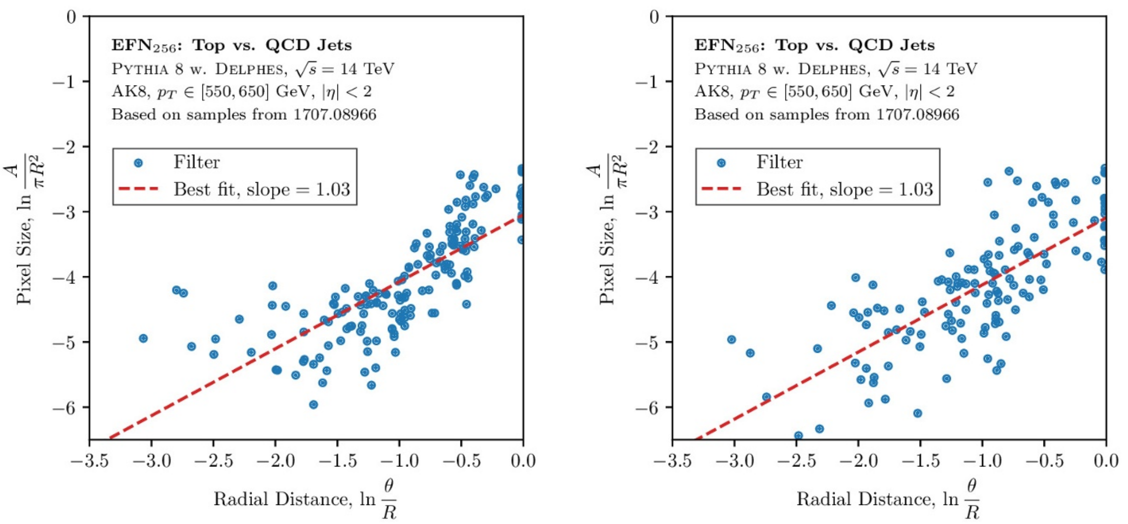
Table 5. Quantified classification performance (AUC, $1/\varepsilon_b$ at $\varepsilon_s = 0.5$, $1/\varepsilon_b$ at $\varepsilon_s = 0.3$) for each of the models trained in figure 17. Reported uncertainties are half of the interquartile range over ten trainings. The PFNs achieve the best performance, with improvements seen by including the rotation and reflection preprocessing. The EFPs slightly outperform the EFNs in AUC and background rejection at 0.5 signal efficiency but perform more poorly in background rejection at 0.3 signal efficiency.

JHEP01(2019)121



(a) **(b)**

Figure 18. Visualization of the learned filters of a top tagging $\ell = 256$ EFN (a) without any additional preprocessing and (b) with additional rotation/reflection. In (b), the rotational symmetry around the center is broken compared to (a).



(a) **(b)**

Figure 19. The filter sizes plotted as a function of their distance from the jet center for a top tagging $\ell = 256$ EFN (a) without any additional preprocessing and (b) with additional rotation/reflection. Though a general trend is seen that pixels closer to the center are smaller, the points are generally not as well fit by a line as in the quark/gluon case in figure 12.

JHEP01(2019)121

For the EFNs, we can also visualize the learned filters using the technique of section 4.4. The resulting filter visualizations for latent dimension 256 are shown in figure 18. The learned filters have some tendency to be smaller near the center and larger near the periphery, but not nearly as much as in the quark/gluon discrimination case in figure 10. This is expected because given the typical three-prong topology of a boosted top jet, the jet axis does not have any distinguished radiation associated with it, unlike for a QCD jet where the jet axis tends to lie along a core of radiation. To quantify this effect, we look at the size of the filters as a function of their distance from the origin, shown in figure 19. The relationship in the top tagging case is much weaker than the linear relationship present in the quark/gluon discrimination study, with significantly worse linear fits.

C Additional visualizations

For the quark/gluon EFN model in section 4, the observed relationship between the size and location of the filters qualitatively (in figure 11) and quantitatively (in figure 12) suggested that the model learned a uniform pixelization in the $\left(\ln \frac{R}{\theta}, \varphi\right)$ emission plane. (The top tagging EFN model in appendix B did not exhibit as clear of a relationship in either figure 18 or figure 19.) To directly visualize the EFN filters in the appropriate space, we implement a change of variables from Cartesian (y, ϕ) coordinates to polar (θ, φ) coordinates. The visualizations in figure 20 use the same contouring and overlaying technique of figure 10 to demonstrate the roughly uniform pixelation in the emission plane for the quark/gluon EFN models with latent dimensions of 16, 64, and 256. In this emission plane, the learned filters can be seen to be much more uniform in size and location compared to the rapidity-azimuth plane. The uniformity as a function of φ indicates that the set of filters approximately has rotational symmetry. We also checked that the model output typically changed by less than 0.1 after applying a random rotation about the jet axis in the rapidity-azimuth plane.

Finally, we show a larger version of the quark/gluon $\ell = 256$ EFN model from figure 11 in figure 21, suitable for framing.

JHEP01(2019)121

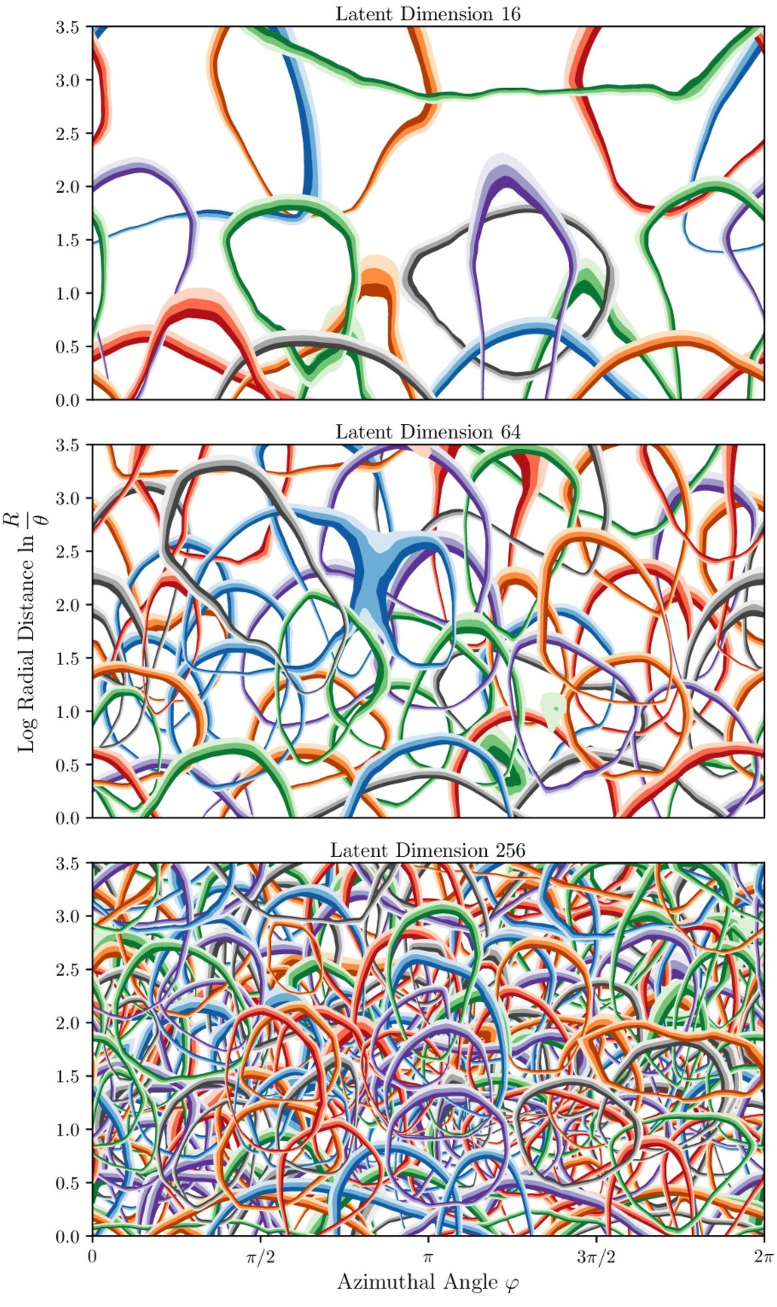


Figure 20. Visualization of the learned filters of a quark/gluon EFN in the $(\ln\frac{R}{\theta}, \varphi)$ plane with latent dimensions of (top) 16, (middle) 64, and (bottom) 256. The pixelization learned by the network is much more regular and uniform in this space compared to the rapidity-azimuth plane in figure 11.

JHEP01(2019)121

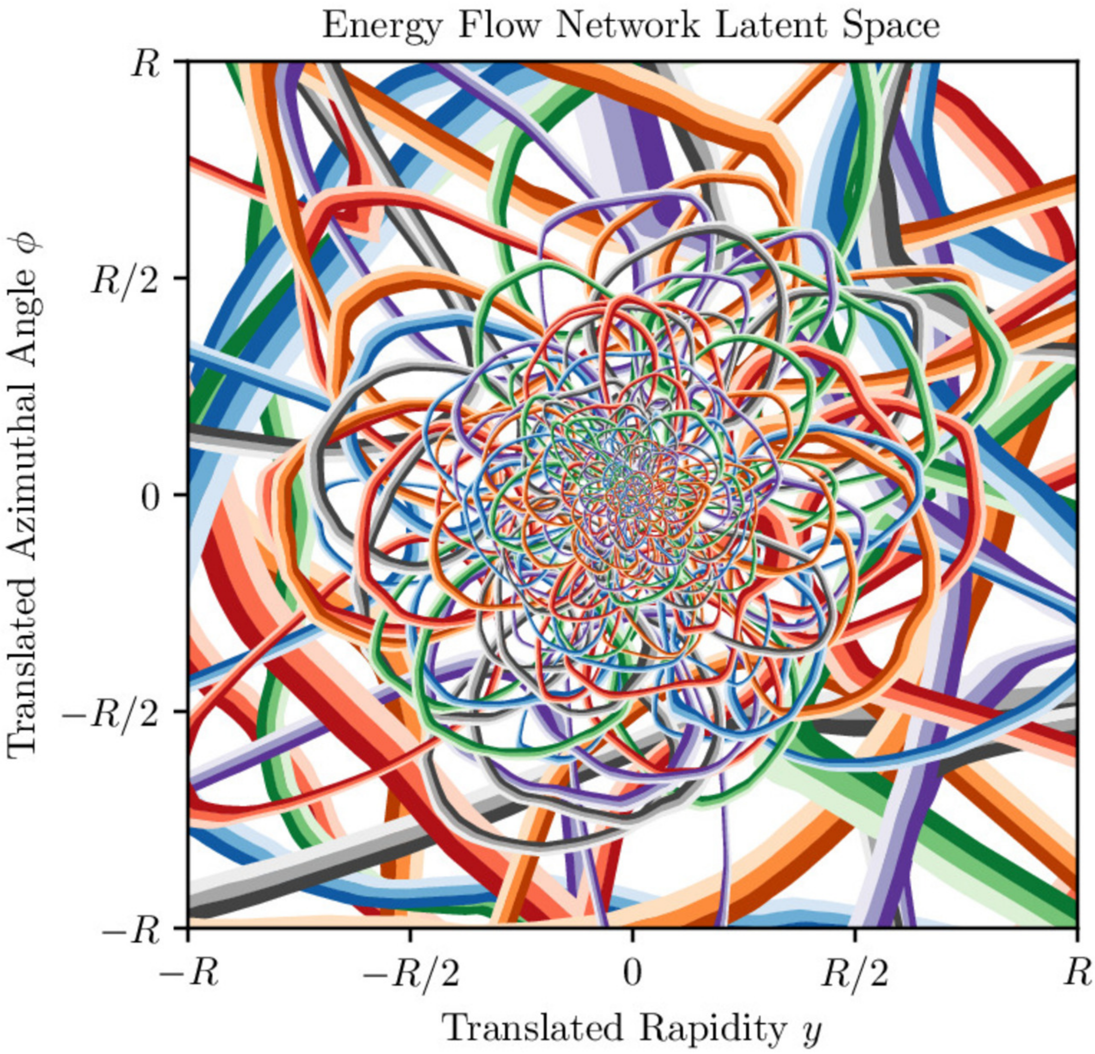


Figure 21. A larger scale visualization of the filters for the quark/gluon $\ell = 256$ EFN model, shown originally in figure 11.

References

[1] CMS collaboration, *The CMS particle flow algorithm*, in *Proceedings, International Conference on Calorimetry for the High Energy Frontier (CHEF* 2013*)*, Paris, France, 22–25 April 2013, pg. 295 [arXiv:1401.8155] [INSPIRE].

[2] CMS collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, 2017 *JINST* **12** P10003 [arXiv:1706.04965] [INSPIRE].

JHEP01(2019)121

[3] ATLAS collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, *Eur. Phys. J.* **C 77** (2017) 466 [arXiv:1703.10485] [INSPIRE].

[4] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-images: computer vision inspired techniques for jet tagging*, *JHEP* **02** (2015) 118 [arXiv:1407.5675] [INSPIRE].

[5] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, *Jet-images — deep learning edition*, *JHEP* **07** (2016) 069 [arXiv:1511.05190] [INSPIRE].

[6] P. Baldi, K. Bauer, C. Eng, P. Sadowski and D. Whiteson, *Jet substructure classification in high-energy physics with deep neural networks*, *Phys. Rev.* **D 93** (2016) 094034 [arXiv:1603.09349] [INSPIRE].

[7] J. Barnard, E.N. Dawe, M.J. Dolan and N. Rajcic, *Parton shower uncertainties in jet substructure analyses with deep neural networks*, *Phys. Rev.* **D 95** (2017) 014018 [arXiv:1609.00607] [INSPIRE].

[8] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, *JHEP* **01** (2017) 110 [arXiv:1612.01551] [INSPIRE].

[9] ATLAS collaboration, *Quark versus gluon jet tagging using jet images with the ATLAS detector*, ATL-PHYS-PUB-2017-017, CERN, Geneva, Switzerland (2017).

[10] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning top taggers or the end of QCD?*, *JHEP* **05** (2017) 006 [arXiv:1701.08784] [INSPIRE].

[11] W. Bhimji, S.A. Farrell, T. Kurth, M. Paganini, Prabhat and E. Racah, *Deep neural networks for physics analysis on low-level whole-detector data at the LHC*, *J. Phys. Conf. Ser.* **1085** (2018) 042034 [arXiv:1711.03573] [INSPIRE].

[12] S. Macaluso and D. Shih, *Pulling out all the tops with computer vision and deep learning*, *JHEP* **10** (2018) 121 [arXiv:1803.00107] [INSPIRE].

[13] J. Guo, J. Li, T. Li, F. Xu and W. Zhang, *Deep learning for R-parity violating supersymmetry searches at the LHC*, *Phys. Rev.* **D 98** (2018) 076017 [arXiv:1805.10730] [INSPIRE].

[14] F.A. Dreyer, G.P. Salam and G. Soyez, *The Lund jet plane*, *JHEP* **12** (2018) 064 [arXiv:1807.04758] [INSPIRE].

[15] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban and D. Whiteson, *Jet flavor classification in high-energy physics with deep neural networks*, *Phys. Rev.* **D 94** (2016) 112002 [arXiv:1607.08633] [INSPIRE].

[16] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-aware recursive neural networks for jet physics*, arXiv:1702.00748 [INSPIRE].

[17] T. Cheng, *Recursive neural networks in quark/gluon tagging*, *Comput. Softw. Big Sci.* **2** (2018) 3 [arXiv:1711.02633] [INSPIRE].

[18] S. Egan, W. Fedorko, A. Lister, J. Pearkes and C. Gay, *Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC*, arXiv:1711.09059 [INSPIRE].

[19] K. Fraser and M.D. Schwartz, *Jet charge and machine learning*, *JHEP* **10** (2018) 093 [arXiv:1803.08066] [INSPIRE].

JHEP01(2019)121

[20] L.G. Almeida, M. Backović, M. Cliche, S.J. Lee and M. Perelstein, *Playing tag with ANN: boosted top identification with pattern recognition*, JHEP **07** (2015) 086 [arXiv:1501.05968] [INSPIRE].

[21] J. Pearkes, W. Fedorko, A. Lister and C. Gay, *Jet constituents for deep neural network based top quark tagging*, arXiv:1704.02124 [INSPIRE].

[22] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned top tagging with a Lorentz layer*, SciPost Phys. **5** (2018) 028 [arXiv:1707.08966] [INSPIRE].

[23] T. Roxlo and M. Reece, *Opening the black box of neural nets: case studies in stop/top discrimination*, arXiv:1804.09278 [INSPIRE].

[24] K. Datta and A. Larkoski, *How much information is in a jet?*, JHEP **06** (2017) 073 [arXiv:1704.08249] [INSPIRE].

[25] J.A. Aguilar-Saavedra, J.H. Collins and R.K. Mishra, *A generic anti-QCD jet tagger*, JHEP **11** (2017) 163 [arXiv:1709.01087] [INSPIRE].

[26] H. Lüo, M.-X. Luo, K. Wang, T. Xu and G. Zhu, *Quark jet versus gluon jet: deep neural networks with high-level features*, arXiv:1712.03634 [INSPIRE].

[27] L. Moore, K. Nordström, S. Varma and M. Fairbairn, *Reports of my demise are greatly exaggerated: N-subjettiness taggers take on jet images*, arXiv:1807.04769 [INSPIRE].

[28] K. Datta and A.J. Larkoski, *Novel jet observables from machine learning*, JHEP **03** (2018) 086 [arXiv:1710.01305] [INSPIRE].

[29] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy flow polynomials: a complete linear basis for jet substructure*, JHEP **04** (2018) 013 [arXiv:1712.07124] [INSPIRE].

[30] P.T. Komiske, E.M. Metodiev, B. Nachman and M.D. Schwartz, *Pileup mitigation with machine learning (PUMML)*, JHEP **12** (2017) 051 [arXiv:1707.08600] [INSPIRE].

[31] J.H. Collins, K. Howe and B. Nachman, *Anomaly detection for resonant new physics with machine learning*, Phys. Rev. Lett. **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

[32] R.T. D'Agnolo and A. Wulzer, *Learning new physics from a machine*, arXiv:1806.02350 [INSPIRE].

[33] A. De Simone and T. Jacques, *Guiding new physics searches with unsupervised learning*, arXiv:1807.06038 [INSPIRE].

[34] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty detection meets collider physics*, arXiv:1807.10261 [INSPIRE].

[35] M. Farina, Y. Nakai and D. Shih, *Searching for new physics with deep autoencoders*, arXiv:1808.08992 [INSPIRE].

[36] T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson, *QCD or what?*, arXiv:1808.08979 [INSPIRE].

[37] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, *Constraining effective field theories with machine learning*, Phys. Rev. Lett. **121** (2018) 111801 [arXiv:1805.00013] [INSPIRE].

[38] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, *A guide to constraining effective field theories with machine learning*, Phys. Rev. **D 98** (2018) 052004 [arXiv:1805.00020] [INSPIRE].

[39] J. D'Hondt, A. Mariotti, K. Mimasu, S. Moortgat and C. Zhang, *Learning to pinpoint effective operators at the LHC: a study of the $t\bar{t}b\bar{b}$ signature*, JHEP **11** (2018) 131 [arXiv:1807.02130] [INSPIRE].

[40] L. de Oliveira, M. Paganini and B. Nachman, *Learning particle physics by example: location-aware generative adversarial networks for physics synthesis*, Comput. Softw. Big Sci. **1** (2017) 4 [arXiv:1701.05927] [INSPIRE].

[41] M. Paganini, L. de Oliveira and B. Nachman, *Accelerating science with generative adversarial networks: an application to 3D particle showers in multilayer calorimeters*, Phys. Rev. Lett. **120** (2018) 042003 [arXiv:1705.02355] [INSPIRE].

[42] L. de Oliveira, M. Paganini and B. Nachman, *Controlling physical attributes in GAN-accelerated simulation of electromagnetic calorimeters*, in 18^{th} *International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2017)*, Seattle, WA, U.S.A., 21–25 August 2017 [J. Phys. Conf. Ser. **1085** (2018) 042017] [arXiv:1711.08813] [INSPIRE].

[43] M. Paganini, L. de Oliveira and B. Nachman, *CaloGAN: simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks*, Phys. Rev. **D 97** (2018) 014021 [arXiv:1712.10321] [INSPIRE].

[44] A. Andreassen, I. Feige, C. Frye and M.D. Schwartz, *JUNIPR: a framework for unsupervised machine learning in particle physics*, arXiv:1804.09720 [INSPIRE].

[45] P. Baldi, P. Sadowski and D. Whiteson, *Searching for exotic particles in high-energy physics with deep learning*, Nature Commun. **5** (2014) 4308 [arXiv:1402.4735] [INSPIRE].

[46] P. Baldi, P. Sadowski and D. Whiteson, *Enhanced Higgs boson to $\tau^+\tau^-$ search with deep learning*, Phys. Rev. Lett. **114** (2015) 111801 [arXiv:1410.3469] [INSPIRE].

[47] J. Searcy, L. Huang, M.-A. Pleier and J. Zhu, *Determination of the WW polarization fractions in $pp \to W^\pm W^\pm jj$ using a deep machine learning technique*, Phys. Rev. **D 93** (2016) 094033 [arXiv:1510.01691] [INSPIRE].

[48] R. Santos et al., *Machine learning techniques in searches for $t\bar{t}h$ in the $h \to b\bar{b}$ decay channel*, 2017 JINST **12** P04014 [arXiv:1610.03088] [INSPIRE].

[49] E. Barberio, B. Le, E. Richter-Was, Z. Was, D. Zanzi and J. Zaremba, *Deep learning approach to the Higgs boson CP measurement in $H \to \tau\tau$ decay and associated systematics*, Phys. Rev. **D 96** (2017) 073002 [arXiv:1706.07983] [INSPIRE].

[50] J. Duarte et al., *Fast inference of deep neural networks in FPGAs for particle physics*, 2018 JINST **13** P07027 [arXiv:1804.06913] [INSPIRE].

[51] M. Abdughani, J. Ren, L. Wu and J.M. Yang, *Probing stop with graph neural network at the LHC*, arXiv:1807.09088 [INSPIRE].

[52] J. Lin, M. Freytsis, I. Moult and B. Nachman, *Boosting $H \to b\bar{b}$ with machine learning*, JHEP **10** (2018) 101 [arXiv:1807.10768] [INSPIRE].

[53] Y.S. Lai, *Automated discovery of jet substructure analyses*, arXiv:1810.00835 [INSPIRE].

[54] 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*, arXiv:1709.04464 [INSPIRE].

[55] D. Guest, K. Cranmer and D. Whiteson, *Deep learning and its application to LHC physics*, Ann. Rev. Nucl. Part. Sci. **68** (2018) 161 [arXiv:1806.11484] [INSPIRE].

JHEP01(2019)121

[56] K. Albertsson et al., *Machine learning in high energy physics community white paper*, *J. Phys. Conf. Ser.* **1085** (2018) 022008 [arXiv:1807.02876] [INSPIRE].

[57] A. Radovic et al., *Machine learning at the energy and intensity frontiers of particle physics*, *Nature* **560** (2018) 41 [INSPIRE].

[58] P. Sadowski and P. Baldi, *Deep learning in the natural sciences: applications to physics*, Springer International Publishing, Cham, Switzerland (2018), pg. 269.

[59] M. Iyyer, V. Manjunatha, J.L. Boyd-Graber and H. Daumé III, *Deep unordered composition rivals syntactic methods for text classification*, in *Proceedings of the* $53^{\rm rd}$ *Annual Meeting of the Association for Computational Linguistics and the* $7^{\rm th}$ *International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, ACL 2015, Beijing, China, 26–31 July 2015, pg. 1681.

[60] R.Q. Charles, H. Su, M. Kaichun and L.J. Guibas, *PointNet: deep learning on point sets for 3d classification and segmentation*, in 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI, U.S.A., 21–26 July 2017, pg. 77.

[61] S.H. Rezatofighi, B.G.V. Kumar, A. Milan, E. Abbasnejad, A. Dick and I. Reid, *DeepSetNet: predicting sets with deep neural networks*, in 2017 *IEEE International Conference on Computer Vision (ICCV)*, IEEE, Venice, Italy, 22–29 October 2017, pg. 5257.

[62] C.R. Qi, L. Yi, H. Su and L.J. Guibas, *Pointnet++: deep hierarchical feature learning on point sets in a metric space*, in *Advances in Neural Information Processing Systems* 30*: annual conference on neural information processing systems* 2017, Long Beach, CA, U.S.A., 4–9 December 2017, pg. 5105 [arXiv:1706.02413].

[63] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R.R. Salakhutdinov and A.J. Smola, *Deep sets*, in *Advances in Neural Information Processing Systems* 30*: annual conference on neural information processing systems* 2017, Long Beach, CA, U.S.A., 4–9 December 2017, pg. 3394 [arXiv:1703.06114].

[64] A. Gardner, J. Kanno, C.A. Duncan and R.R. Selmic, *Classifying unordered feature sets with convolutional deep averaging networks*, arXiv:1709.03019.

[65] Y. Yang, C. Feng, Y. Shen and D. Tian, *Foldingnet: interpretable unsupervised learning on 3d point clouds*, arXiv:1712.07262.

[66] S.H. Rezatofighi, A. Milan, Q. Shi, A.R. Dick and I.D. Reid, *Joint learning of set cardinality and state distribution*, in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, U.S.A., 2–7 February 2018 [arXiv:1709.04093].

[67] S.H. Rezatofighi et al., *Deep perm-set net: learn to predict sets with unknown permutation and cardinality using deep neural networks*, arXiv:1805.00613.

[68] A. Cotter et al., *Interpretable set functions*, arXiv:1806.00050.

[69] J. Pumplin, *How to tell quark jets from gluon jets*, *Phys. Rev.* **D 44** (1991) 2025 [INSPIRE].

[70] J. Thaler and K. Van Tilburg, *Identifying boosted objects with N-subjettiness*, *JHEP* **03** (2011) 015 [arXiv:1011.2268] [INSPIRE].

[71] J. Thaler and K. Van Tilburg, *Maximizing boosted top identification by minimizing N-subjettiness*, *JHEP* **02** (2012) 093 [arXiv:1108.2701] [INSPIRE].

[72] D. Krohn, M.D. Schwartz, T. Lin and W.J. Waalewijn, *Jet charge at the LHC*, *Phys. Rev. Lett.* **110** (2013) 212001 [arXiv:1209.2421] [INSPIRE].

[73] H.-M. Chang, M. Procura, J. Thaler and W.J. Waalewijn, *Calculating track-based observables for the LHC*, *Phys. Rev. Lett.* **111** (2013) 102002 [arXiv:1303.6637] [INSPIRE].

[74] A.J. Larkoski, J. Thaler and W.J. Waalewijn, *Gaining (mutual) information about quark/gluon discrimination*, *JHEP* **11** (2014) 129 [arXiv:1408.3122] [INSPIRE].

[75] A.J. Larkoski, G.P. Salam and J. Thaler, *Energy correlation functions for jet substructure*, *JHEP* **06** (2013) 108 [arXiv:1305.0007] [INSPIRE].

[76] I. Moult, L. Necib and J. Thaler, *New angles on energy correlation functions*, *JHEP* **12** (2016) 153 [arXiv:1609.07483] [INSPIRE].

[77] A.J. Larkoski, S. Marzani, G. Soyez and J. Thaler, *Soft drop*, *JHEP* **05** (2014) 146 [arXiv:1402.2657] [INSPIRE].

[78] C. Frye, A.J. Larkoski, J. Thaler and K. Zhou, *Casimir meets Poisson: improved quark/gluon discrimination with counting observables*, *JHEP* **09** (2017) 083 [arXiv:1704.06266] [INSPIRE].

[79] A. Abdesselam et al., *Boosted objects: a probe of beyond the Standard Model physics*, *Eur. Phys. J.* **C 71** (2011) 1661 [arXiv:1012.5412] [INSPIRE].

[80] A. Altheimer et al., *Jet substructure at the Tevatron and LHC: new results, new tools, new benchmarks*, *J. Phys.* **G 39** (2012) 063001 [arXiv:1201.0008] [INSPIRE].

[81] A. Altheimer et al., *Boosted objects and jet substructure at the LHC. Report of BOOST*2012*, held at IFIC Valencia,* 23–27 *July* 2012, *Eur. Phys. J.* **C 74** (2014) 2792 [arXiv:1311.2708] [INSPIRE].

[82] D. Adams et al., *Towards an understanding of the correlations in jet substructure*, *Eur. Phys. J.* **C 75** (2015) 409 [arXiv:1504.00679] [INSPIRE].

[83] L. Asquith et al., *Jet substructure at the Large Hadron Collider: experimental review*, arXiv:1803.06991 [INSPIRE].

[84] J. Gallicchio and M.D. Schwartz, *Quark and gluon tagging at the LHC*, *Phys. Rev. Lett.* **107** (2011) 172001 [arXiv:1106.3076] [INSPIRE].

[85] J. Gallicchio and M.D. Schwartz, *Quark and gluon jet substructure*, *JHEP* **04** (2013) 090 [arXiv:1211.7038] [INSPIRE].

[86] ATLAS collaboration, *Light-quark and gluon jet discrimination in pp collisions at* $\sqrt{s} = 7$ *TeV with the ATLAS detector*, *Eur. Phys. J.* **C 74** (2014) 3023 [arXiv:1405.6583] [INSPIRE].

[87] P. Gras et al., *Systematics of quark/gluon tagging*, *JHEP* **07** (2017) 091 [arXiv:1704.03878] [INSPIRE].

[88] E.M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174 [arXiv:1708.02949] [INSPIRE].

[89] P.T. Komiske, E.M. Metodiev, B. Nachman and M.D. Schwartz, *Learning to classify from impure samples with high-dimensional data*, *Phys. Rev.* **D 98** (2018) 011502 [arXiv:1801.10158] [INSPIRE].

[90] P.T. Komiske, E.M. Metodiev and J. Thaler, *An operational definition of quark and gluon jets*, *JHEP* **11** (2018) 059 [arXiv:1809.01140] [INSPIRE].

JHEP01(2019)121

[91] *EnergyFlow webpage*, https://energyflow.network.

[92] M.H. Stone, *The generalized Weierstrass approximation theorem*, *Math. Magazine* **21** (1948) 237.

[93] CMS collaboration, *Performance of quark/gluon discrimination in* 8 *TeV pp data*, CMS-PAS-JME-13-002, CERN, Geneva, Switzerland (2013).

[94] G. Parisi, *Super inclusive cross-sections*, *Phys. Lett.* **B 74** (1978) 65 [INSPIRE].

[95] T. Kinoshita, *Mass singularities of Feynman amplitudes*, *J. Math. Phys.* **3** (1962) 650 [INSPIRE].

[96] T.D. Lee and M. Nauenberg, *Degenerate systems and mass singularities*, *Phys. Rev.* **133** (1964) B1549 [INSPIRE].

[97] CTEQ collaboration, *Handbook of perturbative QCD: version* 1.0, *Rev. Mod. Phys.* **67** (1995) 157 [INSPIRE].

[98] S. Weinberg, *The quantum theory of fields. Volume* 1*: foundations*, Cambridge University Press, Cambridge, U.K. (2005) [INSPIRE].

[99] A.J. Larkoski and J. Thaler, *Unsafe but calculable: ratios of angularities in perturbative QCD*, *JHEP* **09** (2013) 137 [arXiv:1307.1699] [INSPIRE].

[100] A.J. Larkoski, S. Marzani and J. Thaler, *Sudakov safety in perturbative QCD*, *Phys. Rev.* **D 91** (2015) 111501 [arXiv:1502.01719] [INSPIRE].

[101] F.V. Tkachov, *Measuring multi-jet structure of hadronic energy flow or what is a jet?*, *Int. J. Mod. Phys.* **A 12** (1997) 5411 [hep-ph/9601308] [INSPIRE].

[102] N.A. Sveshnikov and F.V. Tkachov, *Jets and quantum field theory*, *Phys. Lett.* **B 382** (1996) 403 [hep-ph/9512370] [INSPIRE].

[103] P.S. Cherzor and N.A. Sveshnikov, *Jet observables and energy momentum tensor*, in *Quantum field theory and high-energy physics. Proceedings, Workshop, QFTHEP '97*, Samara, Russia, 4–10 September 1997, pg. 402 [hep-ph/9710349] [INSPIRE].

[104] G.C. Fox and S. Wolfram, *Observables for the analysis of event shapes in e^+e^- annihilation and other processes*, *Phys. Rev. Lett.* **41** (1978) 1581 [INSPIRE].

[105] J.F. Donoghue, F.E. Low and S.-Y. Pi, *Tensor analysis of hadronic jets in quantum chromodynamics*, *Phys. Rev.* **D 20** (1979) 2759 [INSPIRE].

[106] G. Gur-Ari, M. Papucci and G. Perez, *Classification of energy flow observables in narrow jets*, arXiv:1101.2905 [INSPIRE].

[107] Particle Data Group collaboration, *Review of particle physics*, *Phys. Rev.* **D 98** (2018) 030001 [INSPIRE].

[108] V. Nair and G.E. Hinton, *Rectified linear units improve restricted Boltzmann machines*, in *Proceedings of the* 27^{th} *International Conference on Machine Learning (ICML-*10*)*, Haifa, Israel (2010), pg. 807.

[109] K. He, X. Zhang, S. Ren and J. Sun, *Delving deep into rectifiers: surpassing human-level performance on ImageNet classification*, in 2015 *IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile (2015), pg. 1026.

[110] T. Sjöstrand, S. Mrenna and P.Z. Skands, *PYTHIA* 6.4 *physics and manual*, *JHEP* **05** (2006) 026 [hep-ph/0603175] [INSPIRE].

JHEP01(2019)121

[111] T. Sjöstrand et al., *An introduction to PYTHIA* 8.2, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[112] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[113] M. Cacciari, G.P. Salam and G. Soyez, *FastJet user manual*, *Eur. Phys. J.* **C 72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[114] L.M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly supervised classification in high energy physics*, *JHEP* **05** (2017) 145 [arXiv:1702.00414] [INSPIRE].

[115] T. Cohen, M. Freytsis and B. Ostdiek, *(Machine) learning to do more with less*, *JHEP* **02** (2018) 034 [arXiv:1706.09451] [INSPIRE].

[116] G. Blanchard, M. Flaska, G. Handy, S. Pozzi and C. Scott, *Classification with asymmetric label noise: consistency and maximal denoising*, *Electron. J. Statist.* **10** (2016) 2780 [*Corrigendum ibid.* **12** (2018) 1779] [arXiv:1303.1208].

[117] E.M. Metodiev and J. Thaler, *Jet topics: disentangling quarks and gluons at colliders*, *Phys. Rev. Lett.* **120** (2018) 241602 [arXiv:1802.00008] [INSPIRE].

[118] A.J. Larkoski, D. Neill and J. Thaler, *Jet shapes with the broadening axis*, *JHEP* **04** (2014) 017 [arXiv:1401.2158] [INSPIRE].

[119] T. Likhomanenko, P. Ilten, E. Khairullin, A. Rogozhnikov, A. Ustyuzhanin and M. Williams, *LHCb topological trigger reoptimization*, *J. Phys. Conf. Ser.* **664** (2015) 082025 [arXiv:1510.00572] [INSPIRE].

[120] F. Pedregosa et al., *Scikit-learn: machine learning in python*, *J. Machine Learn. Res.* **12** (2011) 2825.

[121] A.J. Larkoski, I. Moult and D. Neill, *Toward multi-differential cross sections: measuring two angularities on a single jet*, *JHEP* **09** (2014) 046 [arXiv:1401.4458] [INSPIRE].

[122] M. Procura, W.J. Waalewijn and L. Zeune, *Joint resummation of two angularities at next-to-next-to-leading logarithmic order*, *JHEP* **10** (2018) 098 [arXiv:1806.10622] [INSPIRE].

[123] D. Bertolini, P. Harris, M. Low and N. Tran, *Pileup per particle identification*, *JHEP* **10** (2014) 059 [arXiv:1407.6013] [INSPIRE].

[124] *Fastjet contrib webpage*, https://fastjet.hepforge.org/contrib/.

[125] F. Chollet, *Keras*, https://github.com/fchollet/keras, (2015).

[126] M. Abadi et al., *Tensorflow: a system for large-scale machine learning*, in *Proceedings OSDI*, volume 16, (2016), pg. 265 [arXiv:1605.08695].

[127] T. Oliphant, *Guide to NumPy*, Trelgol Publishing, U.S.A. (2006).

[128] D.P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, arXiv:1412.6980 [INSPIRE].

[129] DELPHES 3 collaboration, *DELPHES* 3*, a modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].

JHEP01(2019)121