JHEP

Published for SISSA by Springer

Received: *September 24, 2018*
Revised: *October 25, 2018*
Accepted: *November 2, 2018*
Published: *November 8, 2018*

An operational definition of quark and gluon jets

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.

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

Abstract: While "quark" and "gluon" jets are often treated as separate, well-defined objects in both theoretical and experimental contexts, no precise, practical, and hadron-level definition of jet flavor presently exists. To remedy this issue, we develop and advocate for a data-driven, operational definition of quark and gluon jets that is readily applicable at colliders. Rather than specifying a per-jet flavor label, we aggregately define quark and gluon jets at the distribution level in terms of measured hadronic cross sections. Intuitively, quark and gluon jets emerge as the two maximally separable categories within two jet samples in data. Benefiting from recent work on data-driven classifiers and topic modeling for jets, we show that the practical tools needed to implement our definition already exist for experimental applications. As an informative example, we demonstrate the power of our operational definition using Z+jet and dijet samples, illustrating that pure quark and gluon distributions and fractions can be successfully extracted in a fully well-defined manner.

Keywords: Jets

ArXiv ePrint: 1809.01140

https://doi.org/10.1007/JHEP11(2018)059

JHEP11(2018)059

Contents

1 Introduction **1**

2 Defining quark and gluon jets **4**
- 2.1 Review of a conceptual quark/gluon jet definition 4
- 2.2 Motivating the operational definition 5
- 2.3 An operational definition of quark and gluon jets 8

3 Data-driven jet taggers and topics **9**
- 3.1 Classification without labels: training classifiers on collider data 10
- 3.2 Jet topics: extracting categories from collider data 11
- 3.3 Optimal taggers for optimal topics 12

4 Quark and gluon jets from dijets and Z+jet **13**
- 4.1 Event generation 13
- 4.2 Extracting reducibility factors and fractions 14
- 4.3 Self-calibrating classifiers 18
- 4.4 Obtaining observable distributions from extracted fractions 19

5 Conclusions **20**

A Theoretical exploration of Casimir- and Poisson-scaling observables **23**

B Details of observables and machine learning models **25**

C Sample dependence in parton shower events **27**

JHEP11(2018)059

1 Introduction

Quarks and gluons are fundamental, color-charged particles that are copiously produced at colliders like the Large Hadron Collider (LHC). Despite their ubiquity, these high-energy quarks and gluons are never observed directly. Instead, they fragment and hadronize into sprays of color-neutral hadrons, known as *jets*, via quantum chromodynamics (QCD). As the majority of jets originate from light (up, down, strange) quarks or gluons, a firm understanding of quark and gluon jets is important to many analyses at the LHC. There has been tremendous recent theoretical and experimental progress in analyzing jets and jet substructure [1–11], with a variety of observables [12–22] and algorithms [23–27] developed to expose and probe the underlying physics. Despite decades of using the notions of "quark" and "gluon" jets [28–42], a precise and practical hadron-level definition of jet flavor has not been formulated.

Even setting aside the issue of jet flavor, ambiguity is already present whenever one wants to identify jets in an event [43]. Nonetheless, jets can be made perfectly well-defined: any hadron-level algorithm for finding jets that is infrared and collinear (IRC) safe provides an operational jet definition that can be compared to perturbative predictions. While different algorithms result in different jets, specifying a jet algorithm allows one to make headway into comparing theoretical calculations and experimental measurements. Meanwhile, in the case of jet flavor, the lack of a precise, hadron-level definition of "quark" and "gluon" jets has artificially hindered progress by precluding separate comparisons of quark and gluon jets between theory and experiment.

Typical applications involving "quark" and "gluon" jets in practice often rely on ill-defined or unphysical parton-level information, such as from the event record of a parton shower event generator. Progress has been made in providing sharp definitions at the parton-level [44, 45], in the context of factorization theorems [46–48], and at the conceptual level [49], but an operational definition, to our knowledge, has never been developed (see ref. [50] for a review). A quark/gluon jet definition[1] should ideally work at the hadron level, regardless of whether a rigorous factorization theorem exists, and be practically implementable in both theoretical and experimental settings.

In this paper, we develop an operational definition of quark and gluon jets that is formulated solely in terms of experimentally-accessible quantities, does not rely on specific theoretical constructs such as factorization theorems, and can be readily implemented in a realistic context. Intuitively, we define quark and gluon jets as the "pure" categories that emerge from two different jet samples. Our definition operates at the aggregate level, avoiding altogether the troublesome and potentially impossible notion of a per-jet flavor label in favor of quantifying quark and gluon jets by their distributions.

Specifically, given two jet samples M_1 and M_2 (e.g. Z+jet and dijet) in a narrow transverse momentum (p_T) bin, with M_1 taken to be more "quark"-like, and a jet substructure feature space $\mathcal{O}$, we define quark (q) and gluon (g) jet distributions in the following way:

$$p_q(\mathcal{O}) \equiv \frac{p_{M_1}(\mathcal{O}) - \kappa_{12}\, p_{M_2}(\mathcal{O})}{1-\kappa_{12}}, \qquad p_g(\mathcal{O}) \equiv \frac{p_{M_2}(\mathcal{O}) - \kappa_{21}\, p_{M_1}(\mathcal{O})}{1-\kappa_{21}}, \tag{1.1}$$

where κ_{12} and κ_{21} are known as *reducibility factors* and are directly obtainable from the probability distributions $p_{M_1}(\mathcal{O})$ and $p_{M_2}(\mathcal{O})$. The reducibility factors are defined as:

$$\kappa_{12} \equiv \min_{\mathcal{O}} \frac{p_{M_1}(\mathcal{O})}{p_{M_2}(\mathcal{O})}, \qquad \kappa_{21} \equiv \min_{\mathcal{O}} \frac{p_{M_2}(\mathcal{O})}{p_{M_1}(\mathcal{O})}. \tag{1.2}$$

The reducibility factors in eq. (1.2) identify the most M_1-like and M_2-like regions of the substructure phase space by extremizing the sample likelihood ratio. We take these phase space regions to *define* what it means to be quark-like and gluon-like. The subtractions in eq. (1.1) then proceed to "demix" the two sample distributions as if they were statistical mixtures. The quark and gluon distributions are defined solely in terms of hadronic fiducial cross section measurements of the two samples, ensuring that our definition is manifestly

[1] While in some contexts "jet definition" means a procedure for finding jets in an event, in this paper we use "quark/gluon jet definition" to mean a definition of jet flavor.

JHEP11(2018)059

fully data-driven and non-circular. This definition relies on a jet algorithm to define the jets in the jet samples, which also allows for further hadron-level processing, such as jet grooming techniques [23–27], to be folded directly into the quark/gluon jet definition.

One main goal of this paper is to argue that our operational definition, combined with existing tools, provides a way to obtain information about the likelihood, quark fractions, and quark and gluon distributions in a fully data-driven way, without reference to unphysical notions such as generator labels. The concepts appearing in our definition are directly related to methods already in use in experimental quark/gluon jet analysis efforts [51–56]. Quark-gluon likelihood ratios, obtained from parton shower generators, have been implemented by both ATLAS and CMS as optimal discriminants in low-dimensional feature spaces. Quark fractions, obtained from event generators, for several jet samples have successfully allowed for separate determination of quark and gluon jet properties by solving linear equations. These analyses already use a statistical-mixture picture of quark and gluon jets, which is a direct consequence of our definition.

Many physics analyses at the LHC would benefit from a clear definition of quark and gluon jets that allows for unambiguous extraction of separate quark and gluon jet distributions and fractions. Fully data-driven quark/gluon jet taggers have the potential to increase the sensitivity of a variety of new physics searches [37, 38], and related ideas have been developed for model-independent searches for new physics [57]. Experimentally measuring separate quark and gluon distributions of jet observables would significantly improve attempts to extract the strong coupling constant from jet substructure [58] and to constrain parton shower event generators [50, 59]. Extracting data-driven fractions of quark and gluon jets could improve the determination of parton distribution functions and allow for separate measurement of quark and gluon cross sections. These ideas may also be relevant in the context of heavy ion collisions, where quarks and gluons are expected to be modified differently by the medium and probing the separate modifications to quark and gluon jets would be of significant interest.

We now give a brief summary of the rest of this paper. In section 2, we provide a self-contained overview, motivation, and exploration of our quark/gluon jet definition. We discuss recent work in ref. [50] that developed a "conceptual" definition of quark/gluon jets, falling short of providing a full definition that can be reliably used in practice, but highlighting the key elements required of a sensible quark/gluon jet definition. We then develop the intuition and mathematical tools necessary to construct our operational definition, which satisfies the core conceptual principles while being precise and practically implementable. After stating our operational definition, we examine its physical and statistical properties in detail. An exploration of the definition in the context of simple jet substructure observables at leading-logarithmic accuracy is left to appendix A.

In section 3, we discuss how our quark/gluon jet definition benefits from, and provides a foundation for, recent work on data-driven machine learning for jet physics. The classification without labels (CWoLa) paradigm [60] for training classifiers on mixed samples can be used to approximate the mixed-sample likelihood ratio, a key part of implementing our definition. The jet topics framework [61] extracts underlying mutually irreducible distributions from mixture histograms, yielding a practical method to obtain the reducibility

JHEP11(2018)059

factors in eq. (1.2). Using jet topics with the approximated mixed-sample likelihood ratio, obtained from the data via CWoLa, allows for more robust fraction and distribution extraction. With quark fractions, obtained from the data via jet topics, CWoLa classifiers can be (self-)calibrated in a fully data-driven way. More broadly, the assumptions required for CWoLa and jet topics — that QCD jet samples are statistical mixtures of mutually irreducible quark and gluon jets — are satisfied by construction with our definition.

In section 4, we showcase a practical implementation of our definition using jet samples from two different processes: Z+jet and dijets. Using six trained models detailed in appendix B, we apply the procedure outlined in section 3 to extract quark fractions by combining the CWoLa and jet topics methods, finding more robust performance than when using single jet substructure observables. With the reducibility factors and quark fractions in hand, we extract separate quark and gluon distributions for a variety of jet substructure observables, even those that do not exhibit mutual irreducibility. We compare the results of using our data-driven definition of quark and gluon jets with a per-jet Pythia-parton definition, finding qualitative and quantitative agreement between the two. The potential to self-calibrate CWoLa classifiers is also shown with an explicit example. While our studies are based on parton-shower samples, all of these analyses can in principle be performed in data with the experimental tools already developed for quark and gluon jet physics at the LHC.

We present our conclusions in section 5, discussing potential new applications made feasible by this work. Possible future developments and extensions are highlighted. A study of the similarity of parton-labeled quark and gluon jets between different processes is left to appendix C.

2 Defining quark and gluon jets

2.1 Review of a conceptual quark/gluon jet definition

Due to the complicated radiative showering and fundamentally non-perturbative hadronization that occurs in the course of jets emerging from partons, there is no unambiguous definition of "quark" or "gluon" jets at the hadron-level. Despite this challenge, the importance of a clear, well-defined, and practical definition of quark and gluon jets at modern colliders cannot be overstated. In ref. [50], a significant effort was made to summarize and comment on the concepts of "quark jet" and "gluon jet". The authors of ref. [50] settled on the following statement as the best way to conceptually define quark jets (and, analogously, gluon jets):

Quark and gluon jet definition (conceptual) [50]. *A phase space region (as defined by an unambiguous hadronic fiducial cross section measurement) that yields an enriched sample of quarks (as interpreted by some suitable, though fundamentally ambiguous criterion).*

This definition is attractive for numerous reasons. First, it is explicitly tied to hadronic final states, avoiding dependence, for example, on the unphysical event record of a parton shower generator. Further, it is specific to the context of a particular measurement and is

JHEP11(2018)059

thus defined regardless of whether the observable and processes in question have rigorous factorization theorems. Finally, its goal is to tag a region of phase space as quark- or gluon-like rather than to specify a per-jet truth definition of quark and gluon jets. The main difficulty with this conceptual definition, as noted in ref. [50], is determining the criterion that corresponds to successful quark or gluon jet enrichment.

Despite its attractive qualities, without a practical proposal for implementing this conceptual definition on data, the case studies in ref. [50] operationally fell back on less well-defined definitions, such as using initiating parton information from a parton shower generator to tag a quark/gluon jet. Further, the definition only tags specific regions of phase space as "quark" or "gluon", such as low or high values of some substructure observable, and provides no framework for discussing jet flavor outside of these regions. To remedy this issue, we seek to upgrade the conceptual definition to an operational one by giving a concrete, data-driven method for optimally identifying quark- or gluon-enriched regions of phase space and obtaining full quark and gluon jet distributions.

2.2 Motivating the operational definition

To motivate our definition, suppose that we have two QCD jet samples M_1 and M_2 in a narrow p_T bin. One of the mixed samples (M_1 without loss of generality) should be "quark-enriched" and the other "gluon-enriched" relative to each other according to some qualitative criterion. Ref. [50] took M_1 and M_2 to be, respectively, Z+jet and dijet samples, a case that we further investigate in section 4.

Assume for now that M_1 and M_2 are statistical mixtures of quark and gluon jets — an assumption that will *not* be made in our final definition. Letting the quark fractions of the two mixtures be f_1 and f_2, the relationship between the distribution of substructure observables in mixture M_i in terms of the quark and gluon jet distributions is:

$$p_{M_i}(\mathcal{O}) = f_i\, p_q(\mathcal{O}) + (1 - f_i)\, p_g(\mathcal{O}), \tag{2.1}$$

where the feature space $\mathcal{O}$ is, for our purposes, a set of jet substructure observables taken to be sufficiently rich to encode all relevant information about jet flavor.

Following the outline of the Conceptual Definition, we consider classification of quark and gluon jets and examine the relationship of this task with classification of one mixture from the other. By the Neyman-Pearson lemma [62], an optimal classifier for discriminating two classes is their likelihood ratio (or any monotonically-related quantity). In the case of quark and gluon jets, the likelihood ratio is:

$$L_{q/g}(\mathcal{O}) \equiv \frac{p_q(\mathcal{O})}{p_g(\mathcal{O})}, \tag{2.2}$$

and, similarly, the optimal classifier for discriminating between M_1 and M_2 is:

$$L_{M_1/M_2}(\mathcal{O}) \equiv \frac{p_{M_1}(\mathcal{O})}{p_{M_2}(\mathcal{O})} = \frac{f_1\, L_{q/g}(\mathcal{O}) + (1 - f_1)}{f_2\, L_{q/g}(\mathcal{O}) + (1 - f_2)}. \tag{2.3}$$

It is easily verified that the mixed-sample likelihood ratio in eq. (2.3) is a monotonic function of the quark-gluon likelihood ratio in eq. (2.2) as long as $f_1 \neq f_2$ (see refs. [60, 63]). The

JHEP11(2018)059

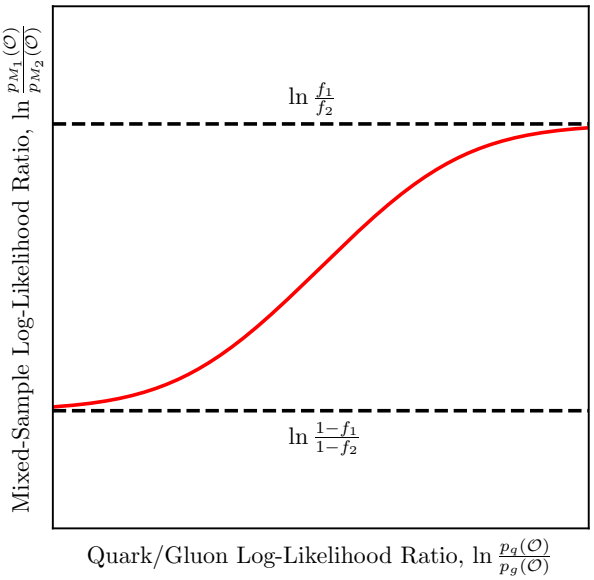


Figure 1. The monotonic relationship between the mixed-sample log-likelihood ratio and the quark-gluon log-likelihood ratio from eq. (2.3) for illustrative fraction values. The relationship between the maximum and minimum values of the mixed-sample and quark/gluon log-likelihoods from eq. (2.8) is visually evident in that the red curve horizontally asymptotes to the two black dashed curves. The plots are shown in terms of the logarithms of the likelihood ratios so that exchanging $M_1 \leftrightarrow M_2$ or $q \leftrightarrow g$ simply corresponds to a reflection of the curve.

relationship between the mixed-sample likelihood ratio and the quark-gluon likelihood ratio of eq. (2.3) is depicted in figure 1. This cleanly demonstrates that the optimal mixed-sample classifier is also the optimal quark-gluon classifier.

Supposing that we can approximate the mixture likelihood ratio sufficiently well, we have distilled the (potentially huge) substructure feature space to a single number which is provably optimal for identifying quark- and gluon-enriched phase space regions. However, we still lack a procedure for actually identifying the enriched regions; we solely know that they are given by some cut on $L_{q/g}(\mathcal{O})$, or equivalently a cut on $L_{M_1/M_2}(\mathcal{O})$. The key insight for moving closer toward an operational definition is that $L_{q/g}(\mathcal{O})$, being the optimal discriminant of quark and gluon jets, can be immediately used to identify the most quark-enriched (gluon-enriched) regions as those where $L_{q/g}(\mathcal{O})$ is at its maximum (minimum). In the case that we can find regions of phase space $\mathcal{O}_q$ and $\mathcal{O}_g$ where quark and gluon jets respectively are pure, we have that $L_{q/g}(\mathcal{O}_g) = 0$ and $L_{g/q}(\mathcal{O}_q) = 0$ and we say that the quark and gluon categories are *mutually irreducible* (see refs. [61, 63]).

The extrema of the quark/gluon likelihood ratio $L_{q/g}$, corresponding to the enriched regions of phase space, are naturally related to the extrema of the mixture likelihood ratio L_{M_1/M_2}. To this end, it is helpful to define the *reducibility factor* between distributions A and B, κ_{AB}, as:

$$\kappa_{AB} \equiv \min_{\mathcal{O}} \frac{p_A(\mathcal{O})}{p_B(\mathcal{O})}, \tag{2.4}$$

JHEP11(2018)059

which is the minimum (or more precisely, the infimum) of the likelihood ratio of A and B. Supposing that quarks and gluons are mutually irreducible in the feature space $\mathcal{O}$, the reducibility factors of quark jets to gluon jets (and vice versa) vanish:

$$\textit{Quark and gluon jet mutual irreducibility}: \qquad \kappa_{qg} = 0, \qquad \kappa_{gq} = 0. \tag{2.5}$$

We now show how, assuming quark/gluon mutual irreducibility, the mixture reducibility factors can be related to mixture fractions. The reducibility factors of the mixed samples can be written down by treating them as mixtures of quarks and gluons as in eq. (2.1):

$$\kappa_{M_i M_j} = \min_{\mathcal{O}} L_{M_i/M_j}(\mathcal{O}) = \min_{\mathcal{O}} \frac{f_i\, L_{q/g}(\mathcal{O}) + (1 - f_i)}{f_j\, L_{q/g}(\mathcal{O}) + (1 - f_j)}. \tag{2.6}$$

Using our assumptions that M_1 is quark-enriched relative to M_2, we can write eq. (2.6) as a relation between the mixed-sample reducibility factors and the quark/gluon reducibility factors:

$$\kappa_{M_1 M_2} = \frac{f_1\, \kappa_{qg} + (1 - f_1)}{f_2\, \kappa_{qg} + (1 - f_2)}, \qquad \kappa_{M_2 M_1} = \frac{f_2 + (1 - f_2)\, \kappa_{gq}}{f_1 + (1 - f_1)\, \kappa_{gq}}, \tag{2.7}$$

where the monotonicity of $L_{M_i/M_j}(\mathcal{O})$ with $L_{q/g}(\mathcal{O})$ has been used to push the minimum operation onto the quark-gluon likelihood ratio in eq. (2.6). If quarks and gluons are mutually irreducible, we can plug eq. (2.5) into eq. (2.7) to find the reducibility factors of the mixtures:[2]

$$\kappa_{12} \equiv \kappa_{M_1 M_2} = \frac{1 - f_1}{1 - f_2}, \qquad \kappa_{21} \equiv \kappa_{M_2 M_1} = \frac{f_2}{f_1}. \tag{2.8}$$

Figure 1 demonstrates that eq. (2.6) defines the asymptotic behavior of the mixed-sample log-likelihood ratio.

Combining the reducibility factors of eq. (2.8) with the mixture relationship of eq. (2.1), we can solve for the underlying quark and gluon jet distributions solely in terms of the well-defined mixture distributions $p_{M_i}(\mathcal{O})$ and mixture reducibility factors κ_{ij}:

$$p_q(\mathcal{O}) = \frac{p_{M_1}(\mathcal{O}) - \kappa_{12}\, p_{M_2}(\mathcal{O})}{1 - \kappa_{12}}, \qquad p_g(\mathcal{O}) = \frac{p_{M_2}(\mathcal{O}) - \kappa_{21}\, p_{M_1}(\mathcal{O})}{1 - \kappa_{21}}. \tag{2.9}$$

Remarkably, eq. (2.9) exposes the underlying quark and gluon jet distributions in terms of experimentally well-defined quantities such as the distribution of jets in mixed samples and their reducibility factors. Notice also that the quark and gluon distributions each depend on only one of the two mixed-sample reducibility factors. Thus, even if only one reducibility factor can be reliably extracted, the corresponding quark or gluon jet distribution can nevertheless be obtained.

Here, we have made several simplifying assumptions, namely that quark and gluon jets can be made well-defined, that M_1 and M_2 are statistical mixtures of quark and gluon jets, and that quark and gluon jets are mutually irreducible in the feature space $\mathcal{O}$. eq. (2.9) then followed as a consequence, demonstrating that, under these assumptions, it is possible to get access to pure quark and gluon distributions. What if, on the contrary, we do not make these assumptions, while also requiring that our definition of quark and gluon jets not be circular? We now proceed to thoroughly explore this idea.

[2] An analogous analysis carries through even if non-zero reducibility factors κ_{qg} and κ_{gq} are specified.

JHEP11(2018)059

2.3 An operational definition of quark and gluon jets

We now provide our *operational definition* of quark and gluon jets that builds upon the Conceptual Definition in section 2.1 but can be used for practical applications at the LHC and future colliders. We begin by stating the definition in terms of the notation developed in section 2.2, and then we proceed to a detailed discussion of its features.

In the absence of any certainty about the underlying structure of samples M_1 and M_2, we choose to start at the end of section 2.2, letting eq. (2.9) provide a fully-operational definition of quark and gluon jets in terms of experimentally well-defined quantities:

Quark and gluon jet definition (operational). *Given two samples M_1 and M_2 of QCD jets at a fixed p_T obtained by a suitable jet-finding procedure, taking M_1 to be "quark-enriched" compared to M_2, and a jet substructure feature space $\mathcal{O}$, the quark and gluon jet distributions are defined to be:*

$$p_q(\mathcal{O}) \equiv \frac{p_{M_1}(\mathcal{O}) - \kappa_{12}\, p_{M_2}(\mathcal{O})}{1-\kappa_{12}}, \qquad p_g(\mathcal{O}) \equiv \frac{p_{M_2}(\mathcal{O}) - \kappa_{21}\, p_{M_1}(\mathcal{O})}{1-\kappa_{21}}, \tag{2.10}$$

where κ_{12}, κ_{21}, $p_{M_1}(\mathcal{O})$, and $p_{M_2}(\mathcal{O})$ are directly obtainable from M_1 and M_2.

There are two immediate points to note about the Operational Definition. First, it does not attempt to define quark and gluon jets at the level of individual jets, but rather it defines them in aggregate as two well-defined probability distributions. This is in keeping with the spirit of the Conceptual Definition in section 2.1, which sought to identify enriched regions of phase space rather than to determine a per-jet truth label. It is also in concert with the basic construction of quantum field theory, which only provides theoretical access to distributional quantities such as cross sections rather than making predictions for individual events.[3]

Second, the Operational Definition does not rely on assumptions of mutual irreducibility of quarks and gluons or the factorization of jet samples as mixtures, instead turning them into derived properties of the definition, as we show below. In the limit where factorization holds and quarks and gluons are mutually irreducible in the feature space $\mathcal{O}$, the Operational Definition returns precisely the quark and gluon jets which make sense in that context. Outside of these potentially-restrictive limits, the definition nonetheless returns two well-defined categories which can be fairly called quark and gluon jets. The Operational Definition essentially takes the vague notion of "quark-like" from the Conceptual Definition and injects mathematical substance by specifying how to extract the quark and gluon distributions.

With the Operational Definition in hand, we now turn the reasoning of section 2.2 on its head to *derive* the mutual irreducibility of quarks and gluons and the mixture nature of the two jet samples M_1 and M_2. Using the quark/gluon jet definition in eq. (2.10), we can write down the quark/gluon reducibility factors as:

$$\kappa_{qg} = \min_{\mathcal{O}} L_{q/g}(\mathcal{O}) = \min_{\mathcal{O}} \frac{(1-\kappa_{21})(L_{M_1/M_2}(\mathcal{O}) - \kappa_{12})}{(1-\kappa_{12})(1-\kappa_{21} L_{M_1/M_2}(\mathcal{O}))} = 0, \tag{2.11}$$

[3] Note that (non-deterministic) per-jet labels can be obtained from this definition if needed. For a jet with observable value O, one can assign it a "quark" label with probability $f\, p_q(O)/(f\, p_q(O) + (1-f)\, p_g(O))$ by using the extracted quark and gluon distributions, p_q and p_g, and extracted quark fraction f of the sample. These labels are universal if the observable is monotonically related to the likelihood ratio.

JHEP11(2018)059

where we have used the monotonicity of $L_{q/g}(\mathcal{O})$ in $L_{M_1/M_2}(\mathcal{O})$ and the definition of κ_{12} to see that the numerator vanishes while the denominator is non-zero. An analogous calculation shows that $\kappa_{gq} = 0$, and therefore that the distributions of quark and gluon jets as defined by the Operational Definition are always mutually irreducible.

Next, we demonstrate that M_1 and M_2 are mixtures of the defined quark and gluon jet distributions. Solving eq. (2.10) for the distributions of M_1 and M_2 in terms of the quark/gluon distributions yields:

$$p_{M_1}(\mathcal{O}) = f_1\, p_q(\mathcal{O}) + (1 - f_1)\, p_g(\mathcal{O}), \qquad f_1 \equiv \frac{1 - \kappa_{12}}{1 - \kappa_{12}\kappa_{21}}, \tag{2.12}$$

$$p_{M_2}(\mathcal{O}) = f_2\, p_q(\mathcal{O}) + (1 - f_2)\, p_g(\mathcal{O}), \qquad f_2 \equiv \frac{\kappa_{21}(1 - \kappa_{12})}{1 - \kappa_{12}\kappa_{21}}, \tag{2.13}$$

where we have introduced two numbers f_1 and f_2 such that $f_1, f_2 \in [0, 1]$. We see from eqs. (2.12) and (2.13) that under the Operational Definition, M_1 and M_2 have the interpretation of being statistical mixtures of quark and gluon jets where the quark fractions of each sample are f_1 and f_2, respectively. Note that while this was entirely anticipated, given the motivation provided in section 2.2, the Operational Definition manages to avoid the circular reasoning of that section, where a well-defined notion of quark and gluon jets and the statistical-mixture nature of M_1 and M_2 were assumed to exist before we were able to specify a rigorous procedure to determine them.

There are several additional properties of the Operational Definition worth noting. First, any additional preprocessing of the jets in M_1 and M_2 which is operationally defined at the hadron level, such as jet grooming, can be folded into the jet-finding procedure and thus incorporated directly into our definition. Second, which of M_1 or M_2 is more "quark-enriched" only serves to label which of the resulting distributions is "quark" and which is "gluon" and does not change the distributions which are produced by this definition. Finally, while eq. (2.10) implies the vanishing of the quark/gluon reducibility factors, if a different, non-zero quark/gluon reducibility factor is desired a priori, then the definition may be suitably modified to accommodate those non-zero values. Thus, the assertion of quark-gluon mutual irreducibility, which is supported by evidence from case studies, can be relaxed to any specified quark/gluon reducibility factors which may then be thought of as inputs to the definition.

In section 3, we connect the Operational Definition to machinery that has already been developed in the jet substructure and statistical literature, finding that the tools needed to implement the Operational Definition, true to the name, are readily available. In appendix A, we gain some additional insight into the Operational Definition by theoretically exploring it with simple jet substructure observables in a tractable limit of perturbative QCD.

3 Data-driven jet taggers and topics

In this section, we connect our Operational Definition of quark and gluon jets to recent developments at the intersection of jet physics and statistical methods, particularly the

JHEP11(2018)059

data-driven paradigms of CWoLa [60] and jet topics [61]. CWoLa provides a method to approximate the quark/gluon likelihood ratio by distilling the available information in a huge feature space of jet substructure observables [60, 64, 65]. The jet topics method was introduced and developed in ref. [61], where it was shown that statistical methods could be used to "disentangle" quark and gluon jets from mixtures. We will show how these methods can be combined to form a concrete implementation of the Operational Definition.

3.1 Classification without labels: training classifiers on collider data

Recently, there has been an effort to train physics classifiers directly on data despite the lack of labeled truth information, going under the broad term of *weak supervision*. Ref. [66] was the first to apply weak supervision methods in a particle physics context, showing that given mixed samples with known signal fractions, a quark/gluon classifier on a few high-level inputs could be trained without access to per-jet truth labels, a paradigm termed learning from label proportions (LLP). Ref. [60] developed CWoLa as a method to train a jet classifier via weak supervision on a few generalized angularities [12–14, 19, 20], where signal fractions do not need to be known in order to train the classifier. Ref. [65] investigated both CWoLa and LLP in the context of high-dimensional, modern machine learning methods, finding that while both methods were performant, CWoLa generalized better and more simply to complex models. CWoLa has since given rise to new techniques to search for signals of new physics in model-independent ways [57]. These methods are an important step towards making classification at colliders fully data-driven. Here, we review the CWoLa paradigm in preparation for incorporating it as part of the implementation of our Operational Definition.

Conceptually, CWoLa is extremely simple: given two mixtures M_1 and M_2 of signal (quark) and background (gluon) jets, train a classifier to distinguish jets in M_1 from jets in M_2. This procedure has the attractive property of being able to immediately use any model which can be trained with full supervision. Furthermore, in the limit that M_1 and M_2 become pure signal and background, CWoLa smoothly approaches full supervision. With enough statistics, a feature space that captures all relevant information, and a suitable training procedure, a CWoLa classifier should approach the optimal discriminant between the two mixed samples.[4] By the Neyman-Pearson lemma [62], the optimal discriminant between two binary classes is the likelihood ratio. As discussed in section 2.2, the mixed-sample likelihood ratio is monotonically related to the quark/gluon jet likelihood ratio. Thus, CWoLa provides a way of approximating the optimal discriminant between quark and gluon jets given access only to mixed samples.

There are potential concerns, though, that one might have regarding CWoLa in particular and weak supervision in general. Are enough statistics and a rich-enough feature space available? Do we have a suitable training procedure? Refs. [60, 64, 65] address these concerns and demonstrate that CWoLa indeed works in realistic cases. For example, CWoLa was used in ref. [65] to obtain a performant quark/gluon jet classifier by discriminating Z+jet and dijet samples using jet images and convolutional neural networks. As described

[4]The generalization to learning from multiple mixtures of signal and background is possible as long as each mixture is assigned a label that is (on average) monotonically related to its signal fraction.

JHEP11(2018)059

in appendix B, there are many other jet representations and machine learning models that are suitable to be trained with CWoLa. Additionally, previous uses of CWoLa have made assumptions about the samples M_1 and M_2 being mixtures of well-defined quark and gluon jets, without specifying which definition is being used or attempting to quantify what happens if quark and gluon jets are not the same in the two samples (i.e. sample dependence). From the perspective of this work, those concerns are removed by using the Operational Definition, which turns the problem on its head and lets the samples M_1 and M_2 *define* quark and gluon jets. The notion of sample dependence manifests in a new way with our Operational Definition, which we discuss more in our conclusions in section 5.

3.2 Jet topics: extracting categories from collider data

Building on a rich analogy between mixed jet samples and textual documents, ref. [61] introduced jet topics and demonstrated how topic modeling could be used to obtain quantitative information about the signal and background distributions from the mixed sample distributions. The present work extends and elaborates on this approach in order to formulate a practical implementation the Operational Definition of quark and gluon jets in section 2.3.

Given two samples of quark and gluon jets M_1 and M_2, the jet topics technique seeks to extract two mutually irreducible categories such that the samples are mixtures of these categories. To the extent that quark and gluon jets are themselves mutually irreducible, they will correspond to the extracted topics. There are various procedures for extracting the topics from mixed samples. Ref. [61] used a method known as "demixing" that was developed in ref. [67] in order to obtain the topics. Other procedures (e.g. non-negative matrix factorization [68]) that are popular for textual topic modeling could in principle also be used. Demixing works by searching for "anchor bins" in the mixed sample distributions over a feature space $\mathcal{O}$, which are histogram bins for which the likelihood of M_1 to M_2 is maximized or minimized.

In the language of section 2.2, demixing returns reducibility factors κ_{12} and κ_{21}. With the reducibility factors in hand, the fractions of topic T_1 in each mixed sample, $f_{T_1}^{(1)}$ and $f_{T_1}^{(2)}$, can be obtained by solving equations analogous to eq. (2.8), and the topic distributions $p_{T_1}(\mathcal{O})$ and $p_{T_2}(\mathcal{O})$ are given by eq. (2.9) where q is replaced by T_1 and g by T_2:

$$p_{T_1}(\mathcal{O}) = \frac{p_{M_1}(\mathcal{O}) - \kappa_{12}\, p_{M_2}(\mathcal{O})}{1-\kappa_{12}}, \qquad f_{T_1}^{(1)} = \frac{1-\kappa_{12}}{1-\kappa_{12}\kappa_{21}}, \tag{3.1}$$

$$p_{T_2}(\mathcal{O}) = \frac{p_{M_2}(\mathcal{O}) - \kappa_{21}\, p_{M_1}(\mathcal{O})}{1-\kappa_{21}}, \qquad f_{T_1}^{(2)} = \frac{\kappa_{21}(1-\kappa_{12})}{1-\kappa_{12}\kappa_{21}}, \tag{3.2}$$

where we have assumed without loss of generality that $f_{T_1}^{(1)} > f_{T_1}^{(2)}$.

The jet topics method provides a simple example of the fascinating mileage one is able to achieve from the picture of jets as statistical mixtures. If the signal (quark) and background (gluon) distributions are mutually irreducible, the topic fractions are the signal fractions, $f_S^{(1)} = f_{T_1}^{(1)}$ and $f_S^{(2)} = f_{T_1}^{(2)}$, from which a number of other useful quantities may be computed. First, consider some observable O that we wish to cut on to make a signal/background classifier. For a given threshold t, let the fraction of jets in M_i for which O is greater than t be $f_{M_i}(O > t)$. Let $\varepsilon_s(t)$ be the rate that the signal is correctly

JHEP11(2018)059

identified (the true positive rate) and $\varepsilon_b(t)$ be the rate that the background is identified as signal (the false positive rate) by the classifier (O, t). We then have the equations:

$$f_{M_1}(O > t) = f_S^{(1)} \varepsilon_s(t) + (1 - f_S^{(1)})\, \varepsilon_b(t), \tag{3.3}$$

$$f_{M_2}(O > t) = f_S^{(2)} \varepsilon_s(t) + (1 - f_S^{(2)})\, \varepsilon_b(t), \tag{3.4}$$

which can be solved to give signal and background efficiencies at the given threshold:

$$\varepsilon_s(t) = \frac{f_{M_1}(O > t)(1 - f_S^{(2)}) - f_{M_2}(O > t)(1 - f_S^{(1)})}{f_S^{(1)} - f_S^{(2)}}, \tag{3.5}$$

$$\varepsilon_b(t) = \frac{f_{M_2}(O > t) f_S^{(1)} - f_{M_1}(O > t) f_S^{(2)}}{f_S^{(1)} - f_S^{(2)}}. \tag{3.6}$$

In this way, the extracted fractions can be used to calibrate the classifier. Additionally, the pure signal and background distributions of any observable can be obtained from the reducibility factors (or equivalently the extracted fractions): simply change the feature space $\mathcal{O}$ in eqs. (3.1) and (3.2) to whatever observable is desired.

There are several issues to address in attempting to use topic modeling for quark and gluon jets. How do we know that quark and gluon jets are mutually irreducible in our feature space? In appendix A, we show that quark and gluon jets are *not* mutually irreducible in the leading-logarithmic limit of individual Casimir-scaling or Poisson-scaling observables, though this calculation strongly suggests that mutual irreducibility could be achieved in a larger feature space. Ref. [61] showed that quark and gluon jets appear to be mutually irreducible in practice for the constituent multiplicity observable, but did not offer a way to fold in additional information. If we attempt to use multiple observables in the topic modeling procedure, how do we deal with the curse of dimensionality that results from attempting to fill multi-dimensional histograms? As we now discuss, CWoLa can be combined with jet topics to efficiently use arbitrarily large feature spaces to determine the optimal quark and gluon jet topics.

3.3 Optimal taggers for optimal topics

To summarize, the CWoLa framework allows trained models to approximate a function monotonic to the quark/gluon likelihood ratio, which is the optimal quark/gluon jet classifier. Further, the jet topics technique allows for signal and background distributions to be extracted from a given low-dimensional feature space. Here, we demonstrate how CWoLa and jet topics can be combined into a direct implementation of the Operational Definition of quark and gluon jets from section 2.3.

When viewed as a likelihood-ratio approximator, a CWoLa-trained model can do more than per-jet classification: it is an efficient method for compressing information in a (potentially) huge but sparsely-populated feature space down to the provably optimal single observable for quark/gluon jet separation. This approach of taking a CWoLa-trained model output as an interesting observable in its own right solves the curse of dimensionality mentioned at the end of section 3.2. Furthermore, the guarantee of optimality for the likelihood

JHEP11(2018)059

ratio by the Neyman-Pearson lemma carries over to the jet topics context in that the mutual irreducibility of quark and gluon jets is maximized when the optimal discriminant is used. In this sense, optimal taggers give rise to optimal topics.

The marriage of CWoLa and jet topics yields more fruit: since the signal fractions extracted by the topics procedure can be used to calibrate a classifier, the requirement that a CWoLa-trained model be calibrated using known signal fractions is removed. A CWoLa model now has the potential to be *self-calibrating* in the sense that the model is used to extract the signal fractions, and then the fractions are used to calibrate that same model (other models can also be calibrated). Furthermore, the optimal topic fractions can be used to extract the pure distribution of any desired observable in a straightforward manner.

This combined paradigm provides a new way to use fully data-driven classifiers in high-energy particle physics, namely as optimal observables for topic fraction extraction. The fully data-driven aspect of the entire procedure cannot be emphasized enough as application of these methods to data is the ultimate goal. The black-box nature of complex classifiers becomes less disturbing in this context, since we can think of the role of the classifier as simply to regress onto the likelihood ratio, without much concern as to how this is done. As with ref. [69], understanding of both the inputs and outputs of a machine learning model allows us to be agnostic with respect to the internal details.

Where does the Operational Definition in section 2.3 fit into this picture? If we adopt the Operational Definition and define quark and gluon jets to be the categories returned by the topic-finding procedure, this addresses the first issue with jet topics referenced at the end of section 3.2, that we do not know the relation between the extracted topics and quark and gluon jets. Also, since under this definition the samples M_1 and M_2 are mixtures of exactly the same quark and gluon jets, the sample dependence concerns mentioned at the end of section 3.1 are alleviated. The optimality guarantee resulting from the Neyman-Pearson lemma and the good practical performance lend support to the Operational Definition being useful both in theory and practice. It is no coincidence that the Operational Definition, CWoLa, and jet topics share the same property: they work well when notions of sample independence and mutual irreducibility exist, but still return something sensible as the situation is detuned away from this nice limit.

4 Quark and gluon jets from dijets and Z+jet

In this section, we apply the combined paradigm of CWoLa and jet topics to the realistic context of Z+jet and dijet samples, obtaining the distributions of quark and gluon jets via the Operational Definition.[5]

4.1 Event generation

We generated events using Pythia 8.230 [71] with the default tunings and shower parameters at $\sqrt{s} = 14\,\text{TeV}$. Hadronization and multiple parton interactions (i.e. underlying event) were included and a parton-level p_T cut of 400 GeV was applied. The Z+jet sample was

[5]We also investigated applying the Operational Definition to CMS jet mass measurements on similar samples [70]. In the dijet sample, though, only average jet mass (instead of individual jet mass) is reported.

JHEP11(2018)059

Symbol	Name	Short Description
n_{const}	Constituent Multiplicity	Number of particles in the jet
n_{SD}	Soft Drop Multiplicity	Probes number of perturbative emissions
N_{95}	Image Activity	Number of pixels containing 95% of jet p_T
$\tau_2^{(\beta=1)}$	2-subjettiness	Probes the two-prong nature of the jet
w	Width	Angularity measuring the girth of the jet
m	Jet Mass	Mass of the jet
PFN-ID	Particle Flow Network with ID	Particle three-momentum + ID inputs
PFN	Particle Flow Network	Particle three-momentum inputs
EFN	Energy Flow Network	Using only IRC-safe information
EFPs	Energy Flow Polynomials	Linear classification with EFPs
CNN	Convolutional Neural Network	Trained on 33×33 2-channel jet images
DNN	Deep Neural Network	Trained on an N-subjettiness basis

Table 1. The individual jet substructure observables (top) and machine learning models (bottom) considered in this study, along with their corresponding symbols and short descriptions. A full discussion of the observables and models is given in appendix B.

obtained using the `WeakBosonAndParton:qg2gmZq` and `WeakBosonAndParton:qqbar2gmZg` processes, ignoring the photon contribution and requiring the Z to decay invisibly. The dijet sample was obtained using the `HardQCD:all` process, excluding bottom quarks.

Final state, non-neutrino particles were clustered with FastJet 3.3.0 [72] using the anti-k_T algorithm [73] with a jet radius of $R = 0.4$. All jets were required to have $p_T \in [500, 550]$ GeV and rapidity $|y| < 2.5$. The hardest jet for Z+jet and the hardest two jets for dijets were considered and kept if they passed the above specified cuts. The unphysical parton-shower-labeled jet flavor was determined by matching the clustered jet to the Pythia parton(s) by requiring that the jet lie within $2R$ of the parton direction from the hard process. Events in which none of the jets passed this criteria were not considered. One million jets passing all cuts were retained for both the dijet and Z+jet samples. The Pythia-labeled quark fraction was 86.3% for the Z+jet sample and 49.8% for the dijet sample.

4.2 Extracting reducibility factors and fractions

For the jet substructure feature space $\mathcal{O}$, we consider a variety of individual jet substructure observables and trained models. In table 1, we summarize the observables and models used for our study. Details of the observable computation, model training, and model architectures are given in appendix B.

For each of the observables and trained models, we proceed to extract the topic fractions from the Z+jet and dijet samples. We implement a version of the demixing procedure used in ref. [61] and described in ref. [67]. Below, we describe the practical procedure used for the studies in this section, including the determination of uncertainties. Here, we let O indicate either a single observable or the output of a trained model.

JHEP11(2018)059

1. *Histograms*: the histograms for $p_{Z+\text{jet}}(O)$ and $p_{\text{dijets}}(O)$ are computed for a specified binning. Statistical uncertainties are taken to be $\sqrt{N_{Z+\text{jet}}}$ and $\sqrt{N_{\text{dijets}}}$ coming from one-sigma count uncertainties within each bin.[6]

2. *Likelihood ratios*: the mixed-sample log-likelihood ratio $\ln p_{\text{dijets}}(O)/p_{Z+\text{jet}}(O)$ is calculated. The statistical uncertainty is estimated from uncertainty propagation per bin to be:

$$\sigma_{\ln p_{\text{dijets}}/p_{Z+\text{jet}}} = \sqrt{\frac{1}{N_{\text{dijets}}} + \frac{1}{N_{Z+\text{jet}}}}. \tag{4.1}$$

3. *Anchor bins*: noisy, low-statistics bins are neglected by only considering bins with more than 50 events in each sample. The upper (lower) anchor bin is obtained by finding the maximum (minimum) bin for the log-likelihood ratio minus (plus) its uncertainty.

4. *Reducibility factors*: the lower (upper) reducibility factor κ_{21} (κ_{12}) is obtained by exponentiating (minus) the log-likelihood ratio evaluated at the lower (upper) anchor bin. Uncertainties on the reducibility factors are obtained by standard uncertainty propagation.

5. *Topics*: the jet topics are obtained from the reducibility factors κ_{12} and κ_{21} according to the definition in eq. (2.10), with uncertainties propagated from the reducibility factors.

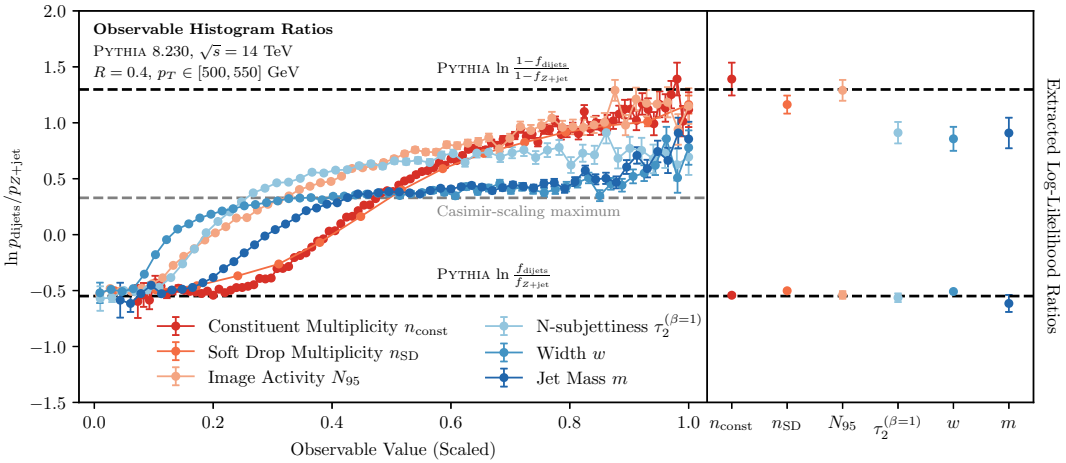
6. *Fractions*: topic fractions are obtained from the reducibility factors κ_{12} and κ_{21} according to eqs. (2.12) and (2.13), with uncertainties propagated from the reducibility factors. In this study, the topic fraction always corresponds to the quark fraction.

While we use the concrete method above to showcase the viability of our method, there may of course be alternative ways to obtain the anchor bins and reducibility factors. For instance, it may be interesting to a pursue a binning-free method, where a cumulative density function is used instead of a binned histogram. Similarly, there may be more suitable ways to ignore low-statistics phase space regions and determine anchor bins. We leave detailed optimizations of the method for future developments.

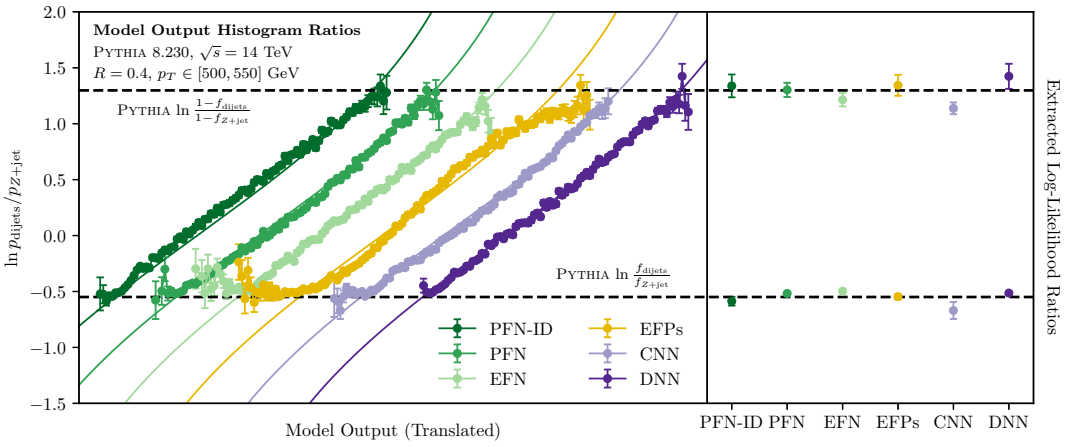
In figure 2, we show the mixed-sample log-likelihood ratios $\ln p_{\text{dijets}}(O)/p_{Z+\text{jet}}(O)$ for various jet substructure observables and model outputs. Overall, we see excellent confirmation that the mixed-sample log likelihood is bounded between the predicted extrema according to the Pythia fractions. To extract these fractions in a data-driven way, we must of course obtain these extrema from the measured log-likelihood ratios. Using the procedure outlined above, the resulting anchor bins are shown in the right-most portion of figure 2. Interestingly and satisfyingly, many of the individual observables and essentially

[6]These uncertainties, and those derived from them, should only be used to give a sense of scale on the plots. Implementing the Operational Definition in LHC data will require careful consideration of other sources of statistical and systematic uncertainties. For instance, using unfolded distributions may mitigate artificial differences in the samples due to detector effects.

JHEP11(2018)059



(a)



(b)

Figure 2. Mixture log-likelihood ratios and their extrema for (a) individual jet substructure observables and (b) trained models, the latter of which have been translated along the horizontal axis for clarity. The black dashed lines indicate the maximum and minimum of the mixture likelihood ratio determined using the Pythia fractions. The gray dashed line in the observable plot indicates the upper bound obtained for a Casimir-scaling observable from appendix A; as expected, jet mass and width approach and remain near the gray line for much of their domain. While all individual observables asymptote well to the lower black line, only the count observables (n_{const}, n_{SD}, N_{95}) come close to the upper black line, indicating that gluons are more irreducible than quarks. By contrast, the minimum and maximum for each trained model appear to achieve extremal values close to the black limits. The solid colored lines in the lower plot indicate the behavior of the optimal classifier, closely related to figure 1.

all of the models extract extrema consistent with the Pythia fractions. It is important to note, though, that the Pythia fractions are not fully well-defined hadron-level concepts and are shown solely to provide a conceptual and semi-quantitative guideline for the performance of the method.

JHEP11(2018)059

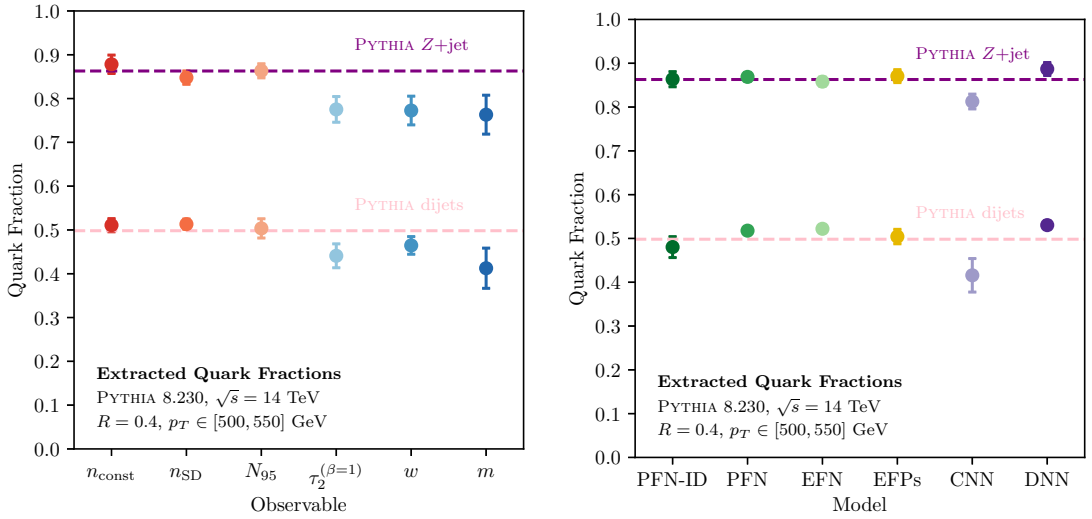
For the substructure observables in figure 2(a), it is evident that the count observables of constituent multiplicity, soft drop multiplicity, and image activity come closest to saturating both the upper and lower bounds. For mass and width, a clear plateau is exhibited close to the leading logarithmic expectation for Casimir-scaling observables (see appendix A). This difference is reflected in the fact that the count observables extract extrema of the log-likelihood ratio consistent with the Pythia fractions, while the remaining observables systematically underestimate the upper bound. One feature worth noting is that the lower bound is accurately extracted by every observable; it is the upper bound that is more difficult to saturate with a generic observable. This indicates that gluon jets are evidently more irreducible than quark jets, and therefore that gluon jet distributions are easier to extract.

For the trained model outputs in figure 2(b), we see that the mixed-sample log-likelihood ratios are clearly bounded as expected and agree with the prediction for a well-trained classifier. The slight deviations from the solid curve in the case of the EFPs arise from the fact that they are trained using Fisher's Linear Discriminant, which optimizes a different objective function, but nonetheless the EFPs exhibit qualitatively similar behavior to the other classifiers. Compared to the individual substructure observables, the models more robustly saturate the upper and lower bounds of the log-likelihood ratio and demonstrate less sensitivity to changes in the binning of the histograms. The extracted extrema of the log-likelihood ratio based on the trained models (with the exception of the CNN) are all consistent with one another as well as with the Pythia fractions. This agreement, present in the variety of different models which process information in very different ways, indicates that there is indeed a robust sense in which "quark" and "gluon", as qualitatively described by the parton-matched labels, are latent within the mixed samples.

Using the extracted extrema of the mixed-sample log-likelihood ratio, the reducibility factors can be obtained by appropriate exponentiation. The quark fractions can then be calculated according to eqs. (2.12) and (2.13). These are shown in figure 3(a) for the individual observables and figure 3(b) for the trained models. We see that the trained models all extract fractions largely consistent with one another and with the Pythia fractions. The count substructure observables also extract consistent fractions, while the shape observables exhibit Casimir-scaling behavior, making them unsuitable for identifying mutually-irreducible quark and gluon jets. The fractions obtained from the trained models were consistently more robust to different choices of topic extraction procedures, such as the histogram binning. Despite having little to no handle on the details of the trained models, we are able to obtain important constraints on their behavior and use them to obtain quark/gluon fractions, which are evidently insensitive to these details.

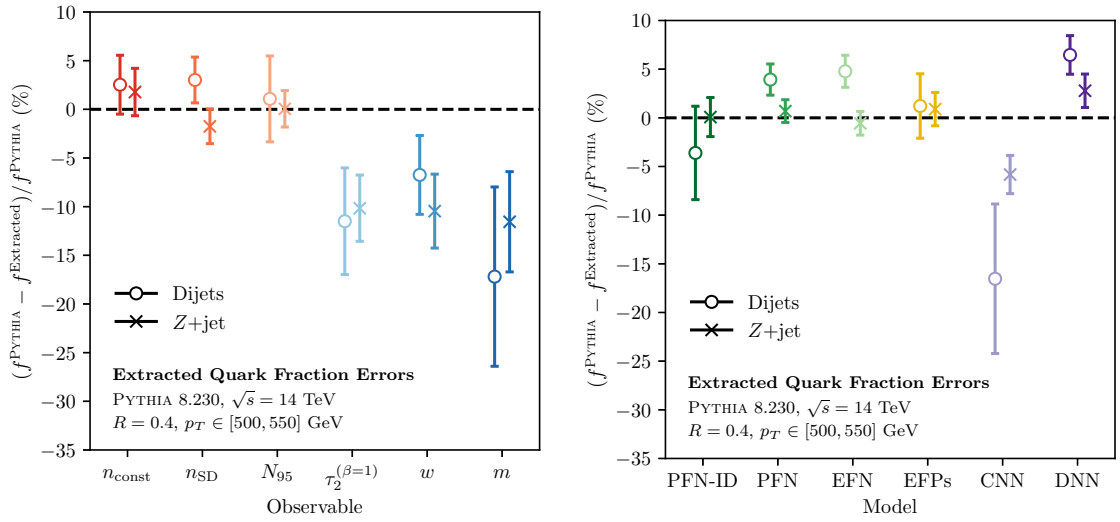
As a more quantitative measure of the quality of the extracted quark fractions, the percent error of the extracted fractions relative to the (unphysical) Pythia fractions is shown in figures 4(a) and 4(b). The count observables and trained models agree within several statistical uncertainties of one another and the Pythia fractions, in many cases achieving $\mathcal{O}(1\%)$ fidelity. Again, we caution that the Pythia fractions solely provide a heuristic to demonstrate the performance of the method and should not be taken as fundamental to quark and gluon jets.

JHEP11(2018)059



(a) (b)

Figure 3. Extracted quark fraction values for the (a) individual observables and (b) trained models as calculated using the log-likelihood extrema of figure 2 inserted into in eqs. (2.12) and (2.13) to obtain the fractions.



(a) (b)

Figure 4. The percent error of the extracted quark fractions (see figure 3) relative to the Pythia fractions, obtained using the (a) individual observables and (b) trained models. By this measure, the best jet observable appears to be N_{95} and the best model is the linear EFP model.

4.3 Self-calibrating classifiers

With the quark fractions of the mixtures in hand, one immediate application is to use them to calibrate the quark/gluon classifiers, as discussed in section 3.3. Since uncalibrated classifiers can be used to obtain these fractions, this allows for self-calibrating classifiers in

JHEP11(2018)059

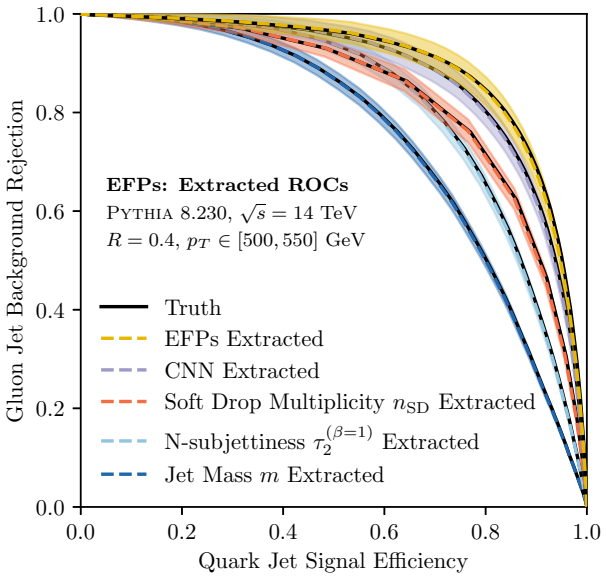


Figure 5. The ROC curves for several substructure observables and trained models using the quark fractions estimated from the EFPs. The "Truth" corresponds to using the Pythia fractions to obtain the ROC curve. We see good agreement between the data-driven ROC curves and the Pythia-labeled ROC curves. Further, we see that the CWoLa-trained EFP classifier has effectively self-calibrated itself in this way.

the CWoLa framework. This liberates the CWoLa framework from necessarily requiring a small test set with known fractions (cf. ref. [60]). In the present picture, this ability to self-calibrate is conceptually clear since a sample with "known" fractions is equivalent to providing a definition of the underlying categories.

Beyond solely self-calibration of classifiers, the extracted fractions can be used to obtain the receiver operating characteristic (ROC) curves for other trained models or substructure observables, even those that do not themselves exhibit quark/gluon mutual irreducibility. The extracted ROC curves of a variety of trained model and substructure observables using the EFP-extracted quark fractions are shown in figure 5, with estimated uncertainty bands coming from uncertainties in the extracted fractions. They are compared to the calibrated ROC curve using the Pythia-labeled fractions, achieving very good agreement between the two. Note that the uncertainties are smaller for worse classifiers, which is intuitive given the limit that a perfectly-random classifier can be identified as such without any fraction information. Overall, this concretely demonstrates that the self-calibration of CWoLa-trained classifiers can be achieved in a purely data-driven way.

4.4 Obtaining observable distributions from extracted fractions

With the reducibility factors of the mixtures, the distributions of substructure observables can be extracted for quark and gluon jets separately. This corresponds to a direct application of the Operational Definition of quarks and gluons in eq. (2.10). This is similar in spirit to the procedure implemented in refs. [52, 55] of using quark/gluon fractions estimated by

JHEP11(2018)059

convolving matrix elements and parton distribution functions and then solving systems of linear equations. The key distinction is that, in our case, the fractions (and reducibility factors) themselves are data-driven.

In figure 6, we use the reducibility factors defined by the EFP classifier to extract quark and gluon distributions for the six individual substructure observables. We see excellent agreement between the data-driven, operationally-defined quark and gluon distributions and the ones specified by the Pythia fractions. Importantly, this procedure works for any substructure observable, even ones such as jet mass and width which do not manifest quark/gluon mutual irreducibility.

5 Conclusions

In this paper, we provided an Operational Definition of quark and gluon jets, based solely on physical cross section measurements. We connected our definition to the existing CWoLa and jet topics paradigms, showing how they each fit naturally into the implementation of the definition. Taking two mixed samples, for which there is a qualitative notion that one is more "quark-like" than the other, the Operational Definition returns a quantitive understanding through mutually-irreducible quark and gluons distributions. Practically, we implemented this definition by approximating the mixed-sample likelihood ratio, relating it to the pure quark/gluon likelihood ratio, and finding its extrema to determine mixed-sample reducibility factors. With the reducibility factors in hand, the quark fractions for the mixed samples can be readily obtained. In a broad sense, our Operational Definition harmonizes with the statistical picture of jet samples at colliders, where individual jets do not carry intrinsic flavor labels and one only ever has access to mixed samples in data.

To illustrate the power of the Operational Definition, we tested it in the realistic context of Z+jet and dijet processes. We applied our quark/gluon jet definition to twelve different observables: six individual substructure observables, and six trained machine learning models which distilled a huge feature space down to a single optimal observable. The six individual observables naturally fall into two categories, count and shape observables, and we confirmed that the count observables yield much more accurate quark fractions (relative to a Pythia baseline). With the minor exception of the CNN, the machine learning models all did well at extracting the fractions. While the performance of the best individual observable (N_{95}) and the best machine learning model (linear EFPs) were comparable, the machine learning models were overall more robust to changes in histogram binning and to the technique used for determining the reducibility factors; this in turn contributes to the robustness of the Operational Definition. Having determined the quark fractions, we extracted pure quark and gluon distributions for various jet substructure observables. Crucially, this worked even for observables that do not exhibit quark/gluon mutually irreducibility, as long as the observable used to extract the fractions does. Additionally, we demonstrated that CWoLa classifiers could be self calibrated using fractions obtained from an uncalibrated classifier, thereby removing a potential hurdle in using CWoLa in practice.

The techniques in this paper represent a novel use of classification in particle physics. Instead of tagging quark and gluon jets, we used a CWoLa-trained deep learning classifier

JHEP11(2018)059

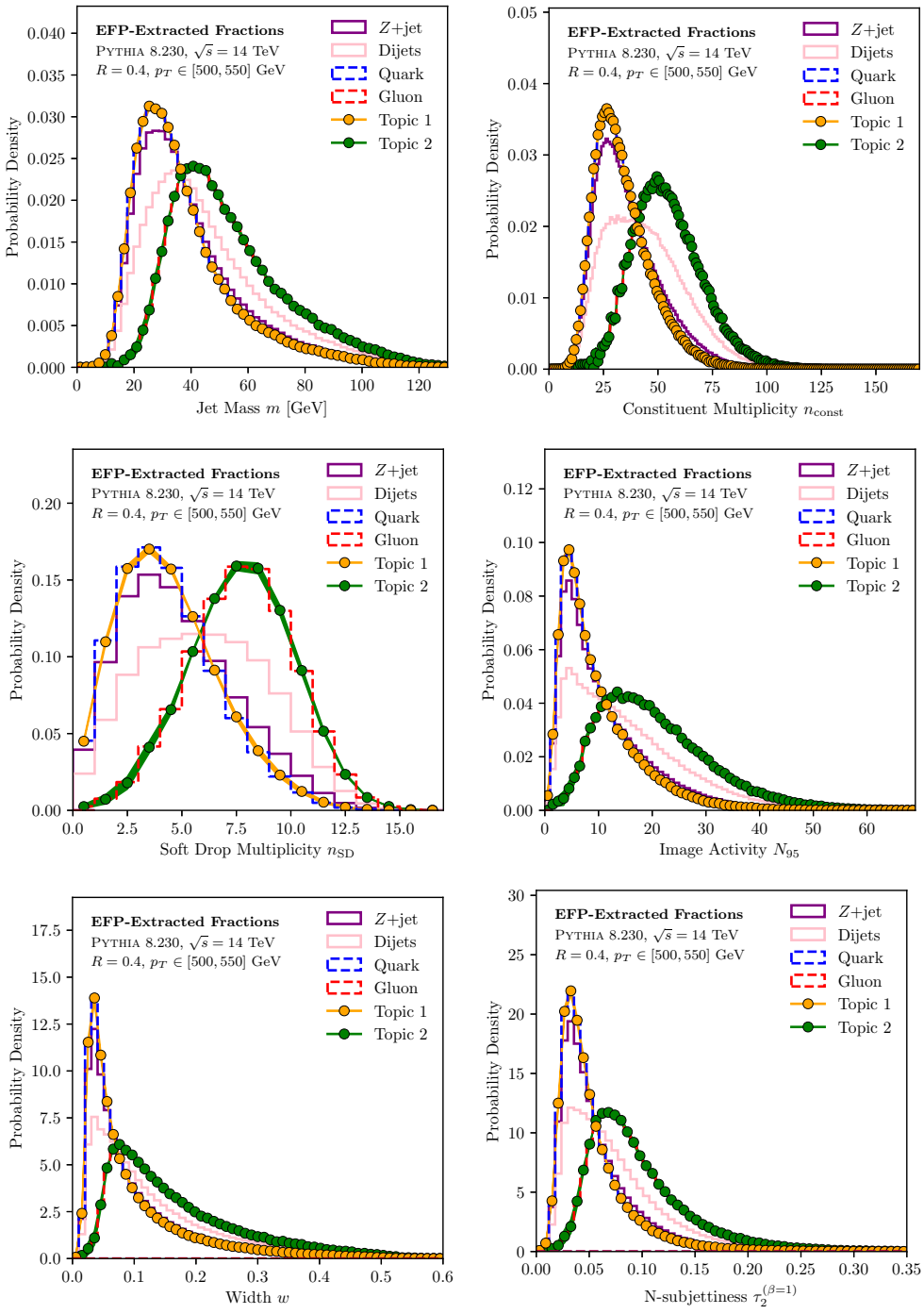


Figure 6. The distributions of the six substructure observables in the Z+jet sample (purple) and dijet sample (pink), with the quark and gluon distributions determined from the Pythia fractions (blue and red, respectively) and the jet topics (orange and green) using EFP-extracted reducibility factors. We see excellent agreement between the jet topics and the Pythia-determined distributions of quark and gluon jets.

JHEP11(2018)059

to approximate the full mixed-sample likelihood ratio. This is in the same spirit of recent work on deep learning [22, 69, 74–80], where the "black box" nature of the trained model is not of central importance to the success or understanding of the method. No longer is the output of a neural network viewed as an arbitrary quantity used only for discrimination, but rather as a robust approximation to the likelihood ratio, which turns out also to be optimal for extracting categories from the data. Surprisingly, while individual quark and gluon jets cannot be tagged perfectly, we were able to use a data-driven classifier to extract the full quark and gluon distributions of an observable to percent-level accuracy. This approach paves the way for fully data-driven collider physics, making use of machine learning techniques trained directly on data while producing results insensitive to the details of the "black box".

We conclude by discussing potential extensions of the methods used in this paper. As mentioned in section 3, a key concern in jet tagging is sample dependence, i.e. whether a "quark jet" in one sample is the same as a "quark jet" in another. While the Operational Definition sidesteps the issue of sample dependence in the case of two mixed samples, it is natural to ask what happens with three or more mixed samples. Concretely, once the Operational Definition is applied to two mixed jet samples, one can ask the degree to which a third sample M is explained by the existing quark and gluon distributions. It turns out that there is a unique and well-defined generalization of the reducibility factor, discussed in ref. [67], that precisely captures this notion and yields a quantifiable notion of sample dependence:

$$\kappa \equiv \max_{f_q,\, f_g} \{ f_q + f_g \,|\, \exists \text{ dist. } p_o(\mathcal{O}) \text{ s.t. } p_M(\mathcal{O}) = f_q p_q(\mathcal{O}) + f_g p_g(\mathcal{O}) + (1 - f_q - f_g) p_o(\mathcal{O}) \}, \quad (5.1)$$

where $0 \leq f_q, f_g \leq 1$ and $f_q + f_g \leq 1$. In eq. (5.1), κ is the maximum amount of M explainable by the quark and gluon distributions, requiring minimal addition of an "other" distribution $p_o(\mathcal{O})$. Understanding sample dependence is a general challenge, even with parton-shower-extracted templates, so it is gratifying that our framework naturally suggests a tool to address this problem. Sample dependence can also be studied by directly comparing the quark and gluon jet definitions provided by different pairs of jet samples (Z+jet, dijets, γ+jet, etc.) at different transverse momenta and jet radii. We leave explorations of these important ideas, as well as more detailed optimizations of the method, to future work.

Extending this thinking, one might attempt to provide a concrete jet flavor definition beyond the two-category case of quarks and gluons. For instance, while the difference in radiation patterns between different-flavor light-quark jets is much smaller than between quark and gluon jets, it may be possible to use the techniques described in this paper to define differently-flavored quark jets. The subtle difference in radiation patterns between different light-quark has been studied in the context of jet charge observables in ref. [17] and in the context of machine learning in ref. [81]. To use our methods in this case would require advances in multiple-category CWoLa and jet topics, though the conceptual underpinnings would be the same as for the two-category case studied here. Further, one could extend such a definition to provide well-defined jet flavor definitions for a variety of

JHEP11(2018)059

other boosted hadronic objects, potentially including subtle distinctions like longitudinal versus transverse polarization of boosted W/Z bosons. More broadly, the concept of mutual irreducibility as a means of defining categories may find additional applications in high-energy physics due to its utility in disentangling overlapping distributions using pure phase space signatures.

Acknowledgments

The authors would like to thank Jonathan Butterworth, Philip Harris, Andrew Larkoski, Benjamin Nachman, Matthew Schwartz, and Clayton Scott for insightful comments and helpful discussions. This work was supported by the Office of Nuclear Physics of the U.S. Department of Energy (DOE) under grant DE-SC-0011090 and the DOE Office of High Energy Physics under grant DE-SC-0012567. Cloud computing resources were provided by a Microsoft Azure for Research Award.

A Theoretical exploration of Casimir- and Poisson-scaling observables

In this appendix, we explore the Operational Definition of quark and gluon jets in the leading-logarithmic (LL) limit, focusing on two theoretically-tractable classes of jet observables: casimir-scaling and Poisson-scaling observables. Though we only work to lowest non-trivial order, these calculations demonstrate that our framework for defining quark and gluon jets is suitable to theoretical exploration in addition to practical experimental implementation. In the LL limit of perturbative QCD, quarks and gluons differ in their emission profiles only by their color charges: $C_F = 4/3$ for quarks and $C_A = 3$ for gluons. Thus, in the LL limit, quarks and gluons are well-defined (at least at the parton level), providing a simplified context to explore the Operational Definition. We find different non-zero quark/gluon reducibility factors for Casimir-scaling and Poisson-scaling observables, substantiating the need to use a richer space of jet substructure observables to approximate the full likelihood ratio.

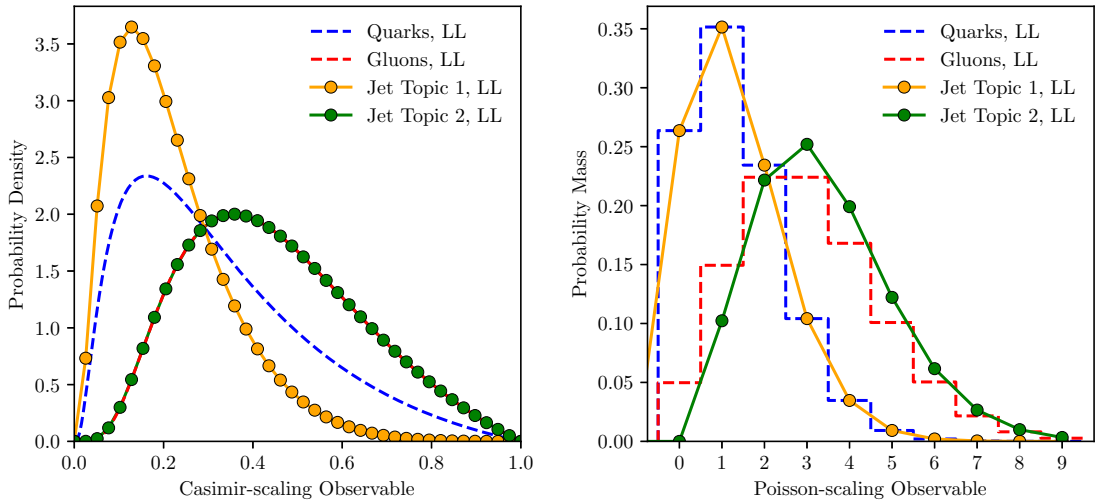
Casimir-scaling observables include common jet substructure observables, such as the jet mass m or IRC-safe angularities [12–14, 19, 20], that are dominated at LL accuracy by a single hard emission. Their cumulative distributions satisfy $\Sigma_g(m) = \Sigma_q(m)^{C_A/C_F}$, where $p_i(m) = d\Sigma_i/dm$. Solely using this scaling property, the quark/gluon reducibility factors of Casimir-scaling observables are:

$$\kappa_{qg}^{\text{Cas.}} = \min_m \frac{p_q(m)}{p_g(m)} = \min_m \frac{\frac{d\Sigma_q}{dm}}{\frac{C_A}{C_F}\Sigma_q^{C_A/C_F-1}\frac{d\Sigma_q}{dm}} = \frac{C_F}{C_A}\min_m \Sigma_q^{1-C_A/C_F} = \frac{C_F}{C_A}, \tag{A.1}$$

$$\kappa_{gq}^{\text{Cas.}} = \min_m \frac{p_g(m)}{p_q(m)} = \min_m \frac{\frac{C_A}{C_F}\Sigma_q^{C_A/C_F-1}\frac{d\Sigma_q}{dm}}{\frac{d\Sigma_q}{dm}} = \frac{C_A}{C_F}\min_m \Sigma_q^{C_A/C_F-1} = 0, \tag{A.2}$$

where $C_A/C_F > 1$ and $\min_m \Sigma_i(m) = 0$ have been used to obtain the last equality. These results are universal to all Casimir-scaling observables and are independent of the remaining details of the observables at LL accuracy.

JHEP11(2018)059



(a) (b)

Figure 7. Quark and gluon distributions at LL accuracy for (a) Casimir-scaling and (b) Poisson-scaling observables, together with the corresponding jet topics. The reducibility of the quark Casimir-scaling distribution and the gluon Poisson-scaling distribution are evident. While neither of these observables individually results in mutually irreducible quarks and gluons, considering them jointly does.

The non-zero reducibility factor in eq. (A.1) indicates that quark and gluon jets are not mutually irreducible in the space of Casimir-scaling observables. In particular, the quark distribution of any Casimir-scaling observable is a mixture of the (irreducible) gluon distribution and some other distribution, as shown in figure 7(a). Note that this *does not* imply that quark jets are fundamentally reducible, since this is just a property derived from Casimir-scaling observables in the LL limit. That said, as noted at the end of section 2.3, if eq. (A.1) were fundamental to quark and gluon jets, one could simply include this reducibility factor in the Operational Definition.

We next consider Poisson-scaling observables, which count the number of perturbative emissions and have qualitatively different quark-gluon reducibility factors. One example is the soft drop multiplicity n_{SD} [82], which counts the number of emissions restricted to a certain phase space region. At LL, Poisson-scaling observables are distributed according to Poissonian distributions with means $C_F\lambda$ for quarks and $C_A\lambda$ for gluons, where λ is a constant proportional to the area of the emission plane that is counted. The quark-gluon reducibility factors corresponding to these distributions are then:

$$\kappa_{qg}^{\text{Pois.}} = \min_n \frac{p_q(n)}{p_g(n)} = \min_n \frac{(C_F\lambda)^n e^{-C_F\lambda}}{(C_A\lambda)^n e^{-C_A\lambda}} = e^{-(C_F-C_A)\lambda} \min_n \left(\frac{C_F}{C_A}\right)^n = 0, \tag{A.3}$$

$$\kappa_{gq}^{\text{Pois.}} = \min_n \frac{p_g(n)}{p_q(n)} = \min_n \frac{(C_A\lambda)^n e^{-C_A\lambda}}{(C_F\lambda)^n e^{-C_F\lambda}} = e^{-(C_A-C_F)\lambda} \min_n \left(\frac{C_A}{C_F}\right)^n = e^{-(C_A-C_F)\lambda}, \tag{A.4}$$

since $C_A/C_F > 1$ and n can take any non-negative integer value.

JHEP11(2018)059

Evidently, Poisson-scaling observables display the *opposite* behavior of Casimir-scaling observables: the gluon distribution is a mixture of the (irreducible) quark distribution and some other distribution, as shown in figure 7(b). Further, the reducibility factor is not universal to all Poisson-scaling observables but rather depends exponentially on the parameter λ. Though $\lambda \sim \mathcal{O}(1)$ was considered in ref. [82], perturbative QCD allows for arbitrarily large λ by counting emissions in larger and larger regions. As λ increases, the reducibility factor falls to zero much more quickly than the overlap in the distributions decreases, and thus quark and gluon jets rapidly approach mutual irreducibility. While perturbative control is lost for large λ due to non-perturbative effects, considering this limit suggests that there is no fundamental impediment to the mutual irreducibility of quarks and gluons from the perspective of perturbative QCD, at least at LL accuracy.

From these two classes of observables, we see that enriching the feature space beyond individual Casimir-scaling and Poisson observables to $\mathcal{O} = \{m, n_{\text{SD}}\}$ yields $\kappa_{qg} = \kappa_{gq} = 0$ for the combined feature space in the LL limit. This benefit of using a rich feature space motivates our approach of training data-driven classifiers on complete substructure information to probe the full quark/gluon jet likelihood ratio, rather than relying on individual specially-crafted substructure observables.

B Details of observables and machine learning models

In this appendix, we give details for the jet substructure study in section 4, describing the observables, machine learning models, and model training used.

For the individual substructure observables, three of them use custom implementations: constituent multiplicity n_{const}, image activity N_{95} [33] (number of pixels in a 33×33 jet image containing 95% of the p_T), and jet mass m. The remaining three observables are computed using FastJet contrib 1.033 [83]. The RecursiveTools 2.0.0-beta1 module is used to calculate soft drop multiplicity n_{SD} [82] with parameters $\beta = -1$, $z_{\text{cut}} = 0.005$, and $\theta_{\text{cut}} = 0$. The Nsubjettiness 2.2.4 module is used to calculate the N-subjettiness [15, 16] observables $\tau_N^{(\beta)}$ with k_T axes as recommended in ref. [84], in particular $\tau_2^{(\beta=1)}$ and jet width w (implemented as $\tau_1^{(\beta=1)}$).

For our trained models, we use several different jet representations and machine learning architectures. In reverse order compared to table 1, they are:

- *DNN*: the N-subjettiness basis [84] is a phase space basis in the sense that $3K-4$ independent N-subjettiness observables map non-linearly onto K-body phase space. We use 20-body phase space consisting of the following set of N-subjettiness basis elements:

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

  i.e. $\tau_N^{(\beta)}$ with $N \in \{1, \ldots, 19\}$ and $\beta \in \{1/2, 1, 2\}$, except $\tau_{19}^{(2)}$ is absent, all computed using the Nsubjettiness 2.2.4 module of FastJet contrib 1.033. A DNN consisting of three 100-unit fully-connected layers and a 2-unit softmaxed output was trained on the N-subjettiness basis inputs.

JHEP11(2018)059

- *CNN*: the jet images approach [85] treats calorimeter deposits as pixel intensities and represents the jet as an image. Convolutional neural networks (CNNs) are the typical model of choice when learning from such a representation, and have been successfully implemented for quark/gluon discrimination [39], W tagging [86], and top tagging [87, 88]. We calculate 33×33 jet images spanning $2R \times 2R$ in the rapidity-azimuth plane. In the language of ref. [39], we formulate "color" jet images with two channels: the p_T per pixel and the multiplicity per pixel. Images were standardized by subtracting the mean and dividing by the per-pixel standard deviation of the training set.

  A CNN architecture similar to that used in ref. [39] was employed: three convolutional layers with 48, 32, and 32 filters and filter sizes of 8×8, 4×4, and 4×4, respectively, followed by a 128-unit dense layer. Maxpooling of size 2×2 was performed after each convolutional layer with a stride length of 2. The dropout rate was taken to be 0.1 for all convolutional layers and was not used for the dense layer.

- *EFPs*: the Energy Flow basis [22] is a linear basis for IRC-safe observables in the sense that any IRC-safe observable is arbitrarily well approximated by a linear combination of Energy Flow Polynomials (EFPs). As a result of this remarkable property, linear methods can be used for classification and regression and are highly competitive with modern machine learning methods. The `EnergyFlow` 0.8.2 package [89] was used to compute EFPs up to $d \leq 7$, $\chi \leq 3$ with $\beta = 0.5$ using the normalized default hadronic measure. This yields 996 EFPs in total, including the trivial constant EFP. This set was used to train a Fisher's Linear Discriminant model with scikit-learn [90].

- *EFN, PFN, PFN-ID*: various particle-level network architectures have been proposed to take advantage of the structure of events or jets as sequences of vectors [41, 69, 91–94]. We choose to focus on the Energy Flow Networks (EFNs) recently introduced in ref. [94] and shown to be competitive with other particle-level models. The EFN architecture is designed to have the properties desirable of a model that takes jet constituents as inputs: it is able to handle variable length lists but, critically, is manifestly symmetric under permutations of the elements in the input. The inputs to an EFN are lists of particles, where a particle is described by its energy fraction, rapidity, and azimuthal angle (the latter two translated to the origin according to the E-scheme jet axis). EFNs construct an internal latent representation of the jet using the particle-level inputs, weighting each particle's contribution by its energy fraction in order to ensure the IRC safety of the internal observables, and then combine the internal jet observables using a DNN backend. The `EnergyFlow` package contains an implementation of EFNs.

  The EFN architecture can be generalized to learn potentially IRC-unsafe internal observables. This variant is termed a Particle Flow Network (PFN), which can easily incorporate additional particle features such as flavor information; see ref. [94] for a more thorough discussion. In addition to the IRC-safe EFN, our study uses a PFN with only kinematic inputs, and a PFN-ID with both kinematic and particle flavor (or

JHEP11(2018)059

ID) information. For each network, the per-particle frontend subnetwork has three fully-connected 100-unit layers corresponding to an internal latent representation of 100 jet observables, and the per-jet backend has three fully-connected 100-unit layers that combines the internal latent observables. The EFN, PFN, and PFN-ID networks differ only in their inputs and whether the energy fractions are used as weights for the internal sum over particles (for the EFN) or passed to the frontend subnetwork (for the PFN and PFN-ID).

All of the above models (excepting the linear EFPs) were implemented and trained using Keras [95] with the TensorFlow [96] backend. Training/validation and test datasets were each constructed using 500,000 events for each jet sample being considered. The training/validation dataset is further divided with 90% used for training and the remaining 10% used for validation. Properties common to all networks were the use of ReLU activations [97] for each non-output layer, a 2-unit softmaxed output layer, He-uniform initialization [98] of the model weights, the categorical crossentropy loss function, the Adam optimization algorithm [99], a learning rate of 0.001, and a patience parameter of 10 epochs monitoring the validation loss. Models are trained 25 times, making use of different random weight initializations, and the best one is selected according to the maximum Area Under the (mixed sample ROC) Curve. The hyperparameters of each model were not optimized for either classification performance or accuracy of the ultimately extracted fractions but rather are demonstrative of typical performance that can be achieved. Practical users of the Operational Definition should tune the hyperparameters for their own purpose.

Finally, it should be noted that other data-driven criteria can be used to select optimal trained models, though we do not explore this further here. One idea is that since the regions of the ROC curve that are relevant for topic extraction are those with very low and very high signal efficiency, in practice it may be beneficial to optimize training for these regions directly. A method for optimizing loss-function based training by operating point is described in ref. [100], and it would be fascinating to explore this for training better models for topic extraction.

C Sample dependence in parton shower events

In this appendix, we do a basic study of sample dependence of Pythia-labeled quark and gluon jets arising from the Z+jet and dijets processes. While this is largely tangential to the main direction of the paper, it lends evidence that our case study is not far from the limit of factorized and universal notions of "quark" and "gluon" jets. Of course, these conclusions are limited by the fact that they come from jets generated in Pythia, which itself relies on notions of factorization in its generation process. A study of these effects in data would be an important addition to our understanding of sample independence and factorization more broadly. We leave a study using our flavor definition to probe sample dependence in a more realistic collider setting to future work.

In figure 8, we plot distributions for the six individual substructure observables, from both the Z+jet and dijet samples, showing the distributions separately for quarks and

JHEP11(2018)059

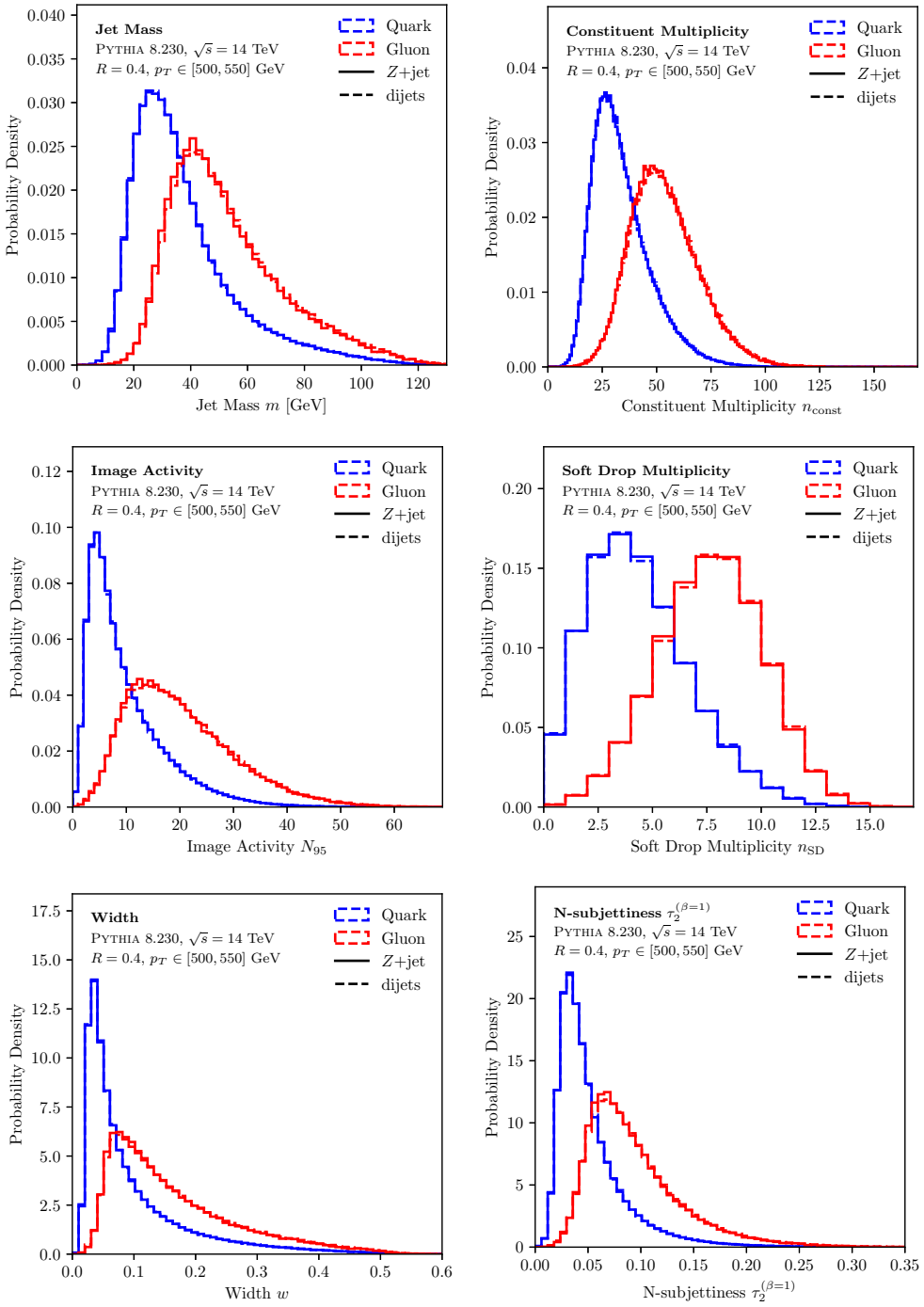


Figure 8. Distributions for the six individual jet observables for Z+jet quarks (solid blue), Z+jet gluons (solid red), dijet quarks (dashed blue), and dijet gluons (dashed red). That the quark and gluon histograms for the two different samples are remarkably similar for this array of observables indicates a high degree of sample independence, at least for the notion of quarks and gluons in Pythia.

JHEP11(2018)059

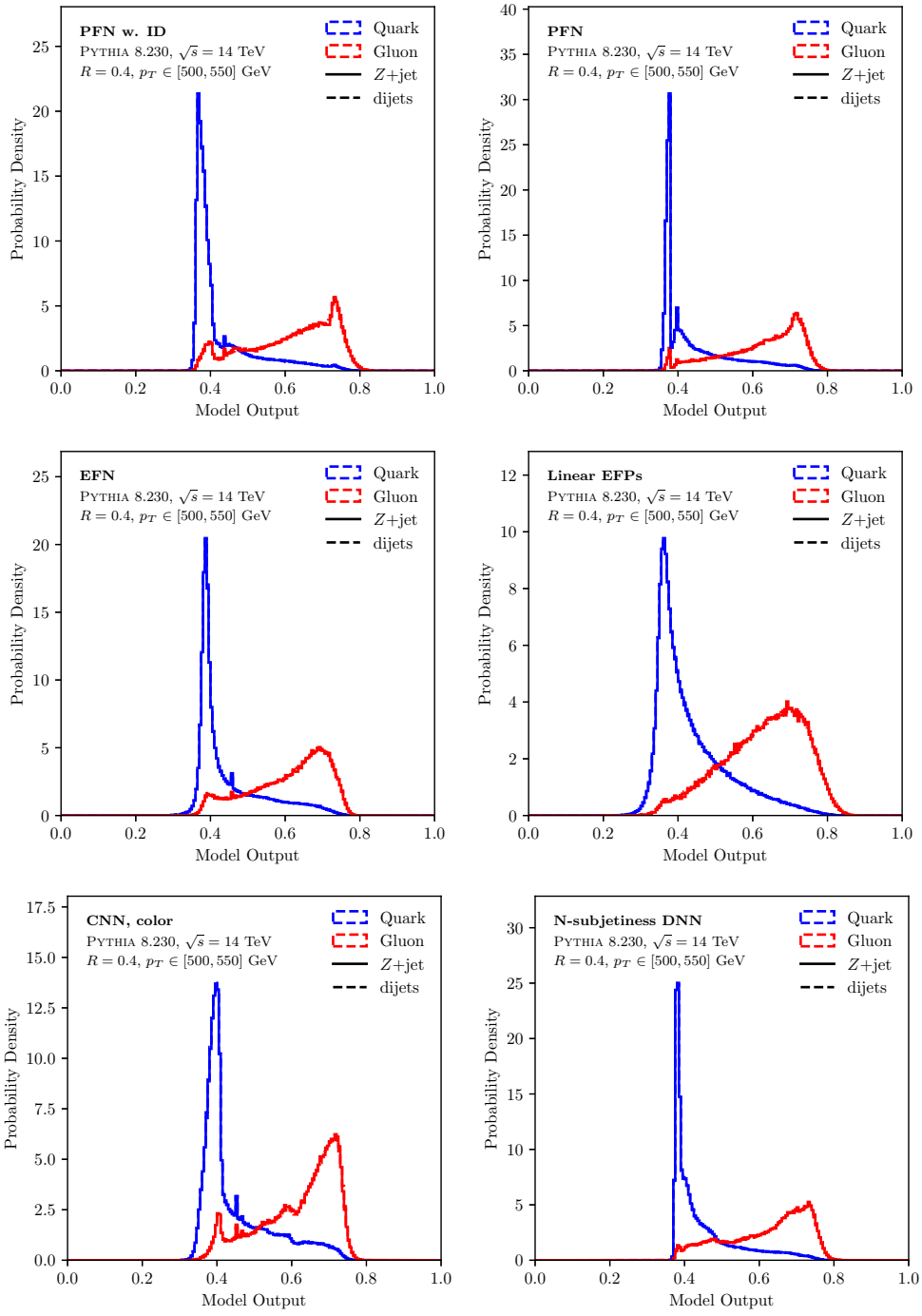


Figure 9. Same as figure 8 but for the six trained model outputs.

JHEP11(2018)059

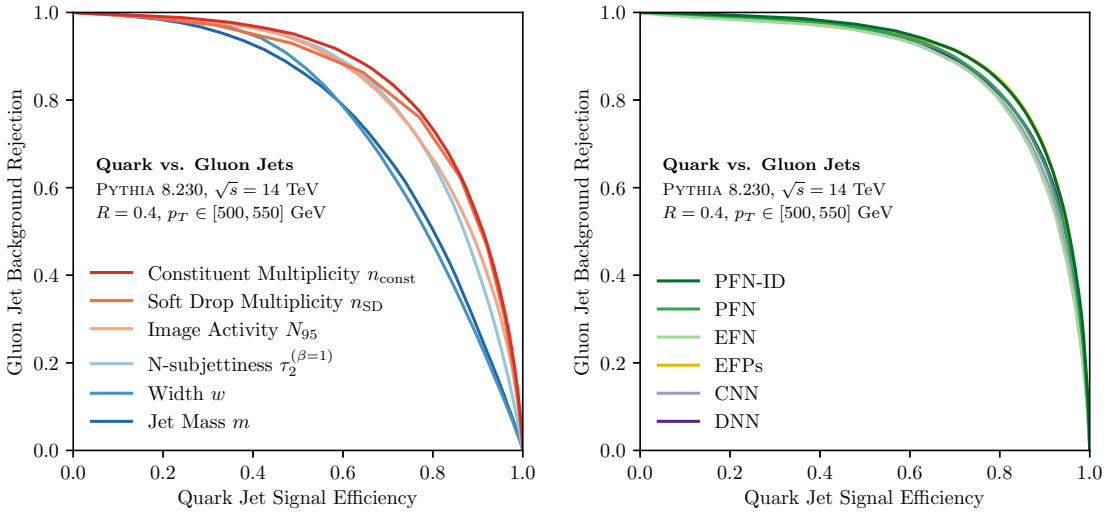


Figure 10. ROC curves for (a) the individual jet observables and (b) the different models trained in the CWoLa paradigm to discriminate Z+jet and dijets events. These are calibrated using Pythia truth fractions. The two best models are PFN-ID and EFPs, which are essentially on top of each other.

gluons as labeled by the Pythia hard scattering process. Importantly, these distributions show a high degree of sample independence: the Z+jet and dijet quarks and gluons have very similar distributions. In figure 9, we plot the distributions of the trained model outputs for quarks and gluons from both the dijet and Z+jet samples. Similar to the standard jet observables in figure 8, a high degree of sample independence is observed. This is perhaps more surprising than for the individual observables because these models have the ability to pick up on very slight differences as part of their training. The observed amount of sample independence is encouraging for using CWoLa and jet topics with complicated models.

For completeness, we also show ROC curves for each of the observables and trained models in figure 10, calibrated using the Pythia fractions. Specifically, we use the Pythia-labeled quark fractions of the Z+jet and dijet samples to calibrate the classifier ROC curve via eqs. (3.6) and (3.5). In figure 10(a), we show ROC curves for each individual observable. In figure 10(b), we show ROC curves for each of the trained models.

References

[1] M.H. Seymour, *Tagging a heavy Higgs boson*, in the proceedings of the *Large Hadron Collider Workshop*, October 4–9, Aachen, Germany (1990).

JHEP11(2018)059

[2] M.H. Seymour, *Searches for new particles using cone and cluster jet algorithms: a comparative study*, *Z. Phys.* **C 62** (1994) 127 [INSPIRE].

[3] J.M. Butterworth, B.E. Cox and J.R. Forshaw, *WW scattering at the CERN LHC*, *Phys. Rev.* **D 65** (2002) 096014 [hep-ph/0201098] [INSPIRE].

[4] J.M. Butterworth, J.R. Ellis and A.R. Raklev, *Reconstructing sparticle mass spectra using hadronic decays*, *JHEP* **05** (2007) 033 [hep-ph/0702150] [INSPIRE].

[5] J.M. Butterworth, A.R. Davison, M. Rubin and G.P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys. Rev. Lett.* **100** (2008) 242001 [arXiv:0802.2470] [INSPIRE].

[6] 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].

[7] 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].

[8] A. Altheimer et al., *Boosted objects and jet substructure at the LHC. Report of BOOST2012, held at IFIC Valencia, 23rd-27th of July 2012*, *Eur. Phys. J.* **C 74** (2014) 2792 [arXiv:1311.2708] [INSPIRE].

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

[10] 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].

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

[12] C.F. Berger, T. Kucs and G.F. Sterman, *Event shape/energy flow correlations*, *Phys. Rev.* **D 68** (2003) 014012 [hep-ph/0303051] [INSPIRE].

[13] L.G. Almeida et al., *Substructure of high-p_T jets at the LHC*, *Phys. Rev.* **D 79** (2009) 074017 [arXiv:0807.0234] [INSPIRE].

[14] S.D. Ellis et al., *Jet shapes and jet algorithms in SCET*, *JHEP* **11** (2010) 101 [arXiv:1001.0014] [INSPIRE].

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

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

[17] 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].

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

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

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

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

JHEP11(2018)059

[22] 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].

[23] D. Krohn, J. Thaler and L.-T. Wang, *Jet trimming*, *JHEP* **02** (2010) 084 [arXiv:0912.1342] [INSPIRE].

[24] S.D. Ellis, C.K. Vermilion and J.R. Walsh, *Recombination algorithms and jet substructure: pruning as a tool for heavy particle searches*, *Phys. Rev.* **D 81** (2010) 094023 [arXiv:0912.0033] [INSPIRE].

[25] S.D. Ellis, C.K. Vermilion and J.R. Walsh, *Techniques for improved heavy particle searches with jet substructure*, *Phys. Rev.* **D 80** (2009) 051501 [arXiv:0903.5081] [INSPIRE].

[26] M. Dasgupta, A. Fregoso, S. Marzani and G.P. Salam, *Towards an understanding of jet substructure*, *JHEP* **09** (2013) 029 [arXiv:1307.0007] [INSPIRE].

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

[28] H.P. Nilles and K.H. Streng, *Quark-gluon separation in three jet events*, *Phys. Rev.* **D 23** (1981) 1944 [INSPIRE].

[29] L.M. Jones, *Tests for determining the parton ancestor of a hadron jet*, *Phys. Rev.* **D 39** (1989) 2550 [INSPIRE].

[30] Z. Fodor, *How to see the differences between quark and gluon jets*, *Phys. Rev.* **D 41** (1990) 1726 [INSPIRE].

[31] L. Jones, *Towards a systematic jet classification*, *Phys. Rev.* **D 42** (1990) 811 [INSPIRE].

[32] L. Lönnblad, C. Peterson and T. Rognvaldsson, *Using neural networks to identify jets*, *Nucl. Phys.* **B 349** (1991) 675 [INSPIRE].

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

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

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

[36] B. Bhattacherjee et al., *Associated jet and subjet rates in light-quark and gluon jet discrimination*, *JHEP* **04** (2015) 131 [arXiv:1501.04794] [INSPIRE].

[37] D. Ferreira de Lima, P. Petrov, D. Soper and M. Spannowsky, *Quark-gluon tagging with shower deconstruction: unearthing dark matter and Higgs couplings*, *Phys. Rev.* **D 95** (2017) 034001 [arXiv:1607.06031] [INSPIRE].

[38] B. Bhattacherjee et al., *Quark-gluon discrimination in the search for gluino pair production at the LHC*, *JHEP* **01** (2017) 044 [arXiv:1609.08781] [INSPIRE].

[39] 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].

[40] J. Davighi and P. Harris, *Fractal based observables to probe jet substructure of quarks and gluons*, *Eur. Phys. J.* **C 78** (2018) 334 [arXiv:1703.00914] [INSPIRE].

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

[42] Y. Sakaki, *Quark jet rates and quark/gluon discrimination in multi-jet final states*, arXiv:1807.01421 [INSPIRE].

JHEP11(2018)059

[43] G.P. Salam, *Towards jetography*, *Eur. Phys. J.* **C 67** (2010) 637 [arXiv:0906.1833] [INSPIRE].

[44] A. Banfi, G.P. Salam and G. Zanderighi, *Infrared safe definition of jet flavor*, *Eur. Phys. J.* **C 47** (2006) 113 [hep-ph/0601139] [INSPIRE].

[45] A. Buckley and C. Pollard, *QCD-aware partonic jet clustering for truth-jet flavour labelling*, *Eur. Phys. J.* **C 76** (2016) 71 [arXiv:1507.00508] [INSPIRE].

[46] J. Gallicchio and M.D. Schwartz, *Pure samples of quark and gluon jets at the LHC*, *JHEP* **10** (2011) 103 [arXiv:1104.1175] [INSPIRE].

[47] C. Frye, A.J. Larkoski, M.D. Schwartz and K. Yan, *Precision physics with pile-up insensitive observables*, arXiv:1603.06375 [INSPIRE].

[48] C. Frye, A.J. Larkoski, M.D. Schwartz and K. Yan, *Factorization for groomed jet substructure beyond the next-to-leading logarithm*, *JHEP* **07** (2016) 064 [arXiv:1603.09338] [INSPIRE].

[49] J.R. Andersen et al., *Les Houches 2015: physics at TeV colliders standard model working group report*, arXiv:1605.04692 [INSPIRE].

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

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

[52] 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].

[53] ATLAS collaboration, *Measurement of the charged-particle multiplicity inside jets from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector*, *Eur. Phys. J.* **C 76** (2016) 322 [arXiv:1602.00988] [INSPIRE].

[54] CMS collaboration, *Performance of quark/gluon discrimination in 13 TeV data*, CMS-DP-2016-070 (2016).

[55] ATLAS collaboration, *Quark versus gluon jet tagging using charged particle multiplicity with the ATLAS detector*, ATL-PHYS-PUB-2017-009 (2017).

[56] CMS collaboration, *Measurement of jet substructure observables in $t\bar{t}$ events from proton-proton collisions at $\sqrt{s} = 13$ TeV*, arXiv:1808.07340 [INSPIRE].

[57] J.H. Collins, K. Howe and B. Nachman, *CWoLa hunting: extending the bump hunt with machine learning*, arXiv:1805.02664 [INSPIRE].

[58] J.R. Andersen et al., *Les Houches 2017: physics at TeV colliders standard model working group report*, talk given at the 10$^{\text{th}}$ *Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2017)*, June 5–23, Les Houches, France (2018), arXiv:1803.07977 [INSPIRE].

[59] D. Reichelt, P. Richardson and A. Siodmok, *Improving the simulation of quark and gluon jets with HERWIG 7*, *Eur. Phys. J.* **C 77** (2017) 876 [arXiv:1708.01491] [INSPIRE].

[60] 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].

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

JHEP11(2018)059

[62] J. Neyman and E.S. Pearson, *On the problem of the most efficient tests of statistical hypotheses*, *Phil. Trans. Roy. Soc. London* **A 231** (1933) 289.

[63] G. Blanchard et al., *Classification with asymmetric label noise: Consistency and maximal denoising*, *Electron. J. Stat.* **10** (2016) 2780.

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

[65] 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].

[66] 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].

[67] J. Katz-Samuels, G. Blanchard and C. Scott, *Decontamination of mutual contamination models*, arXiv:1710.01167.

[68] S. Arora, R. Ge and A. Moitra, *Learning topic models — Going beyond SVD*, in the proceedings of the *2012 IEEE* 53rd *Annual Symposium on Foundations of Computer Science (FOCS'12)*, October 20–23, New Brunswick, U.S.A. (2012).

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

[70] CMS collaboration, *Studies of jet mass in dijet and W/Z + jet events*, *JHEP* **05** (2013) 090 [arXiv:1303.4811] [INSPIRE].

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

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

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

[74] 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].

[75] S. Chang, T. Cohen and B. Ostdiek, *What is the machine learning?*, *Phys. Rev.* **D 97** (2018) 056009 [arXiv:1709.10106] [INSPIRE].

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

[77] 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].

[78] 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].

[79] 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].

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

JHEP11(2018)059

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

[82] 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].

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

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

[85] 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].

[86] L. de Oliveira et al., *Jet-images — Deep learning edition*, *JHEP* **07** (2016) 069 [arXiv:1511.05190] [INSPIRE].

[87] P. Baldi et al., *Jet substructure classification in high-energy physics with deep neural networks*, *Phys. Rev.* **D 93** (2016) 094034 [arXiv:1603.09349] [INSPIRE].

[88] D. Guest et al., *Jet flavor classification in high-energy physics with deep neural networks*, *Phys. Rev.* **D 94** (2016) 112002 [arXiv:1607.08633] [INSPIRE].

[89] *EnergyFlow*, https://energyflow.network.

[90] F. Pedregosa et al., *Scikit-learn: machine learning in Python*, *J. Mach. Learning Res.* **12** (2011) 2825.

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

[92] 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].

[93] S. Egan et al., *Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC*, arXiv:1711.09059 [INSPIRE].

[94] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy flow networks: deep sets for particle jets*, arXiv:1810.05165 [INSPIRE].

[95] F. Chollet, *Keras*, https://github.com/fchollet/keras (2017).

[96] M. Abadi et al., *Tensorflow: a system for large-scale machine learning*, *OSDI* **16** (2016) 265.

[97] V. Nair and G.E. Hinton, *Rectified linear units improve restricted Boltzmann machines*, in the proceedings of the 27th *International Conference on Machine learning (ICML-10)*, June 21–24, Haifa, Israel (2010).

[98] K. He, X. Zhang, S. Ren and J. Sun, *Delving deep into rectifiers: surpassing human-level performance on imagenet classification*, in the proceedings *IEEE International Conference on Computer Vision (ICCV2015)*, December 11–18, Santiago, Chile (2015).

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

[100] S. Gao, C.H. Lee and J.H. Lim, *An ensemble classifier learning approach to ROC optimization*, 18th *International Conference on Pattern Recognition (ICPR'06)*, August 20–24, Hong Kong (2006).

JHEP11(2018)059