Editors' Suggestion | Featured in Physics

Jet Topics: Disentangling Quarks and Gluons at Colliders

Eric M. Metodiev* and Jesse Thaler†
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(Received 22 February 2018; published 12 June 2018)

We introduce jet topics: a framework to identify underlying classes of jets from collider data. Because of a close mathematical relationship between distributions of observables in jets and emergent themes in sets of documents, we can apply recent techniques in "topic modeling" to extract jet topics from the data with minimal or no input from simulation or theory. As a proof of concept with parton shower samples, we apply jet topics to determine separate quark and gluon jet distributions for constituent multiplicity. We also determine separate quark and gluon rapidity spectra from a mixed Z-plus-jet sample. While jet topics are defined directly from hadron-level multidifferential cross sections, one can also predict jet topics from first-principles theoretical calculations, with potential implications for how to define quark and gluon jets beyond leading-logarithmic accuracy. These investigations suggest that jet topics will be useful for extracting underlying jet distributions and fractions in a wide range of contexts at the Large Hadron Collider.

DOI: 10.1103/PhysRevLett.120.241602

When quarks and gluons are produced in high-energy particle collisions, their fragmentation and hadronization via quantum chromodynamics (QCD) results in collimated sprays of particles called jets. To extract separate information about quark and gluon jets though, one typically needs to know the relative fractions of quark and gluon jets in the data sample of interest, estimated by convolving matrix element calculations with nonperturbative parton distribution functions (PDFs). Recent progress in jet substructure —the detailed study of particle patterns and correlations within jets [1–10]—has offered new avenues to tag and isolate quark and gluon jets [11–26], with recent applications at the Large Hadron Collider (LHC) [27–35]. Still, there are considerable theoretical uncertainties in the modeling of quark and gluon jets, as well as more fundamental concerns about how to define quark and gluon jets from first principles in QCD [36–39]. In particular, quark and gluon partons carry color charge, while jets are composed of color-singlet hadrons, so there is presently no unambiguous definition of "quark" and "gluon" jet at the hadron level.

In this Letter, we introduce a data-driven technique to extract underlying distributions for different jet types from mixed samples using quark and gluon jets as an example. We call our method "jet topics" because of a mathematical connection to topic modeling, an unsupervised learning paradigm for discovering emergent themes in a corpus of documents [40]. Jet topics are defined directly from measured multidifferential cross sections, requiring no inputs from simulation or theory. In this way, jet topics offer a practical way to define jet classes, allowing us to label quark and gluon jet distributions at the hadron level without reference to the underlying partons.

At colliders like the LHC, it is nearly impossible to kinematically isolate pure samples of different jets (i.e., quark jets, gluon jets, boosted W jets, etc.). Instead, collider data consist of statistical mixtures M_a of K different types of jets. For any jet substructure observable $\mathbf{x}$, such as jet mass, the distribution $p_{M_a}(\mathbf{x})$ in mixed sample M_a is a mixture of the K underlying jet distributions $p_k(\mathbf{x})$:

$$p_{M_a}(\mathbf{x}) = \sum_{k=1}^{K} f_k^{(a)} p_k(\mathbf{x}), \tag{1}$$

where $f_k^{(a)}$ is the fraction of jet type k in sample a, with $\sum_{k=1}^{K} f_k^{(a)} = 1$ for all a and $\int d\mathbf{x} p_k(\mathbf{x}) = 1$ for all k.

For the specific case of quark (q) and gluon (g) jet mixtures, we have

$$p_{M_a}(\mathbf{x}) = f_q^{(a)} p_q(\mathbf{x}) + (1 - f_q^{(a)}) p_g(\mathbf{x}). \tag{2}$$

Of course, there are well-known caveats to this picture of jet generation, which go under the name of "sample dependence." For instance, quark jets from the Z + jet process are not exactly identical to quark jets from the dijet process due to soft color correlations with the entire event [37], though these correlations are power suppressed in the small-jet-radius limit [41–43]. Also, more universal quark-gluon

definitions can be obtained using jet grooming methods [44–52]. Here, we assume that sample-dependent effects can either be quantified or mitigated, taking Eq. (2) as the starting assumption for our analysis.

Mixed quark-gluon samples were previously studied in the context of classification without labels (CWoLa) [53] (see, also, Refs. [54–58]). Via Eq. (2), one can prove that the optimal binary mixed-sample classifier $p_{M_1}(\mathbf{x})/p_{M_2}(\mathbf{x})$ is a monotonic rescaling of the optimal quark-gluon classifier $p_q(\mathbf{x})/p_g(\mathbf{x})$. This means that a classifier trained to optimally distinguish M_1 (e.g., $Z+$jet) from M_2 (e.g., dijets) is optimal for distinguishing quark from gluon jets without requiring jet labels or aggregate class proportions. The CWoLa framework though does not directly yield information about the individual quark and gluon distributions $p_q(\mathbf{x})$ and $p_g(\mathbf{x})$.

With jet topics—and with topic modeling more generally—one can obtain the full distributions $p_k(\mathbf{x})$ and fractions $f_k^{(a)}$ solely from the mixed-sample distributions in Eq. (1), subject to requirements which will be spelled out below. As originally posed, topic modeling aims to expose emergent themes in a collection of text documents (a *corpus*) [40]. A *topic* is a distribution over *words* in the *vocabulary*. *Documents* are taken to be unstructured bags of words. Each document arises from an unknown mixture of topics: a topic is sampled according to the mixture proportions and then a word is chosen according to that topic's distribution over the vocabulary. As long as each topic has words unique to it, known as *anchor words* [59,60], topic-modeling algorithms can learn the underlying topics and proportions from the corpus alone.

Intriguingly, the generative process for producing counts of words in a document is mathematically identical to producing jet observable distributions via Eq. (1), as summarized in Table I. For the case of quark-gluon jet mixtures, we have suggestively depicted the process of writing "jet documents" in Fig. 1. Anchor words are analogous to having phase-space regions where each of the underlying distributions is pure, and the presence of these *anchor bins* is necessary for jet topics to yield the underlying quark and gluon distributions.

TABLE I. The correspondence between topic models and jet distributions. Note that topic modeling treats each document as an unstructured bag of words in the same way that a collection of jets has no intrinsic ordering.

Topic model	Jet distributions
Word	Histogram bin
Vocabulary	Jet observable(s)
Anchor word	Pure phase-space region (*anchor bin*)
Topic	Type of jet (*jet topic*)
Document	Histogram of jet observable(s)
Corpus	Collection of histograms

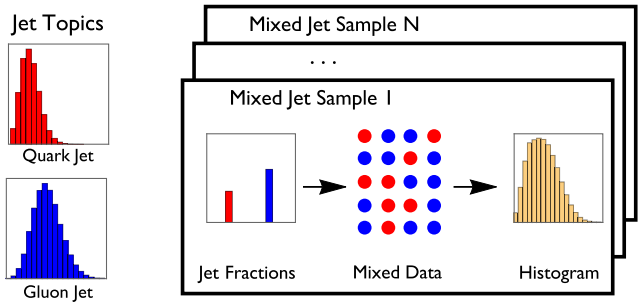

FIG. 1. The generation of mixed samples of quark and gluon jets highlighting the correspondence with topic models. Each jet is either a quark or gluon jet sampled according to the underlying quark fraction. The observable is then sampled according to a universal distribution for that jet type. Each mixed-sample observable distribution is then a mixture of the two universal distributions, giving rise to a jet document.

Because of its theoretical transparency and asymptotic guarantees, we use the `Demix` method [60] to extract jet topics, though other algorithms yield comparable results. The key idea is to undo the mixing of the two fundamental distributions in Eq. (1) by maximally subtracting the two mixtures from one another, such that the zeros of the subtracted distributions correspond to the anchor bins. Adopting the notation of Ref. [60], let $\kappa(M_1|M_2)$ be the largest subtraction amount κ such that $p_{M_1}(\mathbf{x}) - \kappa p_{M_2}(\mathbf{x}) \geq 0$, namely,

$$\kappa(M_i|M_j) = \min_{\mathbf{x}} \frac{p_{M_i}(\mathbf{x})}{p_{M_j}(\mathbf{x})}. \tag{3}$$

We refer to κ as the *reducibility factor* (equivalently, the minimum of the mixed-sample likelihood ratio). The *jet topics* T_1 and T_2 are then the normalized maximal subtractions of M_2 from M_1,

$$p_{T_1}(\mathbf{x}) = \frac{p_{M_1}(\mathbf{x}) - \kappa(M_1|M_2)p_{M_2}(\mathbf{x})}{1-\kappa(M_1|M_2)}, \tag{4}$$

and analogous for $p_{T_2}(\mathbf{x})$. The jet topics are unique and universal in that they are independent of the mixtures used to construct them.

The goal is for the topic distributions $p_{T_1}(\mathbf{x})$ and $p_{T_2}(\mathbf{x})$ to match the underlying quark and gluon jet distributions $p_q(\mathbf{x})$ and $p_g(\mathbf{x})$. There are three required conditions for this to occur. Two of them (shared with CWoLa) are *sample independence* and *different purities*, i.e., that the jet samples are obtained from Eq. (2) with different values of $f_q^{(a)}$. The third condition is the presence of anchor bins, which can be stated more formally as *mutual irreducibility*: Each underlying distribution $p_k(\mathbf{x})$ is not a mixture of the remaining underlying distributions plus another distribution [55].

Note that this is a much weaker requirement than the distributions being fully separated. In the quark-gluon context, a necessary and sufficient condition for mutual irreducibility is that the reducibility factors $\kappa(q|g) = \kappa(g|q) = 0$ for feature representation $\mathbf{x}$. We later explore the implications of this condition for QCD. With these three conditions satisfied, the mixture proportions are uniquely determined via the reducibility factors. Taking $f_q^{(1)} > f_q^{(2)}$, inserting Eq. (2) into Eq. (3) yields

$$\kappa(M_1|M_2) = \frac{1 - f_q^{(1)}}{1 - f_q^{(2)}}, \qquad \kappa(M_2|M_1) = \frac{f_q^{(2)}}{f_q^{(1)}}. \quad (5)$$

Even without mutual irreducibility, the extracted jet topics will still relate to the underlying quark and gluon distributions. Specifically, jet topics yield the "gluon-subtracted quark distribution"

$$p_{q|g}(\mathbf{x}) = \frac{p_q(\mathbf{x}) - \kappa(q|g)p_g(\mathbf{x})}{1 - \kappa(q|g)}, \quad (6)$$

and the "quark-subtracted gluon distribution" defined analogously. By universality, the topics calculated from pure samples via Eq. (6) and from mixtures via Eq. (4) are identical. These may be useful in their own right, particularly if the quark-gluon fractions are uncertain, but $\kappa(q|g)$ and $\kappa(g|q)$ can be determined analytically or from simulation (see Fig. 4 below).

We now turn to a practical demonstration of the jet topics method for realistic quark and gluon samples. Following Ref. [37], we consider two mixed jet processes at the LHC: the quark-enriched Z + jet process and the gluon-enriched dijets process. See Ref. [61] for alternative selections for quark- or gluon-enriched samples. The parton shower PYTHIA 8.226 [62,63] is used to generate 500 000 jets at $\sqrt{s} = 13$ TeV including hadronization and multiple parton interactions (i.e., underlying event). Detector-stable, non-neutrino particles are clustered into anti-k_t jets [64] with radius $R = 0.4$ using FASTJET 3.3.0 [65]. The hardest jet(s) in each event (one jet for Z + jet and up to two jets for dijets) are selected if they have transverse momentum $p_T \in [250, 275]$ GeV and rapidity $|y| \leq 2$. These cuts resulted in the Z + jet process having (PYTHIA-labeled) quark fraction $f_q^{(1)} = 0.88$ and the dijet process having $f_q^{(2)} = 0.37$. We use the constituent multiplicity within a jet as the feature representation $\mathbf{x}$, since it is known to be a good quark-gluon discriminant [18].

In Fig. 2, we present the result of extracting two jet topics from these samples. Shown are the constituent multiplicity distributions from the original Z + jet and dijet samples, from PYTHIA-labeled Z + quark and Z + gluon samples, and from the jet topics T_1 and T_2 using Eq. (4). The uncertainties are estimated by assuming $\pm\sqrt{N}$ bin count uncertainties and only considering bins with more than 30 events. We determine the κ values of Eq. (3)

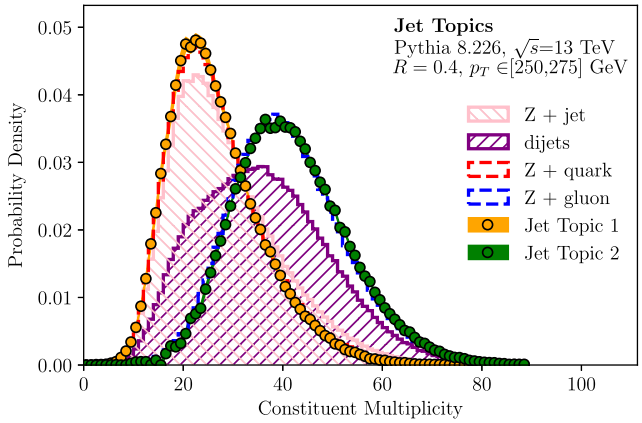

FIG. 2. The jet topics method applied to constituent multiplicity starting with Z + jet (pink) and dijet (purple) distributions from PYTHIA 8.226. There is good agreement between the two extracted jet topics (orange and green) and pure Z + quark and Z + gluon distributions (red and blue).

by selecting the most constraining (anchor) bin: that with the lowest upper uncertainty bar on the ratio. Remarkably, the two extracted jet topics overlap very well with the underlying quark and gluon distributions, providing practical evidence that Eq. (4) works as desired, at least for constituent multiplicity. We verified that similar results could be obtained from samples with different p_T cuts and from mixtures of dijets at different rapidities. This approach is similar to the template extraction procedure in Ref. [24], with the important distinction that the quark-gluon fractions need not be specified *a priori*.

In Fig. 3, we use the extracted jet topics to construct separate jet rapidity spectra for quark and gluon jets in the Z + jet samples. Binning the Z + jet sample into ten rapidity bins in $|y| < 2$, we find the mixture of the two topics extracted above that most closely matches the

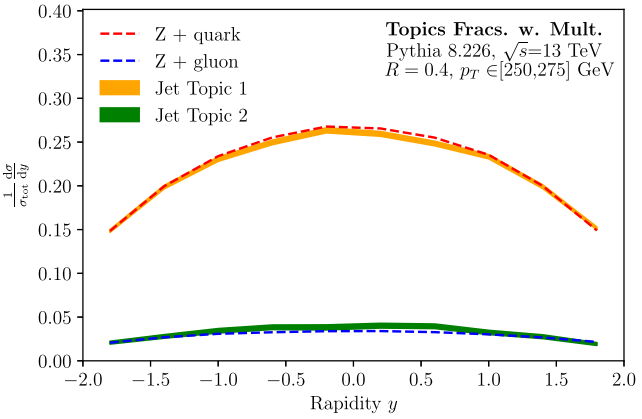

FIG. 3. Cross sections for jet topics (orange and green) using topic fractions extracted from the Z + jet sample across ten rapidity bins. The extracted topic cross sections closely track the underlying Z + quark and Z + gluon cross sections (red and blue).

constituent multiplicity histogram in each rapidity bin, minimizing the squared error to find the best mixture. This is an example of the general problem of extracting sample fractions $f_k^{(a)}$ from various mixed samples. As desired, the extracted topic cross sections in Fig. 3 track the true quark and gluon rapidity cross sections.

Thus, just from a collection of mixed-sample histograms, one can make progress toward extracting both the underlying distributions $p_k(\mathbf{x})$ and the fraction of each jet topic $f_k^{(a)}$. Crucially, Figs. 2 and 3 are just novel projections of the hadron-level multidifferential jet cross section $d^3\sigma/dp_T dy dn_{\text{const}}$ on two independent samples, making jet topics implementable on existing LHC jet measurements (e.g., Ref. [33]). The agreement between the operationally defined jet topics and the theoretically ambiguous quark and gluon distributions may even suggest using mutual irreducibility of the final-state distributions to define quark and gluon jets.

From the perspective of first-principles QCD, the implications of mutual irreducibility are simple yet profound. For the reducibility factors $\kappa(q|g)$ and $\kappa(g|q)$ to be zero, there must be phase-space regions almost entirely dominated by quark or gluon jets. In the leading-logarithmic (LL) limit, mutual irreducibility can be achieved with any jet substructure observable that counts the number of parton emissions, such as "soft drop multiplicity" [26]. In the LL limit, quark and gluon jets have the same emission profile differing only by a color factor in their emission density, $C_F = 4/3$ for quarks and $C_A = 3$ for gluons. Ignoring the Λ_{QCD} regulator, counting these (infinitely many) emissions results in arbitrarily well-separated quark and gluon Poissonian distributions [26] and, therefore, mutual irreducibility. Beyond LL order though, naive quark-gluon definitions may not lead to mutual irreducibility, since running-coupling, higher-order, and nonperturbative effects generically contaminate the anchor bins. That said, as long as these effects maintain sample independence (perhaps achieved via grooming), then one can still use Eq. (6) to define subtracted quark and gluon labels.

Interestingly, many jet substructure observables do not lead to quark-gluon mutual irreducibility, even at LL accuracy. Consider, for instance, the jet mass m (or any jet angularity [66–68]). Jet mass exhibits Casimir scaling at LL order, meaning that the cumulative density functions $\Sigma_i(m)$ are related to each other by $\Sigma_g = \Sigma_q^{C_A/C_F}$ [19,20]. The probability distributions are then given by $p_i = d\Sigma_i/dm$. Substituting this into Eq. (3) immediately yields for all observables with Casimir scaling

$$\kappa(g|q) = \frac{C_A}{C_F} \min \Sigma_q^{\frac{C_A}{C_F}-1} = 0, \tag{7}$$

$$\kappa(q|g) = \frac{C_F}{C_A} \min \Sigma_q^{1-\frac{C_A}{C_F}} = \frac{C_F}{C_A}, \tag{8}$$

since $C_A/C_F = 9/4 > 1$ and Σ takes all values between 0 and 1. Because of Eq. (8), jet mass alone is not sufficient to extract the quark distribution at LL order without additional information.

On the other hand, if the reducibility factors are known, then the subtracted distributions in Eq. (6) can be inverted. This is shown in Fig. 4 for the jet mass, where the quark topic has been corrected using the value $\kappa(q|g) = 0.40$ at 35 GeV determined from the PYTHIA $Z + q/g$ distributions, which is known to differ from the LL expectation [37]. This analysis is performed up to 35 GeV to avoid sample-dependent effects in the high-mass tails of the distributions. The qualitative behavior of the topics agrees with the LL predictions of Eqs. (7) and (8): no correction is needed to obtain the gluon topic, and the quark topic is a nontrivial mixture of the jet topics. Given the good agreement seen here, it would be interesting to apply jet topics to groomed jet mass measurements [69,70], where grooming is an essential ingredient that allows $\kappa(q|g)$ to be calculated to high precision [49–52].

There are many potential uses for the jet topics framework at the LHC. Focusing just on quark and gluon jets, one often wants to separately measure quark and gluon distributions from mixed data samples, without relying on theory or simulation for fraction estimates. To determine PDFs, it would be beneficial to isolate different partonic subprocesses, and this could be feasible as long as jet topics is applied both to data and to fixed-order QCD calculations. Similar subprocess isolation might be useful in monojet searches for dark matter by aiding in signal-background discrimination or in setting improved limits on specific new physics models [71,72]. For extracting the strong coupling constant α_s from (groomed) jet shape distributions, it would be beneficial to determine the quark and gluon jet fractions

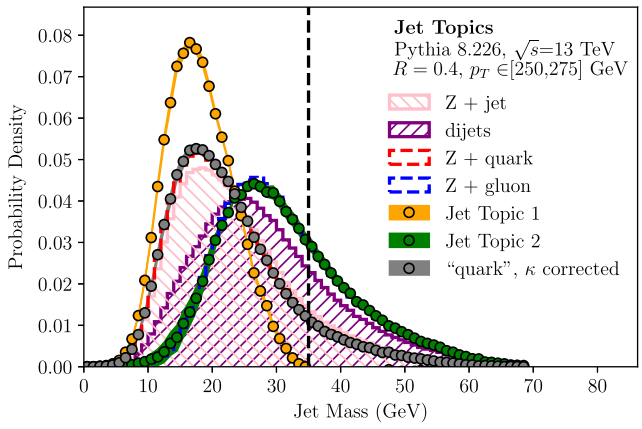

FIG. 4. The jet topics method applied to jet mass up to 35 GeV (up to the black dashed line). The gray curve is the corrected quark topic using PYTHIA to determine $\kappa(q|g)$ extrapolated beyond 35 GeV by letting jet topic 1 go negative. There is good agreement between the κ-corrected quark topic (gray) and the pure $Z +$ quark distribution (red).

using data-driven methods, since there are uncertainties associated with whether α_s comes multiplied by C_F or C_A [73]. The extracted topic fractions could be also be used to augment training with CWoLa, since the classifier operating points could then be determined entirely from the data. In heavy ion collisions, quarks and gluons are expected to be modified differently in medium due to their different color charges, and jet topics may allow for fully data-driven studies of separate quark and gluon jet modifications.

In conclusion, phrasing jet mixtures as a topic-modeling problem makes available a variety of new and more sophisticated statistical and mathematical tools for jet physics (see, e.g., Refs. [59,60,74–88]), including recent efforts to determine the appropriate number of topics to use from the data [89–91]. We emphasize that jet topics can be applied to any set of multidifferential cross sections—in experiment or in theory—as long as the criteria of sample independence, different purities, and mutual irreducibility are met. Furthermore, mutual irreducibility need not be assumed if the subtracted distributions in Eq. (6) are sufficient for the intended application or if the reducibility factors are known from theory or simulation. Of course, experimental studies are needed to understand the systematic and statistical uncertainties associated with jet topics for LHC measurements and searches, and theoretical studies are needed to determine the interplay of jet topics with precision calculations. It would also be interesting to design jet substructure observables specifically targeted for mutual irreducibility. More generally, topic models may find applications in collider physics beyond jets and in other disciplines beyond collider physics, since extracting signal and background distributions from mixtures is a ubiquitous challenge faced when analyzing and interpreting rich data sets.

The authors would like to thank Mario Campanelli, Timothy Cohen, Philip Harris, Patrick Komiske, Andrew Larkoski, Ian Moult, Benjamin Nachman, Gavin Salam, Clayton Scott, and Wouter Waalewijn for illuminating discussions. The authors are grateful to Patrick Komiske for generating the jet samples. This work was supported by the Office of Nuclear Physics of the U.S. Department of Energy (DOE) under Grant No. DE-SC0011090 and the DOE Office of High Energy Physics under Grant No. DE-SC0012567. Computations in this paper were run on the Odyssey cluster supported by the Faculty of Arts and Sciences Division of Science, Research Computing Group at Harvard University. Cloud computing resources were provided through a Microsoft Azure for Research grant.

*metodiev@mit.edu
†jthaler@mit.edu

[1] M. H. Seymour, Tagging a heavy Higgs boson, in *Proceedings of the ECFA Large Hadron Collider Workshop, Aachen, Germany* (CERN, Geneva, 1990), pp. 557–569.
[2] M. H. Seymour, Searches for new particles using cone and cluster jet algorithms: A comparative study, Z. Phys. C **62**, 127 (1994).
[3] J. M. Butterworth, B. E. Cox, and J. R. Forshaw, *WW* scattering at the CERN LHC, Phys. Rev. D **65**, 096014 (2002).
[4] J. M. Butterworth, J. R. Ellis, and A. R. Raklev, Reconstructing sparticle mass spectra using hadronic decays, J. High Energy Phys. 05 (2007) 033.
[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**, 242001 (2008).
[6] A. Abdesselam *et al.*, Boosted objects: A probe of beyond the Standard Model physics, Eur. Phys. J. C **71**, 1661 (2011).
[7] A. Altheimer *et al.*, Jet substructure at the Tevatron and LHC: New results, new tools, new benchmarks, J. Phys. G **39**, 063001 (2012).
[8] A. Altheimer *et al.*, Boosted objects and jet substructure at the LHC, Eur. Phys. J. C **74**, 2792 (2014).
[9] D. Adams *et al.*, Towards an understanding of the correlations in jet substructure, Eur. Phys. J. C **75**, 409 (2015).
[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.
[11] H. P. Nilles and K. H. Streng, Quark-gluon separation in three jet events, Phys. Rev. D **23**, 1944 (1981).
[12] L. M. Jones, Tests for determining the parton ancestor of a hadron jet, Phys. Rev. D **39**, 2550 (1989).
[13] Z. Fodor, How to see the differences between quark and gluon jets, Phys. Rev. D **41**, 1726 (1990).
[14] L. Jones, Towards a systematic jet classification, Phys. Rev. D **42**, 811 (1990).
[15] L. Lönnblad, C. Peterson, and T. Rognvaldsson, Using neural networks to identify jets, Nucl. Phys. **B349**, 675 (1991).
[16] J. Pumplin, How to tell quark jets from gluon jets, Phys. Rev. D **44**, 2025 (1991).
[17] J. Gallicchio and M. D. Schwartz, Quark and Gluon Tagging at the LHC, Phys. Rev. Lett. **107**, 172001 (2011).
[18] J. Gallicchio and M. D. Schwartz, Quark and gluon jet substructure, J. High Energy Phys. 04 (2013) 090.
[19] A. J. Larkoski, G. P. Salam, and J. Thaler, Energy correlation functions for jet substructure, J. High Energy Phys. 06 (2013) 108.
[20] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, Gaining (mutual) information about quark/gluon discrimination, J. High Energy Phys. 11 (2014) 129.
[21] B. Bhattacherjee, S. Mukhopadhyay, M. M. Nojiri, Y. Sakaki, and B. R. Webber, Associated jet and subjet rates in light-quark and gluon jet discrimination, J. High Energy Phys. 04 (2015) 131.
[22] D. F. 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**, 034001 (2017).

[23] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, Deep learning in color: Towards automated quark-gluon jet discrimination, J. High Energy Phys. 01 (2017) 110.

[24] ATLAS Collaboration, Discrimination of light quark and gluon jets in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector, 2016.

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

[26] C. Frye, A. J. Larkoski, J. Thaler, and K. Zhou, Casimir meets Poisson: Improved quark/gluon discrimination with counting observables, J. High Energy Phys. 09 (2017) 083.

[27] CMS Collaboration, Performance of quark/gluon discrimination in 8 TeV pp data, Technical Report No. CMS-PAS-JME-13-002, 2013.

[28] G. Aad *et al.* (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**, 3023 (2014).

[29] G. Aad *et al.* (ATLAS Collaboration), Jet energy measurement and its systematic uncertainty in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector, Eur. Phys. J. C **75**, 17 (2015).

[30] V. Khachatryan *et al.* (CMS Collaboration), Measurement of electroweak production of two jets in association with a Z boson in proton-proton collisions at $\sqrt{s} = 8$ TeV, Eur. Phys. J. C **75**, 66 (2015).

[31] G. Aad *et al.* (ATLAS Collaboration), Search for high-mass diboson resonances with boson-tagged jets in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector, J. High Energy Phys. 12 (2015) 055.

[32] V. Khachatryan *et al.* (CMS Collaboration), Search for the standard model Higgs boson produced through vector boson fusion and decaying to $b\bar{b}$, Phys. Rev. D **92**, 032008 (2015).

[33] G. Aad *et al.* (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**, 322 (2016).

[34] CMS Collaboration, Performance of quark/gluon discrimination in 13 TeV data, Technical Report No. CMS-DP-2016-070, 2016.

[35] M. Aaboud *et al.* (ATLAS Collaboration), Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Rev. D **96**, 072002 (2017).

[36] A. Banfi, G. P. Salam, and G. Zanderighi, Infrared safe definition of jet flavor, Eur. Phys. J. C **47**, 113 (2006).

[37] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer, A. Siódmok, P. Skands, G. Soyez, and J. Thaler, Systematics of quark/gluon tagging, J. High Energy Phys. 07 (2017) 091.

[38] D. Reichelt, P. Richardson, and A. Siodmok, Improving the simulation of quark and gluon jets with Herwig 7, Eur. Phys. J. C **77**, 876 (2017).

[39] J. Mo, F. J. Tackmann, and W. J. Waalewijn, A case study of quark-gluon discrimination at NNLL in comparison to parton showers, Eur. Phys. J. C **77**, 770 (2017).

[40] D. M. Blei, Probabilistic topic models, Commun. ACM **55**, 77 (2012).

[41] A. Banfi, M. Dasgupta, K. Khelifa-Kerfa, and S. Marzani, Non-global logarithms and jet algorithms in high-p_T jet shapes, J. High Energy Phys. 08 (2010) 064.

[42] T. Becher, M. Neubert, L. Rothen, and D. Y. Shao, Factorization and resummation for jet processes, J. High Energy Phys. 11 (2016) 019; Erratum, J. High Energy Phys. 05 (2017) 154.

[43] D. W. Kolodrubetz, P. Pietrulewicz, I. W. Stewart, F. J. Tackmann, and W. J. Waalewijn, Factorization for jet radius logarithms in jet mass spectra at the LHC, J. High Energy Phys. 12 (2016) 054.

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

[45] 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**, 094023 (2010).

[46] D. Krohn, J. Thaler, and L.-T. Wang, Jet trimming, J. High Energy Phys. 02 (2010) 084.

[47] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, Towards an understanding of jet substructure, J. High Energy Phys. 09 (2013) 029.

[48] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft drop, J. High Energy Phys. 05 (2014) 146.

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

[50] C. Frye, A. J. Larkoski, M. D. Schwartz, and K. Yan, Factorization for groomed jet substructure beyond the next-to-leading logarithm, J. High Energy Phys. 07 (2016) 064.

[51] S. Marzani, L. Schunk, and G. Soyez, A study of jet mass distributions with grooming, J. High Energy Phys. 07 (2017) 132.

[52] S. Marzani, L. Schunk, and G. Soyez, The jet mass distribution after soft drop, Eur. Phys. J. C **78**, 96 (2018).

[53] E. M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.

[54] K. Cranmer, J. Pavez, and G. Louppe, Approximating likelihood ratios with calibrated discriminative classifiers, arXiv:1506.02169.

[55] G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott, Classification with asymmetric label noise: Consistency and maximal denoising, Electron. J. Stat. **10**, 2780 (2016).

[56] L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman, Weakly supervised classification in high energy physics, J. High Energy Phys. 05 (2017) 145.

[57] T. Cohen, M. Freytsis, and B. Ostdiek, (Machine) learning to do more with less, J. High Energy Phys. 02 (2018) 034.

[58] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, Learning to classify from impure samples, arXiv:1801.10158.

[59] S. Arora, R. Ge, and A. Moitra, Learning topic models–Going beyond SVD, in *Proceedings of the IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, New York, 2012), pp. 1–10, DOI: 10.1109/FOCS.2012.49.

[60] J. Katz-Samuels, G. Blanchard, and C. Scott, Decontamination of mutual contamination models, arXiv:1710.01167.
[61] J. Gallicchio and M. D. Schwartz, Pure samples of quark and gluon jets at the LHC, J. High Energy Phys. 10 (2011) 103.
[62] T. Sjostrand, S. Mrenna, and P. Z. Skands, PYTHIA 6.4 physics and manual, J. High Energy Phys. 05 (2006) 026.
[63] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015).
[64] M. Cacciari, G. P. Salam, and G. Soyez, The anti-$k(t)$ jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[65] M. Cacciari, G. P. Salam, and G. Soyez, FASTJET user manual, Eur. Phys. J. C **72**, 1896 (2012).
[66] C. F. Berger, T. Kucs, and G. F. Sterman, Event shape/energy flow correlations, Phys. Rev. D **68**, 014012 (2003).
[67] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung, and J. Virzi, Substructure of high-p_T jets at the LHC, Phys. Rev. D **79**, 074017 (2009).
[68] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, and C. Lee, Jet shapes and jet algorithms in SCET, J. High Energy Phys. 11 (2010) 101.
[69] M. Aaboud *et al.* (ATLAS Collaboration), A measurement of the soft-drop jet mass in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, arXiv:1711.08341.
[70] CMS Collaboration, Measurement of the differential jet production cross section with respect to jet mass and transverse momentum in dijet events from pp collisions at $\sqrt{s} = 13$ TeV, Technical Report No. CMS-PAS-SMP-16-010, CERN, Geneva, 2017.
[71] M. Aaboud *et al.* (ATLAS Collaboration), Search for dark matter and other new phenomena in events with an energetic jet and large missing transverse momentum using the ATLAS detector, J. High Energy Phys. 01 (2018) 126
[72] A. M. Sirunyan *et al.* (CMS Collaboration), Search for new physics in final states with an energetic jet or a hadronically decaying W or Z boson and transverse momentum imbalance at $\sqrt{s} = 13$ TeV, arXiv:1712.02345.
[73] S. Chatterjee, F. Dreyer, M. V. Garzelli, P. Gras, A. Larkoski, S. Marzani, I. Moult, B. Nachman, A. Siodmok, A. Papaefstathiou, P. Richardson, T. Samui, G. Soyez, and J. Thaler, Towards extracting the strong coupling constant from jet substructure at the LHC in Les Houches 2017: Physics at TeV colliders standard model working group report, in *10th Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2017) Les Houches, France, 2017* (2018), http://lss.fnal.gov/archive/2018/conf/fermilab-conf-18-122-cd-t.pdf.
[74] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. **3**, 993 (2003).
[75] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature (London) **401**, 788 (1999).
[76] D. Donoho and V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts?, *Advances in Neural Information Processing Systems 16 (NIPS 2003)* (2004), p. 1141.
[77] G. R. Naik and D. K. Kumar, An overview of independent component analysis and its applications, Informatica **35**, 63 (2011).
[78] G. Blanchard and C. Scott, Decontamination of mutually contaminated models, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 33, edited by S. Kaski and J. Corander (PMLR, Reykjavik, 2014), pp. 1–9, http://proceedings.mlr.press/v33/blanchard14.
[79] S. Jain, M. White, M. W. Trosset, and P. Radivojac, Nonparametric semi-supervised learning of class proportions, arXiv:1601.01944.
[80] J. Katz-Samuels and C. Scott, A mutual contamination analysis of mixed membership and partial label models, arXiv:1602.06235.
[81] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications* (Academic Press, New York, 2010).
[82] V. G. Reju, S. N. Koh, and I. Y. Soon, An algorithm for mixing matrix estimation in instantaneous blind source separation, Signal Processing **89**, 1762 (2009).
[83] T. Sanderson and C. Scott, Class proportion estimation with application to multiclass anomaly rejection, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 33, edited by S. Kaski and J. Corander (PMLR, Reykjavik, 2014), pp. 850–858, http://proceedings.mlr.press/v33/sanderson14.
[84] C. Scott, A rate of convergence for mixture proportion estimation, with application to learning from noisy labels, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 38, edited by G. Lebanon and S. V. N. Vishwanathan (PMLR, San Diego, 2015), pp. 838–846, http://proceedings.mlr.press/v38/scott15.html.
[85] H. Ramaswamy, C. Scott, and A. Tewari, Mixture proportion estimation via kernel embeddings of distributions, in *Proceedings of The 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 48, edited by M. F. Balcan and K. Q. Weinberger (PMLR, New York, 2016), pp. 2052–2060, http://proceedings.mlr.press/v48/ramaswamy16.html.
[86] W. Ding, P. Ishwar, and V. Saligrama, Necessary and sufficient conditions and a provably efficient algorithm for separable topic discovery, arXiv:1508.05565.
[87] W. Ding, Learning mixed membership models with a separable latent structure: Theory, provably efficient algorithms, and applications, Ph.D. thesis, Boston University, 2015.
[88] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, A practical algorithm for topic modeling with provable guarantees, in *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, edited by S. Dasgupta and D. McAllester (PMLR, Atlanta, 2013), pp. 280–288, http://proceedings.mlr.press/v28/arora13.html.
[89] D. Greene, D. O'Callaghan, and P. Cunningham, How many topics? Stability analysis for topic models, in *Proceedings of the Joint European Conference on Machine Learning and*

Knowledge Discovery in Databases (Springer, New York, 2014), pp. 498–51q3.

[90] B. Wang, Y. Liu, Z. Liu, M. Li, and M. Qi, Topic selection in latent Dirichlet allocation, in *Proceedings of the 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2014* (IEEE, New York, 2014), pp. 756–760, DOI: 10.1109/FSKD.2014.6980931.

[91] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou, A heuristic approach to determine an appropriate number of topics in topic modeling, BMC Bioinf. **16**, S8 (2015).