Disentangling quarks and gluons in CMS open data

Patrick T. Komiske,[*] Serhii Kryhin,[†] and Jesse Thaler[‡]
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
and The NSF AI Institute for Artificial Intelligence and Fundamental Interactions

(Received 25 May 2022; accepted 10 October 2022; published 16 November 2022)

We study quark and gluon jets separately using public collider data from the CMS experiment. Our analysis is based on 2.3 fb^{-1} of proton-proton collisions at $\sqrt{s} = 7$ TeV, collected at the Large Hadron Collider in 2011. We define two nonoverlapping samples via a pseudorapidity cut—central jets with $|\eta| \leq 0.65$ and forward jets with $|\eta| > 0.65$—and employ jet topic modeling to extract individual distributions for the maximally separable categories. Under certain assumptions, such as sample independence and mutual irreducibility, these categories correspond to "quark" and "gluon" jets, as given by a recently proposed operational definition. We consider a number of different methods for extracting reducibility factors from the central and forward datasets, from which the fractions of quark jets in each sample can be determined. The greatest stability and robustness to statistical uncertainties is achieved by a novel method based on parametrizing the end points of a receiver operating characteristic curve. To mitigate detector effects, which would otherwise induce unphysical differences between central and forward jets, we use the OMNIFOLD method to perform central value unfolding. As a demonstration of the power of this method, we extract the intrinsic dimensionality of the quark and gluon jet samples, which exhibit Casimir scaling, as expected from the strongly ordered limit. To our knowledge, this work is the first application of full phase space unfolding to real collider data (albeit without a full systematics analysis), and the first application of topic modeling using a machine-learned classifier to extract separate quark and gluon distributions at the LHC.

DOI: 10.1103/PhysRevD.106.094021

I. INTRODUCTION

Despite being fundamental ingredients in quantum chromodynamics (QCD), quarks and gluons are ambiguous categories experimentally. While one can cluster sprays of hadrons into jets and use them as proxies for short-distance partons, one cannot definitively say whether a jet was initiated by a quark or gluon due to color confinement. Nevertheless, one can use an *operational definition* of "quarks" and "gluons" to extract interesting jet information from experimental data, as advocated in Ref. [1]. This approach, which is based on the statistical framework of *jet topic modeling* [2], has been applied by the ATLAS experiment to measured jet samples [3] and explored in phenomenological studies [4–7]. See Refs. [8–10] for related topic modeling studies in particle physics.

Because this quark/gluon definition is operational, it depends on the detailed choice of topic modeling algorithm. This is especially true when dealing with finite data samples and their associated statistical uncertainties. The approach described in Ref. [2] was based on histogramming observables and therefore sensitive to the choice of binning, as studied in Ref. [4]. This in turn impacts the uncertainties of the results, since the statistical noise scales roughly with the inverse square root of the number of events in the most constraining bin. To deal with low-statistics datasets, Ref. [5] introduced a more robust method based on functional fitting, though this approach is sensitive to the choice of priors. In addition to these statistical issues, one always has to confront systematic uncertainties in any experimental context.

In this paper, we develop two new approaches to topic modeling that exhibit more stable behavior and apply them to jets extracted from the 2011 CMS open data [11–13]. Jets are divided by pseudorapidity into forward and central datasets, which can be disentangled into "quark" and "gluon" categories using various topic modeling procedures. We compare the quark and gluon fractions obtained from different jet observables, including those based on machine learning [14], and emphasize the relative statistical uncertainties of the different approaches. Using our most

[*]pkomiske@mit.edu
[†]serhin@mit.edu
[‡]jthaler@mit.edu

robust topic modeling method—based on analyzing the end points of the receiver operator characteristic (ROC) curve of a machine-learned classifier—we then present various quark and gluon observables, including their intrinsic dimensionality [15–18]. As a cross-check, we compare results from the CMS open data to those obtained from the parton shower generator PYTHIA 6.4.25 [19]. This PYTHIA 6 sample was provided as part of the CMS open data release.

Because the jet substructure response of the CMS detector [20] depends on the jet pseudorapidity, we find that detector effects have a big impact on topic modeling. Specifically, a necessary condition for topic modeling is *sample independence*, which means that the underlying quark and gluon distributions must be the same in each mixed jet sample. Though a complete study of experimental systematics is beyond the scope of this work, we do implement *central value unfolding* using OMNIFOLD [21,22], which mitigates these detector effects. Weights from this unfolding are publicly available at Ref. [23]. The results of topic modeling on unfolded particle-level CMS open data are comparable to those from PYTHIA, after accounting for differences in the quark/gluon fractions.

Our two new topic modeling algorithms show better robustness to statistical noise and training imperfections than the *anchor bin* method used in Refs. [1–4]. Anchor bins correspond to maximally pure regions of phase space with very high background rejection but correspondingly low signal efficiency. These anchor bins therefore exhibit large statistical uncertainties due to low count rates. To mitigate these uncertainties, we first introduce a *log-likelihood ratio fit* procedure based on probability distribution functions and quantile binning. We then show how to sidestep binning through a *ROC curve fit* defined by cumulative distribution functions. Using a sample of 2×10^6 jets, we find consistent results from all three methods, but better stability from the two fitting procedures.

The remainder of this paper is organized as follows. In Sec. II, we discuss the motivation behind our new topic modeling methods from a statistical perspective and introduce the machine learning models we use. We discuss the features of the CMS open data dataset in Sec. III, and we justify the application of central value unfolding to mitigate sample dependence. In Sec. IV, we apply our topic modeling procedures to the CMS open data dataset and compare the extracted reducibility factors. We then present separate quark and gluon distributions in Sec. V, including ROC curves, rapidity spectra, and the intrinsic dimensionality of quark and gluon jets, which are our main results. We conclude in Sec. VI with a discussion of possible extensions of these methods.

II. JET TOPIC MODELING

In this section, we review the framework of jet topic modeling as well as the anchor bin method used in previous studies. We then introduce our log-likelihood ratio fit (L-fit) and ROC curve fit (R-fit) methods, which exhibit better stability. We then discuss machine-learned classification and regression as inputs to the topic modeling procedure.

A. Review of mixtures and reducibility factors

The following discussion uses the notation and terminology from the original jet topics paper [2], which is based on the DEMIX algorithm of Ref. [24]. Assume you have two mixed samples of jets, denoted as M_1 and M_2, which can be decomposed into their quark (q) and gluon (g) components. This means that for some jet observable $\mathcal{O}$, the observable distributions for the two mixtures can be written as

$$p_{M_1}(\mathcal{O}) = f_1 p_q(\mathcal{O}) + (1 - f_1) p_g(\mathcal{O}), \tag{1}$$

$$p_{M_2}(\mathcal{O}) = f_2 p_q(\mathcal{O}) + (1 - f_2) p_g(\mathcal{O}), \tag{2}$$

where $p_q(\mathcal{O})$ and $p_g(\mathcal{O})$ are the quark and gluon distributions for $\mathcal{O}$, and $f_1 \in [0, 1]$ and $f_2 \in [0, 1]$ are the quark fractions in M_1 and M_2, respectively. We further assume that the mixtures are different and take $f_1 > f_2$, such that M_1 is more quark enriched than M_2.

The above equations are valid under the assumption of *sample independence*. This means that the observable distributions for quark and gluon jets are the same in the phase space regions defined by M_1 and M_2, such that the two mixtures differ only by their quark fractions. In principle, soft color correlations and nonperturbative effects can introduce sample dependence for quarks and gluons, though the analysis in Ref. [1] found this to be a small effect (see also Refs. [25,26]). As we will see in Sec. III C, detector effects will introduce large sample dependence, which we must mitigate for our analysis.

Making the further assumption that quarks and gluons are *mutually irreducible* in $\mathcal{O}$ [see Eq. (8) below], we can invert Eqs. (1) and (2) and solve for the quark and gluon distributions. Defining the likelihood ratio

$$L_{A/B}(\mathcal{O}) = (L_{B/A}(\mathcal{O}))^{-1} = \frac{p_A(\mathcal{O})}{p_B(\mathcal{O})}, \tag{3}$$

the *reducibility factors* are the maximum amount of one mixture that can be subtracted from the other and still leave a valid probability density:

$$\kappa_{21} \equiv \min_{\mathcal{O}} L_{M_2/M_1}(\mathcal{O}), \qquad \kappa_{12} \equiv \min_{\mathcal{O}} L_{M_1/M_2}(\mathcal{O}), \tag{4}$$

In particular, the quark and gluon distributions are given by

$$p_q(\mathcal{O}) = \frac{p_{M_1}(\mathcal{O}) - \kappa_{12} p_{M_2}(\mathcal{O})}{1 - \kappa_{12}}, \tag{5}$$

$$p_g(\mathcal{O}) = \frac{p_{M_2}(\mathcal{O}) - \kappa_{21} p_{M_1}(\mathcal{O})}{1 - \kappa_{21}}, \tag{6}$$

which are non-negative distributions by construction.

We emphasize that the mutual irreducibility of quarks and gluons is an assumption, but a powerful one since it allows us to operationally define what we mean by "quarks" and "gluons" [1]. To state the assumption more explicitly, we can define the reducibility factors for the quark and gluon distributions themselves as

$$\kappa_{gq} \equiv \min_{\mathcal{O}}(L_{g/q}(\mathcal{O})), \qquad \kappa_{qg} \equiv \min_{\mathcal{O}}(L_{q/g}(\mathcal{O})). \quad (7)$$

With this notation, mutual irreducibility corresponds to the assumption that

$$\kappa_{gq} = \kappa_{qg} = 0. \quad (8)$$

In various limits of quantum chromodynamics, one can prove that this relation holds [2,6,27], which is why we are motivated to use it for our studies, despite its potential limitations [28].

A useful consequence of Eq. (8) is that we can derive the quark fractions for M_1 and M_2 by plugging Eqs. (5) and (6) into Eqs. (1) and (2), yielding

$$f_1 = \frac{1-\kappa_{12}}{1-\kappa_{12}\kappa_{21}}, \qquad f_2 = \frac{\kappa_{21}(1-\kappa_{12})}{1-\kappa_{12}\kappa_{21}}. \quad (9)$$

Via Eq. (9), we can extract the quark fractions in a given dataset only knowing κ_{21} and κ_{12}. Via Eqs. (5) and (6), we can also extract the distributions of quark and gluon jets for any observable, even one that was not used to extract the reducibility factors in the first place. In this way, jet topic modeling boils down to determining the reducibility factors in Eq. (4), and we present three different strategies for their extraction in the next three subsections.

B. Anchor bin method

The anchor bin method is the most direct interpretation of Eq. (4) and was used in Refs. [1–4]. If one only has access to finite samples drawn from the probability densities $p_{M_1}(\mathcal{O})$ and $p_{M_2}(\mathcal{O})$, then one needs some way to estimate the likelihood ratio. One way to do so is to bin the data into histograms, yielding the probabilities $p_{M_1}(\mathcal{O}_i)$ and $p_{M_2}(\mathcal{O}_i)$, where i is the bin number. The anchor bin method therefore depends on the choice of binning, as discussed further in Sec. IV B.

Because it is more convenient to work with log-likelihood ratios, we define the *anchor values* as

$$a_{21} \equiv \min_i \ln L_{M_2/M_1}(\mathcal{O}_i), \qquad a_{12} \equiv \max_i \ln L_{M_2/M_1}(\mathcal{O}_i). \quad (10)$$

We refer to the corresponding arguments of the minimum/maximum as the *anchor bins*. The reducibility factors are then given by

$$\kappa_{21} = \exp[a_{21}], \qquad \kappa_{12} = \exp[-a_{12}]. \quad (11)$$

In practice, Eq. (10) is modified to account for statistical uncertainties as described in Sec. IV B. With the reducibility factors in hand, we can use Eq. (9) to extract the quark fractions and Eqs. (5) and (6) to extract the quark and gluon distributions for any observable.

The main advantage of the anchor bin method is its conceptual simplicity, since it is just a binned version of Eq. (4). The main drawback is that it depends sensitively on the end points of the probability densities, where bin counts can be low. The next two methods aim to mitigate this issue by taking into account information from the whole distribution.

C. Log-likelihood ratio fit method

The log-likelihood ratio fit (L-fit) method is a more statistically robust way to extract reducibility factors. Like the anchor bin method, it depends on the choice of histogram binning. Unlike the anchor bin method, it takes advantage of data from the whole observable domain, not only from the neighborhoods around the maximum and minimum values.

After estimating the log-likelihood ratio $\ln L_{M_2/M_1}$, we fit it to a polynomial of degree K:

$$f(\mathcal{O}; b) = \sum_{k=0}^{K} b_k f_k(\mathcal{O}), \quad (12)$$

where b_k are fit coefficients, and $f_k(\mathcal{O})$ is a polynomial of degree k from some polynomial basis. To avoid sensitivity to possible changes of variables, we find it convenient in Sec. IV C to characterize $\mathcal{O}$ in terms of its quantiles (strictly speaking, fractiles), such that the argument of f ranges from 0 to 1. Quantile binning makes it such that different observables have similar shapes to their log-likelihood ratios, simplifying the fitting procedure.

Assuming that the function $f(\mathcal{O}; b)$ faithfully represents $L_{M_2/M_1}(\mathcal{O})$ when $b = b^{\text{opt}}$, the anchor values can be extracted via

$$a_{21} = \min_{\mathcal{O}} f(\mathcal{O}; b^{\text{opt}}), \qquad a_{12} = \max_{\mathcal{O}} f(\mathcal{O}; b^{\text{opt}}), \quad (13)$$

and the reducibility factors are given by Eq. (11) as before. In the special case that the quark/gluon likelihood ratio is a monotonic function of the observable $\mathcal{O}$, with smaller values corresponding to more quarklike, Eq. (13) simplifies to

$$a_{21} = f(\mathcal{O}_{\text{min}}; b^{\text{opt}}), \qquad a_{12} = f(\mathcal{O}_{\text{max}}; b^{\text{opt}}), \quad (14)$$

where $\mathcal{O}_{\text{min}} = \min(\mathcal{O})$ and $\mathcal{O}_{\text{max}} = \max(\mathcal{O})$. When characterizing $\mathcal{O}$ in terms of its quantiles, $\mathcal{O}^{\text{quant}}_{\text{min}} = 0$ and $\mathcal{O}^{\text{quant}}_{\text{max}} = 1$.

The main advantage of the L-fit method is that it allows for a better accounting of statistical noise and is less dependent on binning than the anchor bin method. The main drawback is that it has residual binning dependence, especially if one does not use the quantile parametrization of the observable. It also depends on the precise choice of functional fit form, though this dependence can be assessed as an uncertainty in the method.

D. ROC curve fit method

The ROC curve fit (R-fit) method allows us to extract the reducibility factors without needing to bin the observable. The method assumes that the quark/gluon likelihood ratio is a monotonic function of the observable, which will be the case for the observables studied in this paper.

The R-fit method starts from the cumulative probabilities $P(\mathcal{O}_{\text{cut}})$, defined as

$$P(\mathcal{O}_{\text{cut}}) = \int_{\mathcal{O}_{\text{min}}}^{\mathcal{O}_{\text{cut}}} d\mathcal{O} p(\mathcal{O}), \tag{15}$$

where $\mathcal{O}_{\text{min}}$ is the minimum value of the observable. For the mixtures M_1 and M_2, the ROC curve is the parametric curve traced out by the value of the upper cut $\mathcal{O}_{\text{cut}}$:

$$\{P_{M_1}(\mathcal{O}_{\text{cut}}), P_{M_2}(\mathcal{O}_{\text{cut}})\}. \tag{16}$$

Assuming that the ROC curve is single valued, it tells us the efficiency for selecting the "background" (M_2, more gluonlike) as a function of the efficiency for selecting the "signal" (M_1, more quarklike):

$$P_{M_2}(P_{M_1}). \tag{17}$$

In the statistics language, this is the false positive rate as a function of the true positive rate.[1]

Since the cumulative probabilities $P_{M_1}(\mathcal{O}_{\text{cut}})$ and $P_{M_2}(\mathcal{O}_{\text{cut}})$ contain the same information as the probability densities, we can use them to extract the reducibility factors κ_{21} and κ_{12}. Perhaps less obvious is that the ROC curve built from P_{M_1} and P_{M_2} contains sufficient information for this extraction. Assuming the quark/gluon likelihood ratio is a monotonic function of the observable, then as discussed around Eq. (14), the anchor bins are located at the end points of the observable. One end point has $P_{M_1} \to 0$ and $P_{M_2} \to 0$, and the slope of the ROC curve is

$$\begin{aligned}\lim_{P_{M_1}\to 0} \frac{dP_{M_2}}{dP_{M_1}} &= \lim_{\mathcal{O}_{\text{cut}}\to\mathcal{O}_{\text{min}}} \frac{dP_{M_2}}{d\mathcal{O}_{\text{cut}}} \frac{d\mathcal{O}_{\text{cut}}}{dP_{M_1}}, \\ &= \frac{p_{M_2}(\mathcal{O}_{\text{min}})}{p_{M_1}(\mathcal{O}_{\text{min}})} = \kappa_{21},\end{aligned} \tag{18}$$

where we have used the fundamental theorem of calculus in the second line. Similarly, the other end point has $P_{M_1} \to 1$ and $P_{M_2} \to 1$, and the slope of the ROC curve is

$$\lim_{P_{M_1}\to 1} \frac{d(1-P_{M_2})}{d(1-P_{M_1})} = \frac{p_{M_2}(\mathcal{O}_{\text{max}})}{p_{M_1}(\mathcal{O}_{\text{max}})} = \frac{1}{\kappa_{12}}. \tag{19}$$

We emphasize that these relations assume the monotonicity of the (log-)likelihood ratio as a function of the observable. In this way, the slopes of the ROC curve at the end points directly yield the reducibility factors.

With finite statistics datasets, we can fit the ROC curve to a polynomial and extract the slopes from the obtained functional fit. To ensure the fit intersects the points $\{0, 0\}$ and $\{1, 1\}$, we use the form

$$f(r) = r + r(1-r)\left(\sum_{k=0}^{K} b_k f_k(r)\right), \tag{20}$$

where b_k are the fit coefficients and $f_k(x)$ are polynomials of degree k defined on the interval [0, 1].

Like the L-fit method, the R-fit method is robust to statistical noise since it uses information from the whole observable range. That said, the procedure for accounting for statistical uncertainties with the R-fit method is more involved, as discussed in Sec. IV D. The main caution when using this method is that it relies on the quark/gluon likelihood ratio being a monotonic function of the observable. Fortunately, this is a quite common property (for example, constituent multiplicity and mass), and this also turns out to be the case for the machine-learned classifiers that we use for our default reducibility factor extraction.

E. Machine-learned classification

Jet topic modeling can be applied to any observable, but as advocated in Ref. [1], it is particularly powerful when applied to a machine-learned classifier. As reviewed below, such classifiers are monotonically related to the likelihood ratio $L_{g/q}(\mathcal{O})$ in the asymptotic limit, and therefore well suited for extracting reducibility factors.

Using the classification without labels technique [29], we can train a classifier to distinguish M_1 and M_2, and this yields a function that defines optimal decision boundaries between quark and gluon jets. For concreteness, we use the binary cross entropy loss functional to train a classifier $c(x)$ over the full phase space x:

[1]This definition of the ROC curve is standard in the HEP literature. In the statistics literature, it is more common to plot the true positive rate as a function of the false positive rate.

$$L[c] = \int dx (p(M_1) p_{M_1}(x) \ln c(x) + p(M_2) p_{M_2}(x) \ln[1 - c(x)]), \tag{21}$$

where $p(M_a)$ is the proportion of the events drawn from sample a. This functional is minimized when

$$\frac{1 - c(x)}{c(x)} = \frac{p(M_2)}{p(M_1)} L_{M_2/M_1}(x), \tag{22}$$

so $c(x)$ is indeed a monotonic function of L_{M_2/M_1} in the limit of infinite training data. Following Ref. [29], it is straightforward to show that it is also a monotonic function of $L_{g/q}$. Therefore, the classifier $c(x)$ is a suitable observable for extracting reducibility factors using the L-fit method with Eq. (14) or using the R-fit method with Eqs. (18) and (19).

The specific classifiers we use for our study are energy flow networks (EFNs) and particle flow network (PFNs) [14], based on the Deep Sets formalism [30]. The distinctive feature of EFNs is that they have an infrared-and-collinear-safe (IRC-safe) latent space:

$$c(x) = F\left(\sum_i z_i \Phi(\hat{p}_i)\right), \tag{23}$$

where the sum is taken over all the particles in the jet, z_i is the energy of particle i, and $\hat{p}_i$ is the direction of particle i. For a latent space of dimension ℓ, Φ is a per-particle trainable function with ℓ outputs, and F is a per-jet trainable function with ℓ inputs. Relaxing the requirement of IRC safety yields PFNs:

$$c(x) = F\left(\sum_i \Phi(p_i)\right), \tag{24}$$

where p_i is the full information about particle i, including possible charge/flavor labels. Both EFNs and PFNs can handle variable sized inputs and respect particle permutations, both of which are desirable properties for analyzing jets. For the observables in Sec. III B, we omit charge/flavor information, since otherwise we would be sensitive to the relative fraction of up-type versus down-type quarks as a function of jet rapidity.

F. Machine-learned regression

Classification without labels is actually a special case of a more general strategy to extract discriminants from mixed training data. Instead of defining two jet mixtures M_1 and M_2 and performing binary classification, we can instead label jets via a continuous parameter η and regress η on the jet properties. This strategy requires that the fraction of quark and gluon jets change as a function of η, but the properties of the jets themselves are unaffected. (While η can be any observable that satisfies the requirement above, in the scope of this paper we will constrain ourself to η be a jet pseudorapidity.) This implies the factorized probability structure:

$$p(\eta, x) = f p_q(\eta) p_q(x) + (1 - f) p_g(\eta) p_g(x), \tag{25}$$

where $0 < f < 1$ is the total quark fraction of the sample, and the probability densities in the above expression each integrate to 1.

We perform regression with the mean squared error loss functional:

$$L[h] = \int d\eta dx p(\eta, x)(\eta - h(x))^2, \tag{26}$$

where $h(x)$ is the regression function to be learned. This loss functional is minimized for

$$h(x) = \frac{\int d\eta \eta p(\eta, x)}{\int d\eta p(\eta, x)}, \tag{27}$$

such that $h(x)$ learns the expectation value of η as a function of x.

To see that $h(x)$ is monotonically related to $L_{g/q}$, we can plug Eq. (25) into Eq. (27):

$$r(x) = \frac{f\langle\eta\rangle_q + (1 - f)\langle\eta\rangle_g L_{g/q}(x)}{f + (1 - f) L_{g/q}(x)}, \tag{28}$$

where $\langle\eta\rangle_a = \int d\eta \eta p_a(\eta)$. As long as $\langle\eta\rangle_q$ and $\langle\eta\rangle_g$ differ, this is a monotonic function of $L_{g/q}$ and therefore defines optimal decision boundaries.

In the studies below, we show jet topics results using both classification and regression observables, where we find consistent results within uncertainties. Note that if a change of variables is applied to η, then the extracted $h(x)$ will change by a Jacobian factor. With finite training data, there is an optimal change of variables that minimizes the statistical uncertainty on the inferred $L_{g/q}(x)$, though we leave an investigation of that to future work.

III. JETS IN THE CMS OPEN DATA

In this section, we describe the real and synthetic datasets used in our study, which derive from a public release by the CMS experiment [11]. We list the observables used as inputs for topic modeling and demonstrate the challenge of sample dependence induced by the CMS detector. Going further than previous studies with CMS open data, we perform full phase-space, central-value unfolding using the OMNIFOLD technique [21,22]. Our ability to validate the unfolding is limited due to our having only a single Monte Carlo (MC) dataset with CMS full detector simulation, but we check for detector-level closure in several

distributions of interest and find that the unfolding works well at a qualitative level.

A. Review of CMS 2011A jet primary dataset

The CMS 2011 open data release corresponds to 2.3 fb^{-1} of proton-proton collisions at center-of-mass energy $\sqrt{s} = 7$ TeV [31]. Related trigger streams [32] are grouped together into primary datasets, and our analysis is based on the `Jet` primary dataset for CMS run 2011A [11], which includes single jet and dijet triggers. Each event from CMS has full particle-flow reconstruction [33–35]. This dataset was extensively studied, processed, and simplified in Ref. [12], which applied event geometry techniques based on the energy mover's distance [18] to explore the space of jets in real collider data. Our current study is based on the reprocessed "MIT open data" (MOD) formatted HDF5 files from Ref. [12], which were publicly released on the ZENODO platform [13]. See Ref. [36] for a recent analysis of multipoint correlators with this dataset.

The CMS 2011 release also has associated MC datasets for a variety of physics processes, simulated with a model of CMS based on GEANT 4 [37]. Relevant for our analysis, there is a sample of hard QCD scattering events generated by PYTHIA 6.4.25 [19] with tune Z2 [38]. These generated events are further processed using the CMS simulation and reconstruction software [39–46]. Both generator-level truth particles and simulation-level reconstructed objects are available in the same MOD HDF5 format on ZENODO [47–54].

For both the real and synthetic datasets, we keep up to the two jets in an event with the highest p_T. Jets are clustered by CMS using the anti-k_t algorithm with $R = 0.5$ [55,56]. As a baseline, jets are selected if their transverse momentum and pseudorapidity satisfy:

$$p_T \in [475, 525] \text{ GeV}, \qquad |\eta| < 1.9. \tag{29}$$

Jet substructure observables are computed using all truth particles (for generator-level data) or all particle flow candidates (PFCs, for simulation-level data). For the measured dataset, we applied CMS-provided jet energy correction factors [57] (which include area-median pileup subtraction [58]) and imposed the "medium" jet quality criteria [59,60] (quality ≥ 2 in the MOD format). For this simulated dataset, we only use events that correspond to generated parton-level p_T greater than 300 GeV and less than 1800 GeV, divided into six independent event samples [48–53]. Note that there is a very small tail of events from the $\hat{p}_T \in [170, 300]$ GeV and $\hat{p}_T \geq 1800$ GeV samples [47,54] that satisfy our selection criteria but are ignored for this study. Unlike the analysis in Ref. [12], we place no further cuts on the p_T values of the PFCs, nor do we separate out charged from neutral PFCs.

B. Choice of mixtures and observables

To define the two mixtures for topic modeling, we split the $p_T \in [475, 525]$ GeV jet sample into two mixtures of roughly equal sizes:

(i) Forward (M_1, quark enriched): $|\eta| \in [0.65, 1.9]$,
(ii) Central (M_2, gluon enriched): $|\eta| < 0.65$.

A related strategy of considering the more forward/central jet was pursued by ATLAS in Refs. [3,61,62]. As defined by PYTHIA 6.4.25 parton-level information, forward (central) jets have a quark fraction of

$$\text{PYTHIA Parton}: \begin{matrix} f_1 \simeq 0.689 \\ f_2 \simeq 0.510 \end{matrix}. \tag{30}$$

This relatively modest difference in quark fraction is nevertheless sufficient for topic extraction. For reference, we summarize the various quark fractions we encounter in this study in Table I.

Following the analysis of Ref. [1], we consider a suite of six well-studied jet substructure observables as inputs to topic modeling:

(i) N_{const}: Number of jet constituents,
(ii) N_{95}: Image activity [63]; i.e. number of pixels in a 33×33 jet image that contain 95% of the jet p_T,
(iii) p_T^D: Transverse momentum dispersion [64]; i.e., $\sum_i p_{T,i}^2 / (\sum_i p_{T,i})^2$,
(iv) τ_2: 2-subjettiness [65,66] with $\beta = 1$ and k_t-axes,
(v) τ_1: 1-subjettiness with $\beta = 1$ and k_t-axes,
(vi) m_{jet}: Jet mass.

TABLE I. Various quark fractions used in this analysis. The first row correspond to parton labels, as determined by PYTHIA. The second row is also from PYTHIA but includes weights from central value unfolding; this is a parton-level reference that can be compared to our jet topics result. The third row (in bold) is the main result from our study, using jet topics to extract an operational definition of "quark" fractions in the central-value-unfolded CMS 2011 open data. The fourth row, studied in the Appendix, comes from performing the jet topics procedure on the PYTHIA Monte Carlo samples.

Sample	Forward (f_1)	Central (f_2)	Equation
PYTHIA Parton	0.689	0.510	Eq. (30)
PYTHIA Parton with CVU	0.711	0.543	Eq. (31)
CMS 2011 with OMNIFOLD: **CVU Jet Topics**	$\mathbf{0.708 \pm 0.025}$	$\mathbf{0.561 \pm 0.033}$	Eq. (55)
PYTHIA Jet Topics	0.649 ± 0.012	0.466 ± 0.011	Eq. (A1)

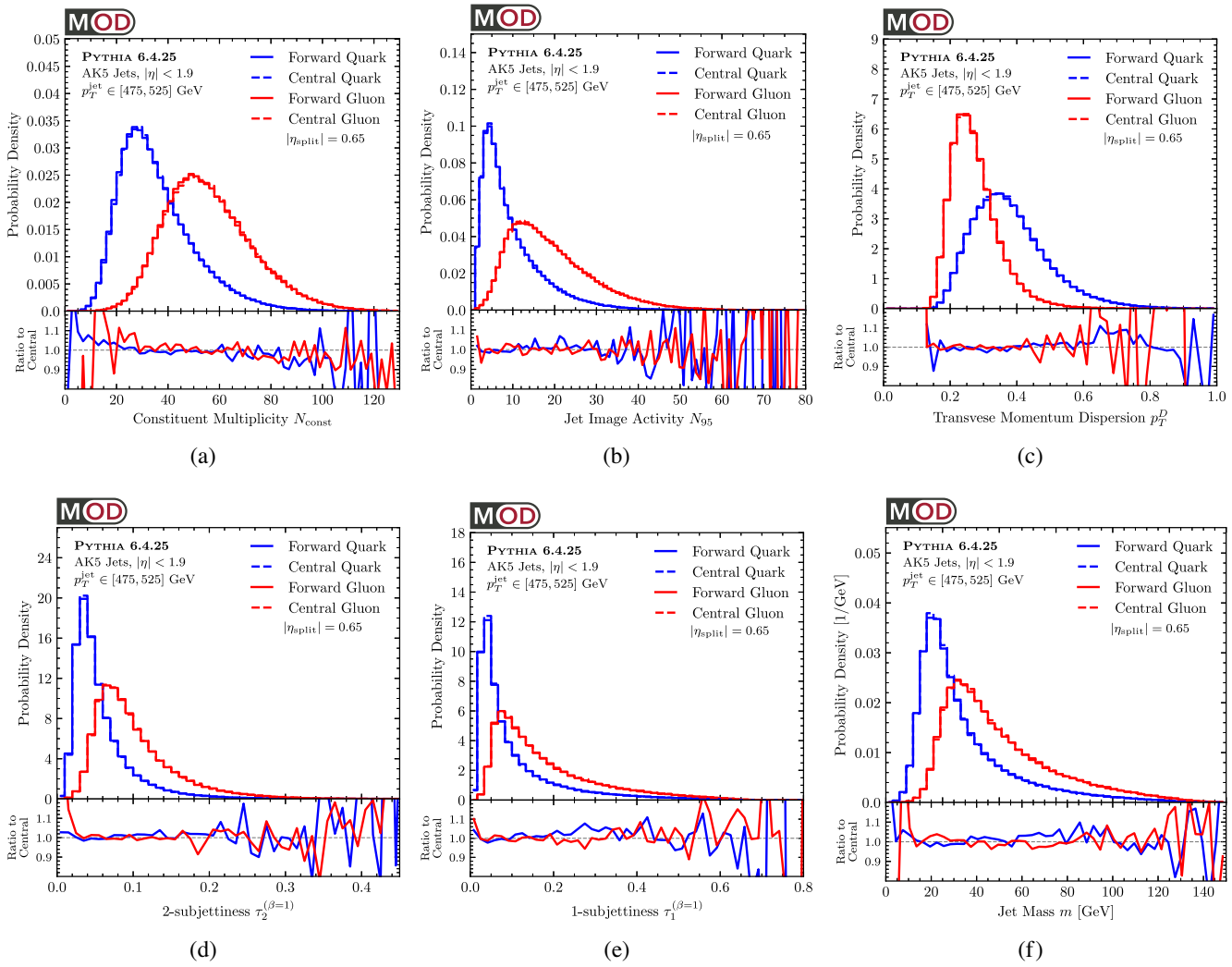

FIG. 1. Distributions in PYTHIA 6.4.25 for the six substructure observables from Sec. III B: (a) constituent multiplicity, (b) image activity, (c) momentum dispersion, (d) 2-subjettiness, (e) 1-subjettiness, and (f) jet mass. Using parton truth labels from the PYTHIA event record, the distributions for quark (blue) and gluon (red) are compared between forward jets (solid, $|\eta| \in [0.65, 1.9]$) and central jets (dashed, $|\eta| < 0.65$). As expected from leading-power factorization, the differences are modest such that sample independence holds to an excellent approximation.

The first observable (N_{const}) is known to be good quark/gluon discriminant. The middle three observables (N_{95}, p_T^D, and τ_2) yield moderate quark/gluon separation power. The last two observables (τ_1 and m_{jet}) exhibit Casimir scaling in their quark/gluon separation [67], limiting their usefulness for jet topic modeling.

We also consider four machine-learned observables, as discussed in Secs. II E and II F:

(i) EFN_{cls}: classification of forward versus central jets with an EFN,
(ii) PFN_{cls}: classification of forward versus central jets with a PFN,
(iii) EFN_{reg}: regression for jet pseudorapidity $|\eta|$ with an EFN,
(iv) PFN_{reg}: regression for jet pseudorapidity $|\eta|$ with a PFN.

For the PFNs, we input the full particle four-vector information, but exclude charge/flavor information since we do not wish to distinguish up-type from down-type quarks. Apart from the input and output values, the network architectures are the same, with a latent space dimension of $\ell = 128$, two hidden layers of 100 nodes for the per-particle function Φ, and three hidden layers of 100 nodes for the per-jet function F. The activation functions are rectified linear unit (ReLU) for classification and leaky ReLU for regression. Networks are built using the ENERGYFLOW package [68] based on KERAS [69] and trained using the ADAM optimizer [70].

To avoid training complications associated with the real CMS datasets, we trained these four observables on the particle-level PYTHIA 6.4.25 samples. This is valid from the perspective of jet topics, since any observable (even a

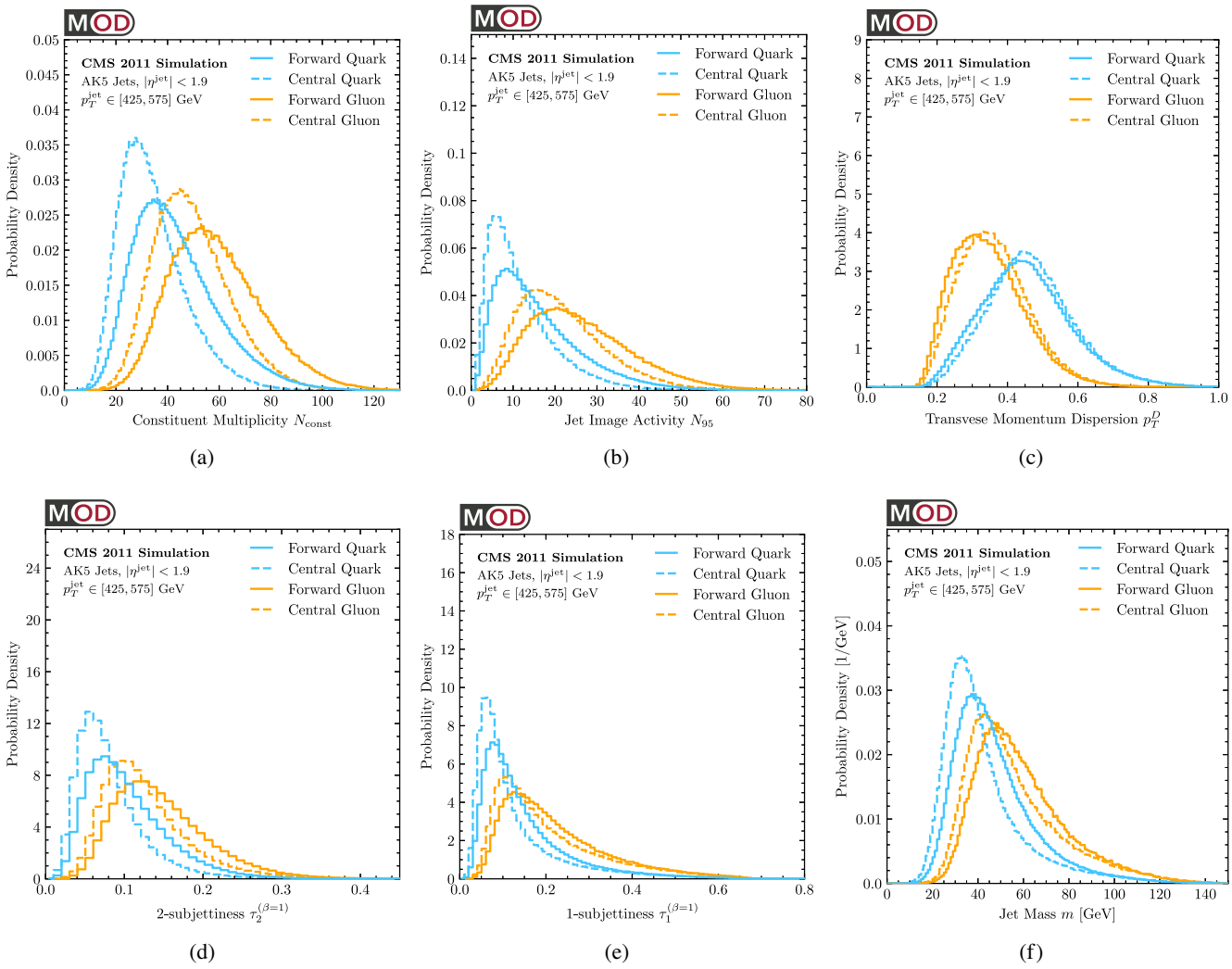

FIG. 2. Similar to Fig. 1 but now for detector-level distributions after PYTHIA events are passed through the CMS detector simulation. We see that quark jets (cyan) and gluon jets (orange) look different in the forward and central parts of the CMS detector, which violates the assumption of sample independence and necessitates the use of unfolding.

machine-learned observable trained on imperfect synthetic data) can be used as an input to the jet topics procedure. Of course, such a suboptimal observable will not provide as good of an estimate of the true likelihood ratio, generically resulting in residual mixing between the categories. We checked that similar results could be obtained by training on the central-value-unfolded samples described below in Sec. III D, though the results were not as stable as we would have liked between different numbers of unfolding iterations.[2]

[2]Because topic modeling is sensitive to the very high and very low purity regions of phase space, instabilities can arise if the machine-learning algorithm picks up on small changes near the end points of the classifier. These instabilities can be amplified in the presence of large weight variations, which we encounter in Fig. 3 below. It would be interesting to see if these instabilities could be reduced if the unfolding were explicitly regularized, as opposed to implicitly regularized by the number of unfolding iterations.

C. The challenge of sample dependence

A key requirement for jet topic modeling is sample independence, which means that each mixture distribution should be a linear combination of the same underlying quark and gluon distributions. To test this, we can use the MC samples, where we have quark/gluon labels based on PYTHIA truth parton information. As shown in Fig. 1, sample independence holds to an excellent approximation for generator-level jets, with forward quarks and central quarks having nearly identical distributions, and similarly for gluons.

At detector-level, though, we see considerable sample dependence in Fig. 2. This sample dependence is due to a number of factors, including changes in the detector geometry and granularity, changes to object reconstruction efficiencies, and differing response to pileup and other sources of noise. In general, forward jets have more reconstructed constituents and therefore larger values of

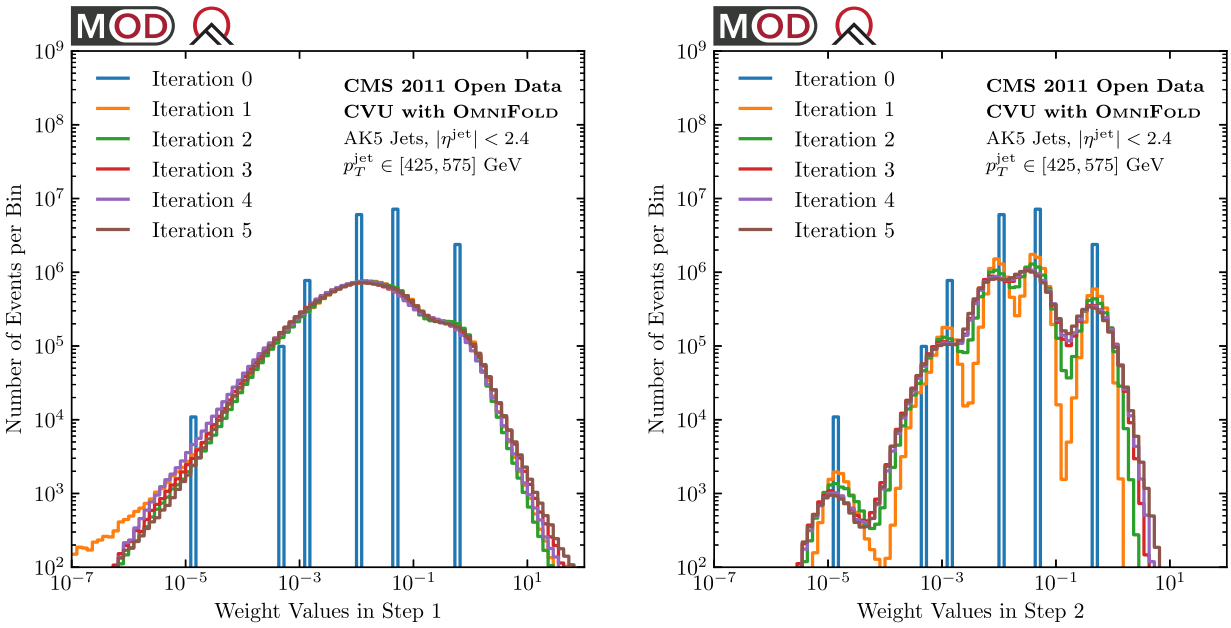

FIG. 3. The weight distribution produced by central value unfolding after (a) step 1 and (b) step 2 of the OMNIFOLD algorithm. The initial weights from the CMS open data are shown in blue, where the spikes arise because there are six independent samples that have been individually unweighted. The results in this paper correspond to using weights from iteration 4, after step 2.

N_{95}, τ_2, τ_1, and $m_{\rm jet}$. (More constituents typically implies smaller values of p_T^D.) This dependence of the CMS detector response on the overall jet kinematics introduces undesirable sample dependence in our mixed samples, motivating the need for (central value) unfolding.

D. Central value unfolding with OMNIFOLD

In a full experimental analysis, one would perform detector unfolding to mitigate the impact of detector effects. Such an unfolding would also account for various sources of systematic uncertainties, including the choice of generator used to extract the response matrix. See Ref. [71] for a review of jet substructure measurements. For this study, we perform *central value unfolding* (CVU), where we correct detector effects based on the response matrix derived from a single MC sample without assigning systematic uncertainties.[3] This is sufficient to address the issues of sample dependence, though not sophisticated enough to derive distributions suitable for quantitative studies, e.g., comparisons to theoretical calculations or tunings of event generators. The statistical uncertainties on the unfolding are estimated in a naive way through the variance of the unfolded weights (i.e., the usual way one would account for uncertainties in a weighted event sample). We checked that these uncertainties are comparable in size to the statistical uncertainties associated with the prior distribution.

Because the absence of systematic uncertainties will be viewed as anathema to some readers (especially the authors of Refs. [3,72]), we want to explain more about our philosophy. We will leverage a full phase space unfolding strategy called OMNIFOLD [21,22], which assigns each MC event a weight at the truth particle level.[4] With these per-event weights, one can in principle compute the unfolded distribution for any observable, including ones not envisioned at the time of the unfolding. Full phase space unfolding is a great opportunity for the field, and there are already discussions about how to release per-event unfolded data in a suitable format [75]. It is also a great challenge, though, since it is not clear how to validate the unfolding for arbitrary down-stream analyses, nor is it clear what the best strategy would be to assign per-event systematic uncertainties. Furthermore, OMNIFOLD is based on neural networks, which introduces additional sources of uncertainties from the initialization and training paradigms. For these reasons, we feel it is prudent to perform a proof-of-concept study that shows how unfolding can indeed mitigate the issue of sample dependence, without attempting to make a quantitative claim

[3]We use the term "central value" to refer to the central value with respect to systematic variations. This differs from the central value associated with statistical or initialization variations.

[4]See Ref. [73] for an application of OMNIFOLD by the H1 collaboration on an eight-dimensional phase space, with proper uncertainty quantification. See Ref. [74] for a preliminary analysis by the H1 collaboration to unfold six jet substructure observables in step 2 of OMNIFOLD by using the full phase space at step 1.

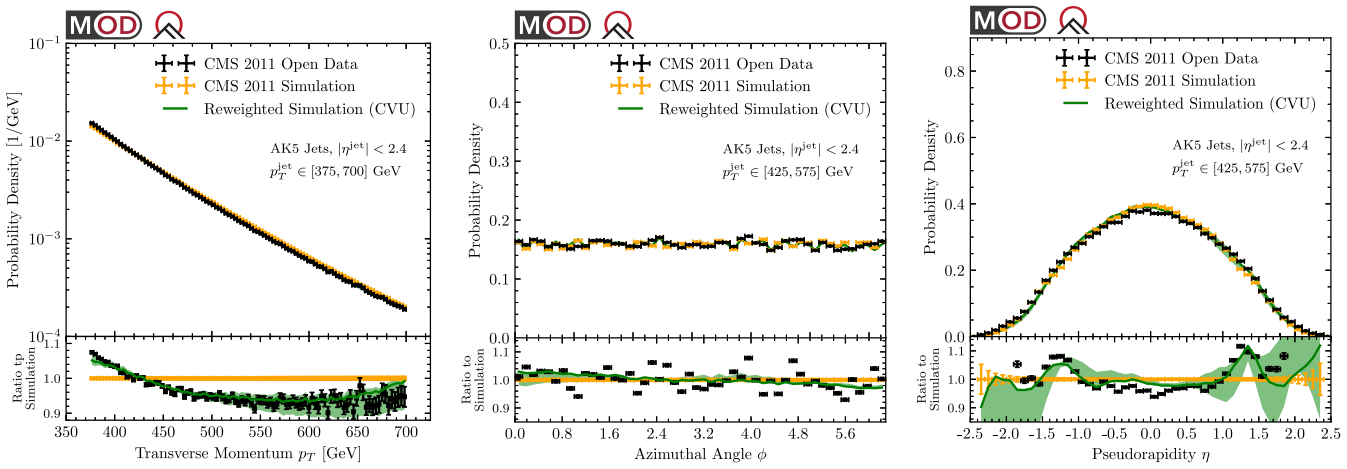

FIG. 4. Validation of central value unfolding with OMNIFOLD applied to the CMS 2011 open data. Shown are distributions for (a) jet p_T, (b) jet azimuth, and (c) jet pseudorapidity. The detector-level distributions after using the unfolding weights (green) match better to the CMS measured distributions (black) than to PYTHIA run through the CMS detector simulation (orange). Note the smaller jet p_T range for the azimuth and pseudorapidity plots. The error bars on the open data/simulation correspond to statistical uncertainties, while the green band corresponds to using three or five iterations of OMNIFOLD compared to the default of four.

about the degree of mitigation or the systematic uncertainties on the final distributions.

With that apologia, we apply the OMNIFOLD algorithm to all reconstructed jets in the range $p_T \in [375, 700]$ GeV and $|\eta| < 2.4$. OMNIFOLD can be viewed as an unbinned version of iterated Bayesian unfolding [76]. In step 1 of the algorithm, event weights are computed such that the detector-level full phase-space density of the simulated sample match that of the observed data. In step 2 of the algorithm, the event weights are refined such that the weights can be expressed as a function of the particle-level inputs. For this analysis, we do not account for migration of events out of the jet selection range [22], though we later restrict our attention to a smaller p_T range to avoid edge complications. The weights distributions are shown in Fig. 3, where one can see that the starting weights for the six initial samples are smeared out in each OMNIFOLD iteration. These weights are publicly available at Ref. [23].

We use four iterations of OMNIFOLD as the default, and checked that the qualitative features were robust with one more or one fewer iteration. More iterations lead to a better fit to the data, at the expense of larger fluctuations. The unfolding weights are computed using a PFN, as parametrized in Eq. (24). The model takes as input a tuple of all particle 4-vectors, experimentally accessible particle-type information ("PFN-Ex," see Ref. [14]), and four global features: jet p_T, jet η, jet ϕ, and jet mass. For step 1, we use a latent space with $\ell = 256$, where the per-particle Φ function has two hidden layers of 128 nodes each, and the per-jet F function has three hidden layers of 128 nodes each. For step 2, we use $\ell = 192$, where Φ has two hidden layers of 100 nodes, and F has three hidden layers of 100 nodes. The ReLU function is used as an activation function on each layer besides the output layer that uses softmax.

We emphasize that, after step 2 of OMNIFOLD, the unfolded weights are assigned to the truth particle-level information. Combining the weights with PYTHIA parton-level information, we can derive the effective quark and gluon fraction of the central-value-unfolded sample:

$$\text{PYTHIA Parton with CVU:} \quad \begin{matrix} f_1 \simeq 0.711 \\ f_2 \simeq 0.543 \end{matrix}. \tag{31}$$

These fractions are unphysical for two reasons. First, they correspond to parton-level information that only have meaning at leading logarithmic accuracy. Second, the OMNIFOLD weights are derived from particle-level alone, without reference to parton-level information, so these fractions do not correspond to a true generative model in the sense of Eqs. (1) and (2). Nevertheless, Eq. (31) is a useful benchmark to guide the eye in interpreting the jet topic modeling results.

E. Unfolding validation at detector level

To validate the unfolding, we compare detector-level distributions before and after OMNIFOLD. For all of the plots in this subsection, we emphasize that distributions are shown at detector level, and we only use unfolding to determine the event weights (particle-level distributions are shown later in Secs. IV and V). In Fig. 4(a), we show the result of using the unfolding weights on the overall jet transverse momentum, in the range of $p_T \in [375, 700]$ GeV and $|\eta| < 2.4$. The spectrum from the MC (orange dots) differs from that of the CMS data (black dots) by upwards of 10%, but the simulation spectrum with unfolding weights (green line) successfully matches the spectrum seen in the CMS data. The shaded

green region is the envelope from considering three, four, or five OMNIFOLD iterations, with four being the default. The weights are obtained after step 2 of OMNIFOLD.

To reduce p_T edge effects, we further restrict to $p_T \in [425, 575]$ GeV for the remaining distributions. Using the unfolding weights has a negligible effect on the azimuthal spectrum in Fig. 4(b), which suggests that the CMS simulation does not match the detailed azimuthal structure of the CMS detector.

For the pseudorapidity spectrum in Fig. 4(c), there is a slight clifflike feature at around $|\eta| \approx 1.4$, which is roughly where the detector barrel meets the detector end cap. Using the unfolding weights adjusts the truth-level pseudorapidity spectrum in the vicinity of this cliff such that the detector-level pseudorapidity structure better matches the data. This is an example of a potential systematic issue that requires a detailed understanding of jet reconstruction at CMS. If the barrel/end cap interface is incorrectly modeled in the CMS simulation or if there are sources of contamination not captured in simulation and not mitigated by the jet quality criterion, then unfolding could erroneously modify the truth-level pseudorapidity spectrum. This is particularly important for our analysis, because quark/gluon composition of the samples are affected by the pseudorapidity spectrum. Note that this issue is potentially present with any unfolding algorithm, not just OMNIFOLD, and requires expert experimental knowledge to assess.

To study the substructure of jets, we now restrict our attention to the detector-level jet kinematic range of

$$p_T \in [475, 525] \text{ GeV}, \qquad |\eta| < 1.9. \tag{32}$$

The tightening of the pseudorapidity range is because the tracker extends only out to $|\eta| < 2.4$, so we want to make sure that all charged particles within the jet radius of $R = 0.5$ have a chance to be reconstructed. In Fig. 5, we show detector-level distributions for the six observables from Sec. III B. Overall, unfolding improves the agreement

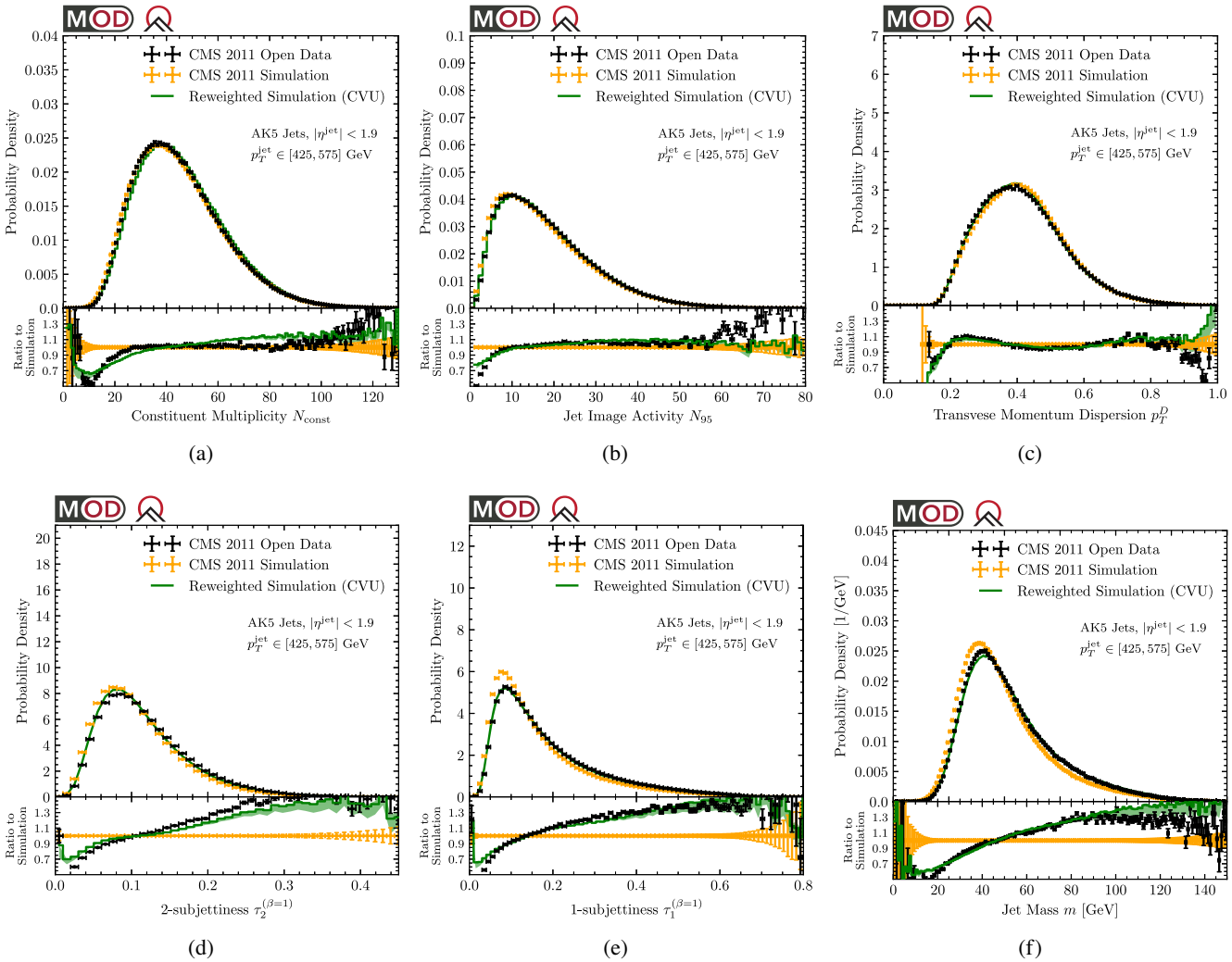

FIG. 5. Similar to Fig. 4 but for the six substructure observables from Sec. III B. Here, we are using a narrower jet p_T and pseudorapidity range to mitigate edge effects in the unfolding.

between the detector-level distributions in regions of phase space with sufficient numbers of events, particularly for N_{95}, p_T^D, τ_1, and m_{jet}. Using the unfolding weights improves the agreement for τ_2, though there are still noticeable discrepancies. Using the unfolding weights makes comparatively little impact on N_{const}, suggesting that the response of this observable may not be perfectly modeled by the CMS simulation or that there are correlations in the PYTHIA sample that cannot be corrected via unfolding.

At a qualitative level, we conclude that using the unfolding weights satisfies this basic detector-level closure test. In the context of a complete measurement, deviations from closure would be assessed as a systematic uncertainty. Though not shown, we checked that the particle-level distributions exhibit the same degree of sample independence as seen in Fig. 1.

IV. EXTRACTING REDUCIBILITY FACTORS

In this section, we apply the methods described in Sec. II to extract reducibility factors and determine the quark fractions of the forward/central CMS open data jet samples. All studies are based on the CVU dataset obtained in Sec. III D. All plots in the remainder of this paper are based on particle-level information.

A. Unfolded forward/central distributions

The starting point for the jet topics analysis is particle-level observable distributions. After central value unfolding, we have a sample of forward jets ($|\eta| \in [0.65, 1.9]$) and central jets ($|\eta| < 0.65$) with full particle-level kinematic information. We restrict our attention to the particle-level jet kinematic range of

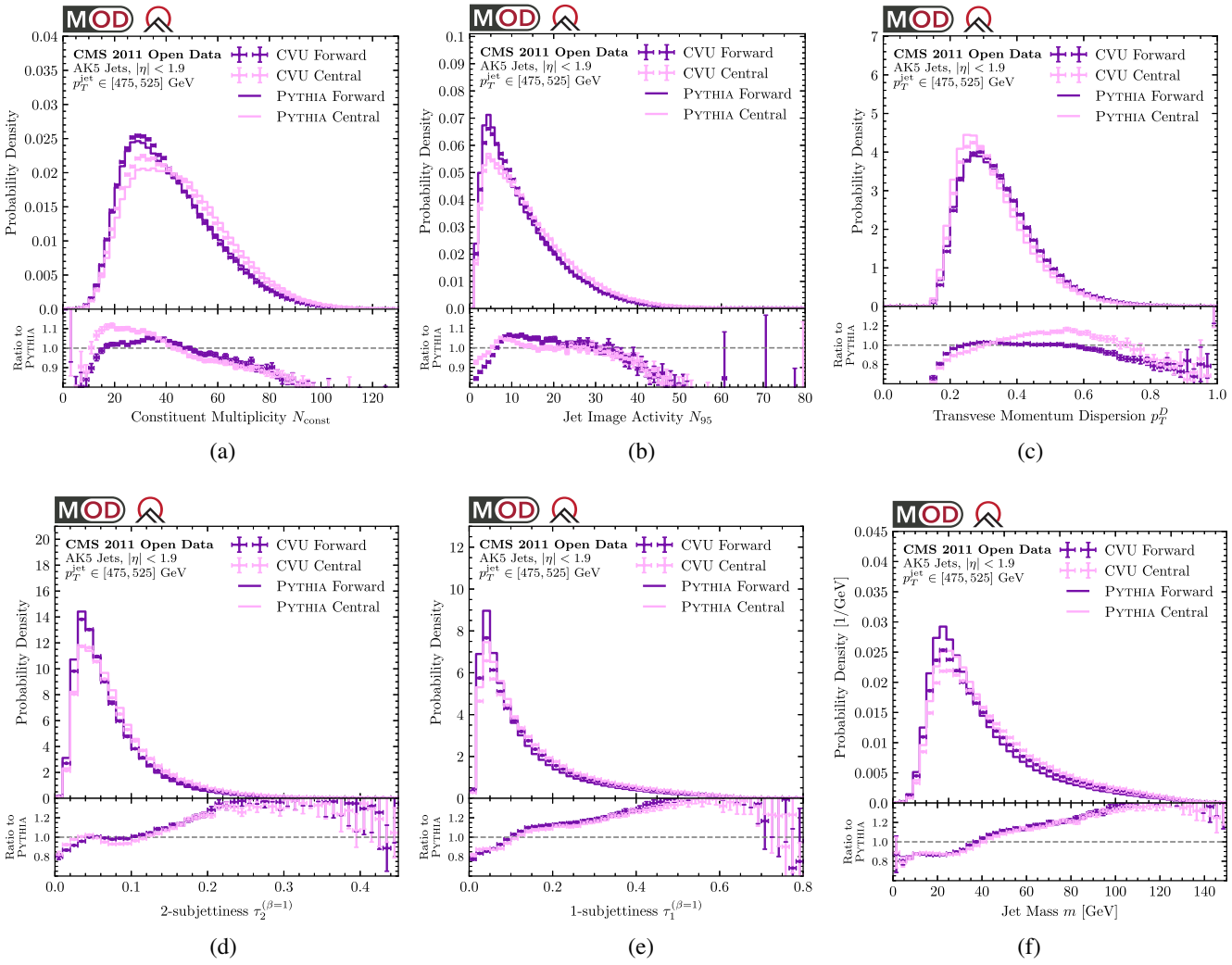

FIG. 6. Particle-level distributions for forward (purple) and central (pink) jets for the six substructure observables from Sec. III B. We compare central-value-unfolded results from the CMS 2011 open data (data points) to PYTHIA 6.4.25 (solid curve), where the differences are subtle but noticeable. The CVU distributions are inputs for the subsequent jet topics analyses. See the Appendix for results obtained from the PYTHIA distributions. The error bars correspond to statistical uncertainties only.

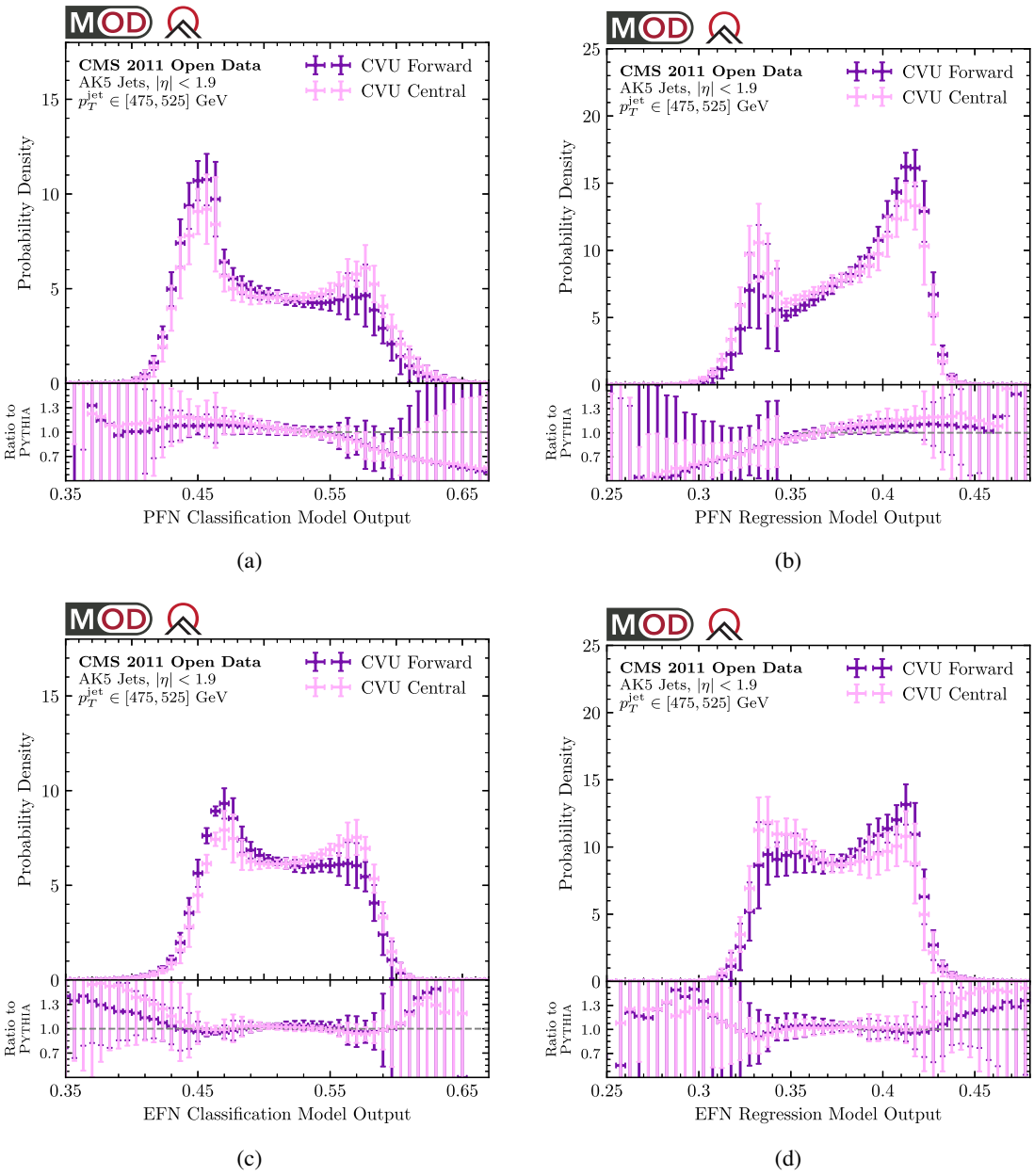

FIG. 7. Similar to Fig. 6, but for the four machine-learned observables from Sec. III B: (a) PFN classification, (b) PFN regression, (c) EFN classification, and (d) EFN regression. The uncertainties include fully correlated uncertainties from the training. For an individual training, there is better separation power than implied here. For visual clarity, we have omitted the PYTHIA distributions from the top panels.

$$p_T \in [475, 525] \text{ GeV}, \qquad |\eta| < 1.9. \tag{33}$$

This yields a sample of 3.25 millions jets for analysis.

In Fig. 6, we plot the six observables from Sec. III B for these two samples. While the forward and central distributions are rather similar, the differences are enough to yield a nontrivial result from jet topic modeling. As expected, the forward jets are more quarklike than the central jets, with fewer jet constituents and smaller invariant masses.

The unfolding induces 10%–20% differences from the baseline PYTHIA distributions, which will have a noticeable impact on the results below. The unfolding primarily affects the constituent multiplicity and jet mass distributions, with more modest differences in the other observables. In the Appendix, we repeat the jet topics analysis on the PYTHIA samples for comparison.

In Fig. 7, we show distributions for the four machine-learned observables from Sec. III B. We emphasize that these observables were trained on the PYTHIA samples but the distributions in the top panel are derived from the CVU

samples. The uncertainties in these distributions include both statistical uncertainties and fully correlated variance from doing 10 different training runs. While the training uncertainties are large in the bulk of the distribution, they are small (and correlated) near the end points, which is why we get relatively small uncertainties on the extracted reducibility factors. We again see 10%–20% differences from the PYTHIA baseline in the core of the distribution, which will show up in the jet topics results.

B. Anchor bin method

As our first method to extract reducibility factors, we apply the anchor bin method from Sec. II B. This method requires choosing a histogram binning for the observable distributions $p_{M_a}(\mathcal{O})$. We use a quantile-based binning to address the statistical noise of the samples. Specifically, we chose the bins to minimize the number of events per bin, while simultaneously ensuring that no fewer than 1.0% of central jets and 1.0% of forward jets are contained in each bin. For continuous observables, this variable binning yields approximately the same number of events per bin. For discrete observables like constituent multiplicity, this forces larger bin widths in the end point regions where statistical noise would otherwise dominate.

With around 15k events from each sample per bin, the Poisson uncertainty in the bin occupancy is around 1%. To compute the statistical uncertainties, we use the standard formulas for weighted events. Letting B_i be the set of all events assigned to the ith histogram bin and p_i be the probability in that bin, we have

$$\bar{p}_i = \sum_{x \in B_i} \omega(x), \tag{34}$$

$$\mathrm{Var}[p_i] = \sum_{x \in B_i} \omega(x)^2, \tag{35}$$

where x represents the full kinematics, and $\omega(x)$ is the event weight. These uncertainties are computed for the forward and central mixtures separately. For the log-likelihood ratio $f(\mathcal{O}) = \ln L_{M_2/M_1}(\mathcal{O})$, we use standard error propagation in the Gaussian limit for each bin:

$$\Delta f(\mathcal{O}) = \sqrt{\left(\frac{\Delta p_{M_1}(\mathcal{O})}{p_{M_1}(\mathcal{O})}\right)^2 + \left(\frac{\Delta p_{M_2}(\mathcal{O})}{p_{M_2}(\mathcal{O})}\right)^2}. \tag{36}$$

In Fig. 8, we show a typical log-likelihood ratio, in this case for constituent multiplicity. To account for statistical noise, we use a slightly different procedure to determine the anchor bins from the one described in Sec. II B. Following Refs. [1,2], instead of directly searching for minimum and maximum values of $f(\mathcal{O})$, we choose the anchor bins based on values adjusted for uncertainty:

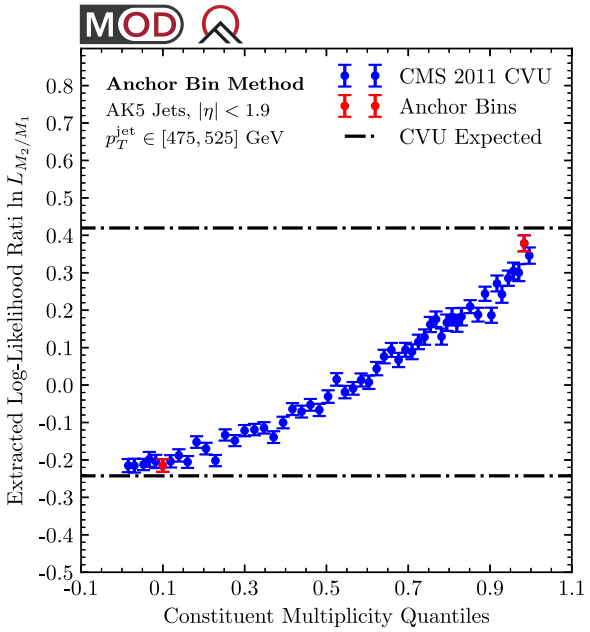

FIG. 8. Demonstration of the anchor bin method on the CMS 2011 open data with central value unfolding. After binning the data into quantiles, we plot the log-likelihood ratio between the central and forward samples. The anchor bins, shown in red, are at the extrema of this ratio after accounting for uncertainties. The horizontal dashed lines correspond to the PYTHIA parton-level expectation in Eq. (31). The error bars correspond to statistical uncertainties only.

$$\kappa_{21} = \min_{\mathcal{O}}(f(\mathcal{O}) + \Delta f(\mathcal{O})), \tag{37}$$

$$\kappa_{12} = \max_{\mathcal{O}}(f(\mathcal{O}) - \Delta f(\mathcal{O})). \tag{38}$$

The red dots in Fig. 8 correspond to the anchor bins. Their values match reasonably well to the horizontal dashed lines, which are the expected reducibility factors given the quark fractions estimated in Eq. (31). Full results from the anchor bin method are summarized in Fig. 11 below.

C. Log-likelihood ratio fit method

We now turn to the log-likelihood ratio fit (L-fit) method from Sec. II C. Like the anchor bin method, it is based on the log-likelihood ratio of the two mixtures, but it uses curve fitting to a functional form.

This method still has some dependence on the choice of binning, though the binning impact is much smaller than for the anchor bin method. Because of the reduced binning sensitivity, we use finer bins than in Sec. IV B, such that each bin contains at least 0.1% of jets from the central and forward mixtures. We use the same central value and statistical uncertainty estimates from Eqs. (34) and (35).

When expressed in quantiles, the log-likelihood ratio $f(\mathcal{O}^{\mathrm{quant}}) = \ln L_{M_2/M_1}(\mathcal{O}^{\mathrm{quant}})$ is fit to a polynomial of the form:

$$f(\mathcal{O}^{\text{quant}}; b) = \sum_{k=0}^{K} b_k \mathcal{L}_k(\mathcal{O}^{\text{quant}}), \tag{39}$$

where K is the maximum degree used in the fit and $\mathcal{L}_k(\mathcal{O}^{\text{quant}})$ is a Legendre polynomial of degree k as a function of $(2\mathcal{O}^{\text{quant}} - 1)$, such that the Legendre polynomial domain $[-1, 1]$ is mapped to the quantile domain $[0, 1]$. The fit coefficients are determined by minimizing the χ^2 statistic:

$$\chi^2 = \sum_{i,j=1}^{|B|} (\bar{f}_i - f(\mathcal{O}_i; b)) W_{ij} (\bar{f}_j - f(\mathcal{O}_j; b)), \tag{40}$$

where $\bar{f}_i$ is the bin value, W_{ij} is the inverse covariance matrix, and $|B|$ is the total number of bins. Because the bins are uncorrelated (up to normalization effects), the covariance matrix is diagonal, with diagonal elements corresponding to the bin variances:

$$W_{ij} = \frac{1}{\Delta f_i^2} \delta_{ij}. \tag{41}$$

To implement χ^2 minimization, we use the `curve_fit()` algorithm from the SCIPY PYTHON library [77] to perform gradient descent on the b_k coefficients.

The L-fit method depends on the maximum polynomial degree K of the fit. Following Ref. [78], we use the Akaike information criterion (AIC) [79] to choose the optimal model that minimizes

$$\chi^2 + 2(K + 1), \tag{42}$$

where $K + 1$ is the number of fit parameters. To compute the reducibility factors, we evaluate the optimal model at the end points. Because the rescaled Legendre polynomials are either 1 or -1 at $\mathcal{O}^{\text{quant}}_{\text{min}} = 0$ and $\mathcal{O}^{\text{quant}}_{\text{max}} = 1$, this yields relatively simple expressions for the anchor values:

$$a_{21} = \sum_{k=0}^{K} b_k^{\text{opt}} (-1)^k, \tag{43}$$

$$a_{12} = \sum_{k=0}^{K} b_k^{\text{opt}}, \tag{44}$$

which can be converted into reducibility factors using Eq. (11).

The χ^2 minimization method not only allows us to extract optimal coefficients b^{opt} but also their covariance matrix near the optimum point $\text{Cov}[b_i^{\text{opt}}, b_j^{\text{opt}}]$. This matrix contains information about the statistical uncertainties associated with the method and allows us to compute the uncertainties of the anchor values via

$$\Delta a_{21}^2 = \sum_{i,j=0}^{K} (-1)^{i+j} \text{Cov}[b_i^{\text{opt}}, b_j^{\text{opt}}], \tag{45}$$

$$\Delta a_{12}^2 = \sum_{i,j=0}^{K} \text{Cov}[b_i^{\text{opt}}, b_j^{\text{opt}}]. \tag{46}$$

An example of the L-fit method is shown in Fig. 9. In Fig. 9(a), we show the optimal polynomial fit for the case of multiplicity. The anchor values, which are indicated by the red horizontal line, are comparable to the estimated values from Eq. (31) In Fig. 9(b), we show the extracted anchor values as a function of the maximum degree K, where the red dots indicate the choice that minimizes the AIC expression in Eq. (42).

In addition to the statistical uncertainty, we assess a systematic uncertainty from model selection. Specifically, we minimize two alternative version of Eq. (42), where the proportionality coefficient of 2 in front of $K + 1$ is replaced by either 1 or 4. The half-difference between the minimum and maximum of the anchor value produced by these three models is used to estimate the systematic uncertainty, which is added to the statistical uncertainty in quadrature. Typically the anchor values are quite robust to this coefficient change. Full results from the L-fit method are summarized in Fig. 11 below.

D. ROC curve fit method

Our last method for extracting reducibility factors is the ROC curve fit (R-fit) method from Sec. II D. In this approach, we build a ROC curve using observable cumulative probability distributions and then fit this curve to a functional form to determine its slope at the end points. While the R-fit method does not require binning the data, for reasons of computational efficiency we use the same binning procedure as in Sec. IV C for continuous observables. For discrete observables, we use the full set of bins.

Unlike for the probability distribution function (PDF), mean values for the cumulative distribution function (CDF) are correlated. Therefore, to assess statistical uncertainties, we need to compute a full covariance matrix for the CDF. For cumulative bin C_i, the cumulative probability $\bar{P}_i$ is estimated in the standard way:

$$\bar{P}_i = \sum_{x \in C_i} \omega(x). \tag{47}$$

The covariance between cumulative bins C_i and C_j with $i \leq j$ (i.e., bin i is fully contained by bin j) is

$$\text{Cov}[P_i, P_j] = \sum_{x \in C_i} \omega^2(x) - \frac{1}{N} \bar{P}_i \bar{P}_j, \tag{48}$$

where N is the total number of events. Note that the symmetry between i and j is broken because the sum in the

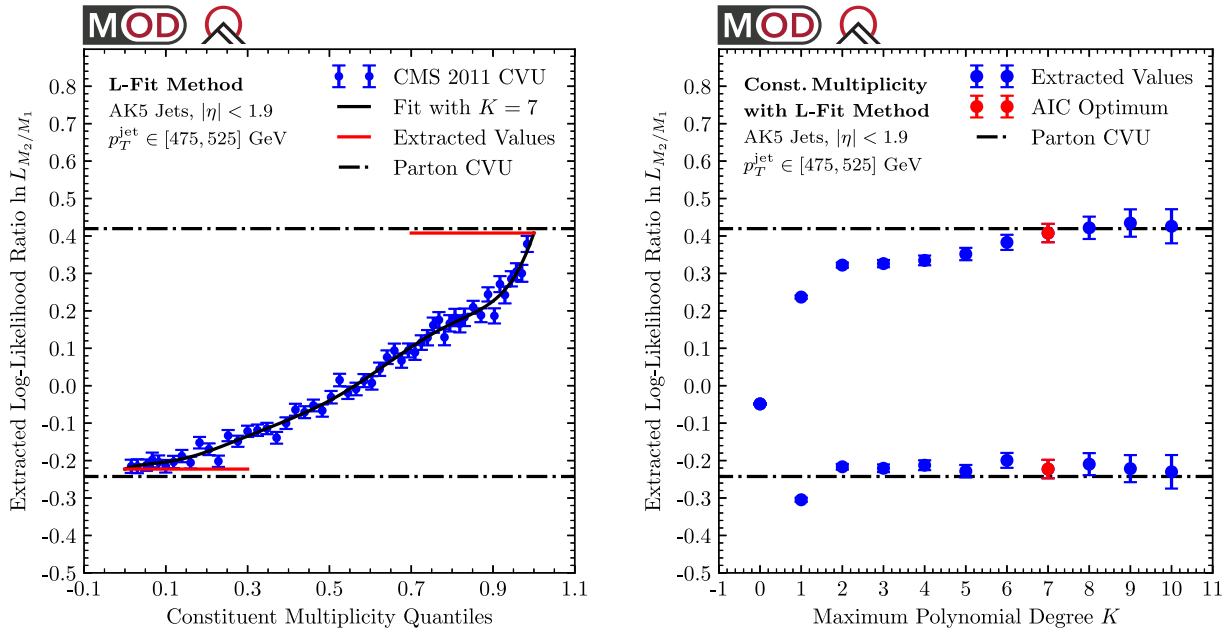

FIG. 9. Demonstration of the L-fit method on the CMS 2011 open data with central value unfolding. (a) The same likelihood ratio as Fig. 8 but now fit to a $K = 7$ degree polynomial. The anchor values (red) are at the extrema of the fitted curve. (b) Extracted anchor values as a function of the degree of the fit polynomial. The red point indicates the optimum determined from the AIC in Eq. (42). The error bars correspond to statistical uncertainties only.

first term only goes over entries in bin i. Unlike for Eq. (35), here we are keeping track of the $1/N$ correction to the covariance.

Evaluating the statistical uncertainty for a ROC curve is more complex than for the methods in Secs. IV B and IV C. The ROC curve expresses the CDF for M_2 as a function of the CDF for M_1. Thus, there are statistical uncertainties in the x coordinate of the ROC curve that need to be converted to an additional effective noise in the y coordinate in order to do error propagation on the fit form. Let r_i be the y value of the ROC curve associated with the ith cumulative bin, and let superscripts $^{(a)}$ refer to the mixture M_a. The covariance between r_i and r_j is

$$\text{Cov}[r_i, r_j] = \text{Cov}[P_i^{(2)}, P_j^{(2)}] + \frac{\bar{p}_i^{(2)}\bar{p}_j^{(2)}}{\bar{p}_i^{(1)}\bar{p}_j^{(1)}}\text{Cov}[P_i^{(1)}, P_j^{(1)}], \tag{49}$$

where $p^{(a)}$ are the PDFs of the mixture distributions (i.e., the slopes of the CDFs). With this derivation, we are assuming that there are no correlations between the two mixtures.

To extract the anchor values, we fit the ROC curve to

$$f(r) = r + r(1-r)\left(\sum_{k=0}^{K} b_k \mathcal{L}_k(r)\right), \tag{50}$$

where $\mathcal{L}_k(x)$ is again a rescaled Legendre polynomial of degree k defined on the interval [0, 1]. Like in Sec. IV C, we perform χ^2 minimization with covariances given by Eq. (49), with the optimal maximum degree K and associated fit coefficients b_k^{opt} determined by the AIC in Eq. (42).

As shown in Eqs. (18) and (19), the reducibility factors are related to the slopes at the end point of the ROC curve:

$$\kappa_{21} = \left.\frac{df}{dr}\right|_{r=0} = 1 + \sum_{k=0}^{K} b_k^{\text{opt}}(-1)^k, \tag{51}$$

$$\frac{1}{\kappa_{12}} = \left.\frac{df}{dr}\right|_{r=1} = 1 + \sum_{k=0}^{K} b_k^{\text{opt}}. \tag{52}$$

The associated uncertainties are

$$\Delta\kappa_{12} = \sum_{i,j=0}^{K}(-1)^{i+j}\text{Cov}[b_i^{\text{opt}}, b_j^{\text{opt}}], \tag{53}$$

$$\Delta\left(\frac{1}{\kappa_{21}}\right) = \sum_{i,j=0}^{K}\text{Cov}[b_i^{\text{opt}}, b_j^{\text{opt}}], \tag{54}$$

where $\text{Cov}[b_i^{\text{opt}}, b_j^{\text{opt}}]$ is determined from the χ^2 fit.

An example of the R-fit methods is shown in Fig. 10. In Fig. 10(a) we show an example ROC curve, with the fitted form and the extracted slope values. The change in reducibility factors as a function of K is shown in

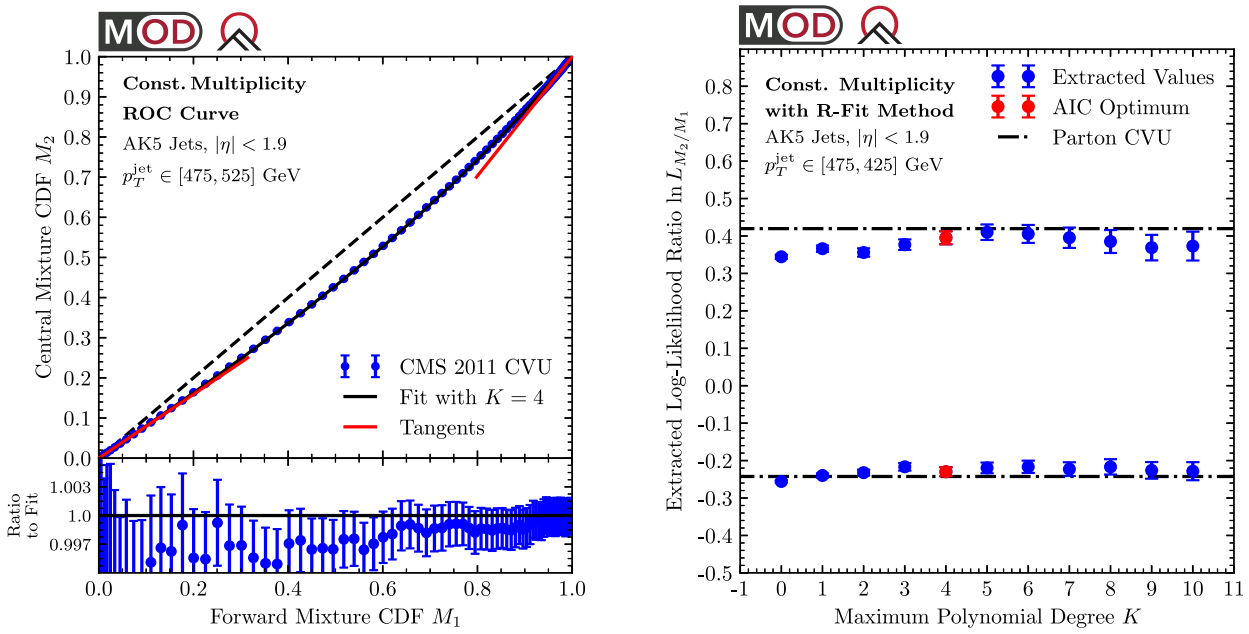

FIG. 10. Demonstration of the R-fit method on the CMS 2011 open data with central value unfolding. (a) ROC curve for distinguishing the central from forward samples, along with a polynomial fit of modified degree K. The tangent lines (red) determine the anchor values. (b) Extracted anchor values as a function of the degree of the fit polynomial. The red point indicates the optimum determined from the AIC in Eq. (42). The error bars correspond to statistical uncertainties only.

Fig. 10(b), where the red dot is the one corresponding to the AIC optimum. Full results from the R-fit method are summarized in Fig. 11 below.

E. Summary of results

We now summarize and compare results obtained from the anchor bin, L-fit, and R-fit methods. In Fig. 11, we show the extracted reducibility factors and their statistical uncertainties for the six fixed substructure observables and the four machine-learned observables. We also checked that the same statistical uncertainties could be obtained through the bootstrap procedure [80] applied to 10k pseudoexperiments.

For a fixed observable, we see that the three methods yield similar results with comparable uncertainties on the reducibility factors. That said, this conclusion is sensitive to our choice of binning, and the agreement can be significantly worse with different binning schemes. The L-fit method is robust to making the bins finer, but the results degrade substantially with coarser bins. For the anchor bin method, the results degrade if the bins are either too fine or too coarse. So while one can adjust the binning in the L-fit and anchor bin methods to get sensible results, the R-fit method works well out of the box, as long as the bins are fine enough to yield a reasonably smooth ROC curve. Thus, our recommended method is the R-fit method, since it yields results that almost entirely independent of the binning choice.

As expected, the reducibility factors depend sensitively on the choice of observable used for jet topic modeling. In Fig. 11(a), we see that the better the observable is for quark/gluon discrimination, the larger the difference is between the two extracted reducibility factors. Among the fixed observables, the best quark/gluon discriminant is constituent multiplicity, so it is encouraging that the reducibility factors match the expectations from Eq. (31). Mass is a poor quark/gluon discriminant, so, not surprisingly, it yields relatively poor results for jet topics.

In Fig. 11(b), we show the results from the machine-learned observables. Here, the uncertainties include both the statistical variance within each training run and the variance from 10 different training runs, which are added in quadrature. The PFNs, which can in principle exploit all of the information available in the jet, yield excellent separation power, for both the classification and regression strategies. The EFNs, which restrict their attention to IRC-safe information, yield worse separation power, in agreement with the expectations from Ref. [1]. Note that the PFNs and EFNs agree on the value of κ_{21}, which is a general expectation from the fact that quarks are irreducible with respect to gluons (but not vice versa), even when considering just IRC-safe information [2,27].

For the results in Sec. V, we focus on PFN classification with the R-fit method as our benchmark approach for extracting reducibility factors. See Fig. 18 in the Appendix for reducibility factors extracted from the PYTHIA samples.

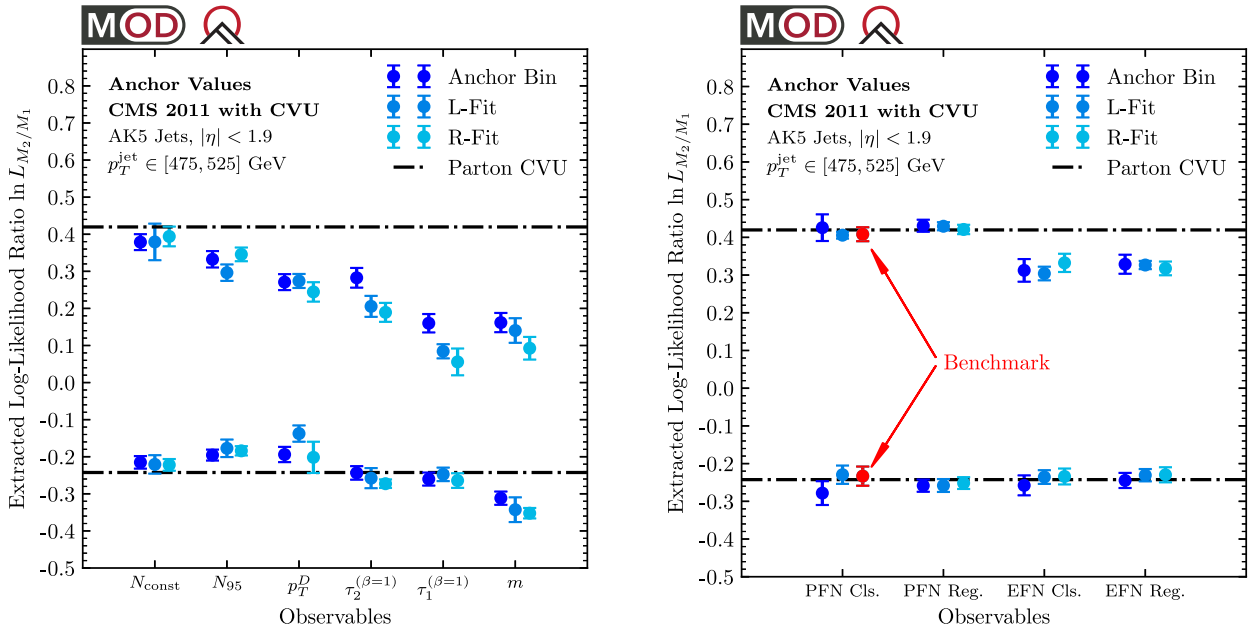

FIG. 11. Comparing the extracted anchor values between the anchor bin method (blue), L-fit method (azure), and R-fit method (cyan). Shown are results for the (a) six substructure and (b) four machine-learned observables from Sec. III B, applied to the CMS 2011 open data with central value unfolding. Our benchmark method using the R-fit method with PFN classification is indicated in red. The error bars correspond to statistical uncertainties combined with the AIC uncertainty in quadrature.

V. QUARK AND GLUON FRACTIONS AND DISTRIBUTIONS

With reducibility factors in hand, we can now isolate separate "quark" and "gluon" properties from the 2011 CMS open data. The reason for the quotes is that our definition of jet flavor is based on the operational procedure described in Sec. IV, which may differ from other definitions of jet flavor that have been proposed (see Ref. [28] for a review).

A. Quark and gluon fractions

We first need to convert reducibility factors into quark fractions of the forward and central mixtures. This conversion is straightforward using Eq. (9), and we use standard Gaussian error propagation to translate the uncertainties on the reducibility factors into uncertainties on the quark fractions. The results are shown in Fig. 12 for the four best discriminants: constituent multiplicity, image activity (N_{95}), PFN classification, and PFN regression. Interestingly, even though the reducibility factors for N_{95} in Fig. 11(a) do not match the CVU expectations from Eq. (31), the quark fractions approximately do, within uncertainties.

In Fig. 13, we show the statistical correlations between extracted f_1 and f_2 values for selected methods. By construction, the anchor bin method assumes no correlation between the anchor bin values, so the tilt of the ellipses is driven entirely by the correlations in Eq. (9). The L-Fit and R-fit methods have a nontrivial correlation in the anchor bin values, but this correlation is relatively small compared to those induced by the $\{\kappa_{12}, \kappa_{21}\} \to \{f_1, f_2\}$ mapping. See Figs. 19 and 20 in the Appendix for the quark factors extracted from the PYTHIA samples.

With all of the methods in Fig. 12 showing similar behavior, we choose the R-fit method with the PFN classifier as our benchmark. As already discussed, the R-fit method is the most statistically robust, and the PFN method has somewhat more conservative uncertainties, due to fluctuations from the different machine learning training runs. The extracted fractions are

$$\text{CMS 2011 CVU Jet Topics}: \begin{matrix} f_1 \simeq 0.708 \pm 0.025 \\ f_2 \simeq 0.561 \pm 0.033 \end{matrix}. \tag{55}$$

We emphasize that these uncertainties do not include any experimental systematics nor uncertainties from the unfolding.

B. Individual substructure distributions

Given the quark fractions, we can use Eqs. (5) and (6) to extract the "quark" and "gluon" distributions for any observable. We emphasize that we do not need to use the same observable for extracting fractions and for deriving distributions. We use standard error propagation to compute the uncertainties on individual flavor

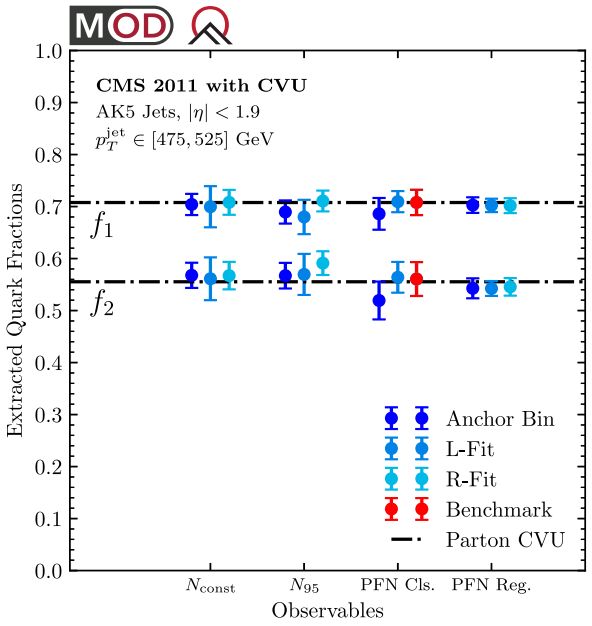

FIG. 12. Extracted quark fractions from the CMS 2011 open data with central value unfolding in the forward (M_1) and central (M_2) samples. We focus on the four observables with the best quark/gluon classification performance and compare the anchor bin (blue), L-fit (azure), and R-fit (cyan) methods. The benchmark method is indicated in red.

distributions, which depends on the uncertainties on the quark fractions from Eq. (55) and the statistical uncertainties on the binned mixture distributions.

We show jet topics results on the CMS 2011 CVU sample for six substructure distributions in Fig. 14. These are compared against distributions from parton truth labels in PYTHIA. The constituent multiplicity distributions are remarkably similar between the CMS 2011 CVU jet topics analysis and PYTHIA. For image activity, the gluon jets look similar, but our extracted quark jets exhibit more radiation. For momentum dispersion, the quark and gluon jets change in opposite directions. For 2-subjettiness, both types of jets show a suppression of the Sudakov peak. For 1-subjettiness and jet mass, both types of jets exhibit more radiation. We conclude that a full analysis, with unfolding uncertainties, has the potential to tell us detailed information about modeling quark and gluon jets in parton showers.

In the Appendix, we repeat the jet topics procedure on PYTHIA directly. Because jet topics is an operational procedure, there is no guarantee that the resulting distributions will match those from parton truth labels. Indeed, we find that the extracted quark fraction from the PYTHIA samples is systematically lower than the estimate from parton truth labels.

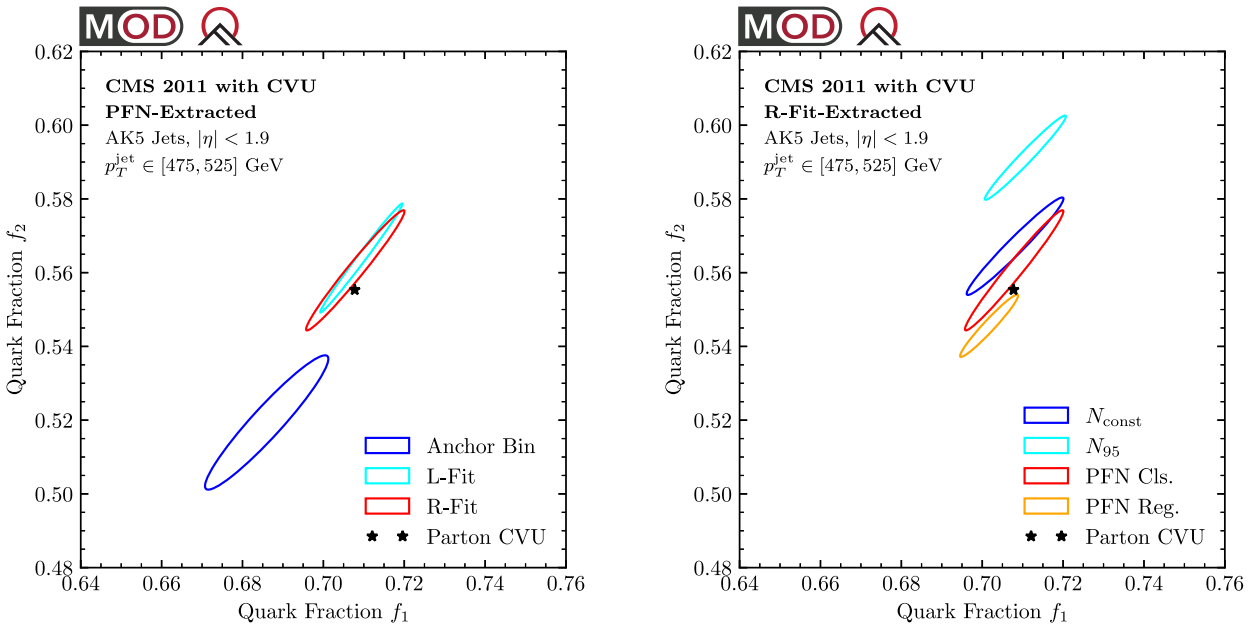

FIG. 13. Statistical correlations for the extracted quark fractions by (a) fixing the PFN observable and sweeping over the reducibility factor method and (b) fixing the R-fit method and sweeping over the four observables in Fig. 12. The diagonal alignment of the ellipses is primarily due to the correlations induced from the $\{\kappa_{12}, \kappa_{21}\} \to \{f_1, f_2\}$ mapping in Eq. (9). For the benchmark method in red, the extracted fractions match the CVU expectation in Eq. (31) within the statistical uncertainties (combined with AIC).

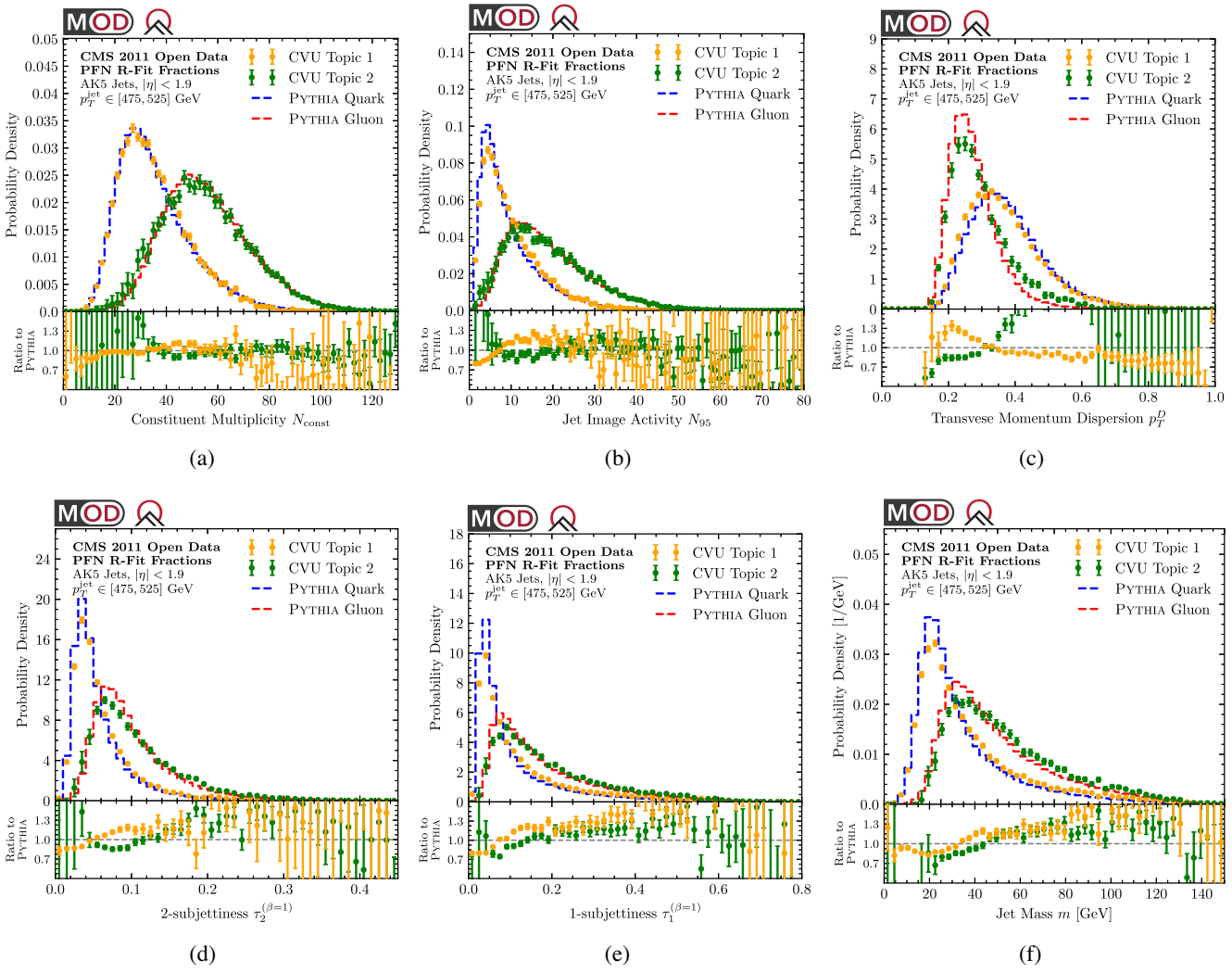

FIG. 14. Extracted "quark" (yellow, topic 1) and "gluon" (green, topic 2) distributions for the six substructure observables from Sec. III B. These are obtained by mixing the CMS 2011 open data CVU distributions from Fig. 6 according the quark fractions from Eq. (55). For comparison, we show the same PYTHIA truth-parton-labeled quark (blue dashed) and gluon (red dashed) distributions from Fig. 6. The error bars correspond to statistical uncertainties on the distributions combined with the reducibility factor uncertainties in quadrature.

C. Tagging performance

From the individual "quark" and "gluon" substructure distributions, we can present the tagging performance of each observable in the form of ROC curves. Following Eqs. (5) and (6), the cumulative distributions are

$$P_q(\mathcal{O}) = \frac{P_{M_1}(\mathcal{O}) - \kappa_{12} P_{M_2}(\mathcal{O})}{1 - \kappa_{12}}, \tag{56}$$

$$P_g(\mathcal{O}) = \frac{P_{M_2}(\mathcal{O}) - \kappa_{21} P_{M_1}(\mathcal{O})}{1 - \kappa_{21}}. \tag{57}$$

When plotting uncertainties below, we ignore statistical correlations between the reducibility factors and the cumulative distributions. The ROC curve for each observable is defined by the set of points

$$\{P_q(\mathcal{O}_{\text{cut}}), P_g(\mathcal{O}_{\text{cut}})\}, \tag{58}$$

in analogy with Eq. (16). The statistical uncertainties in these ROC curves are dominated by those of the reducibility factors κ_{12} and κ_{21}.

In Fig. 15, we show the ROC curves for four representative discriminants: jet mass, 2-subjettiness, constituent multiplicity, and PFN classification. Recall from Sec. II D that the slopes of the ROC curve at its end points are related to the reducibility factors associated with the corresponding observable. In the case of jet mass, the reducibility factors at leading-logarithmic order are [2]

$$\kappa_{gq}^{\text{mass}} = 0, \qquad \kappa_{qg}^{\text{mass}} = \frac{C_q}{C_g} = \frac{4}{9}, \tag{59}$$

where in the last step, we used $C_g = C_A = 3$ and $C_q = C_F = 4/9$. This expected behavior is reflected in

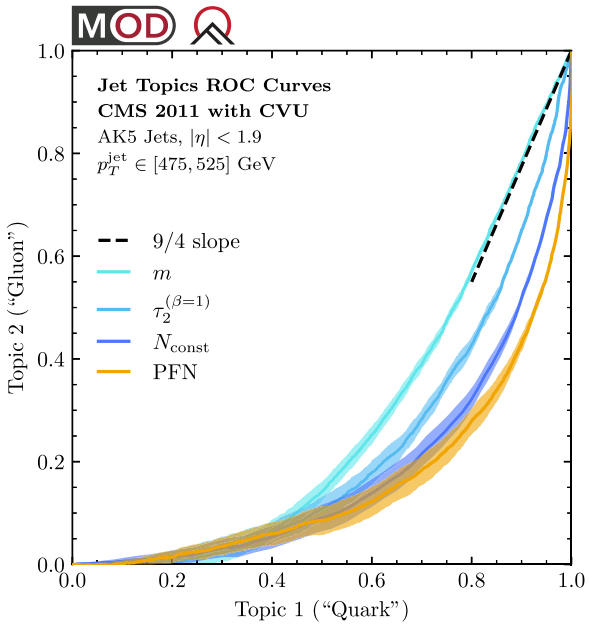

FIG. 15. Extracted ROC curves from the CMS 2011 open data. We consider four discriminants that span the range of quark/gluon tagging performance: jet mass, 2-subjettiness, constituent multiplicity, and PFN classification. The fact that the ROC curves slopes are close to 0 and ∞ at the end points imply that the jet topics are mutually irreducible, as desired. For jet mass, the dashed black slope corresponds to the leading-logarithmic prediction in Eq. (59). The error band corresponds to statistical uncertainties propagated through Eqs. (56) and (57).

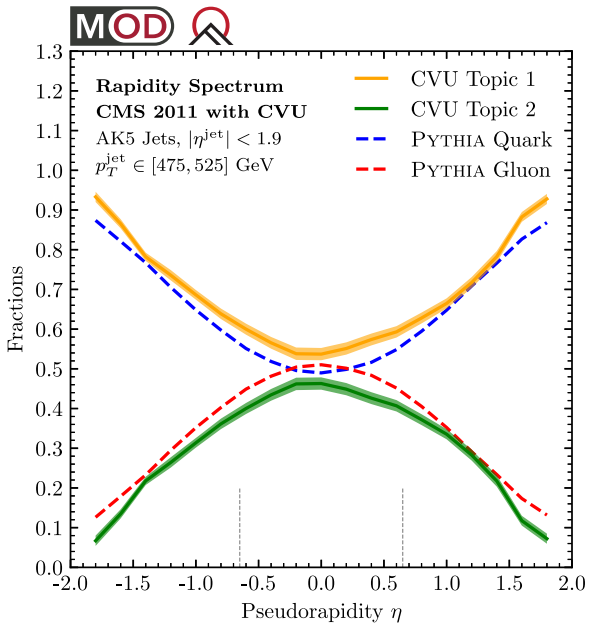

FIG. 16. Extracted quark and gluon fractions as a function of pseudorapidity, comparing the topic extraction in the CMS 2011 open data to parton labels from PYTHIA. The vertical gray lines correspond to the forward/central boundaries at $|\eta| = 0.65$. Compared to the PYTHIA baseline, there is an overall increase in the "quark" content across the spectrum. The error band corresponds to statistical uncertainties in the fitted distribution.

the $1/\kappa_{qg}^{\text{mass}} \approx 9/4$ slope at the right end point of the jet mass ROC curve.

In the case of the PFN, we effectively have a self-calibrating classifier, since the PFN R-fit method was used to operationally define quarks and gluons. As expected, the slopes of the PFN ROC curve are approximately 0 and ∞ at the respective end points, since the samples are, by construction, mutually irreducible.

D. Rapidity spectrum

The analysis thus far has focused on two wide rapidity regions, forward and central separated by the boundary $|\eta| = 0.65$. With the quark and gluon topics in hand, we can revisit Eq. (25) and try to determine the rapidity spectrum for the quark and gluon jets separately. The idea, adapted from Ref. [2], is to bin the data into smaller rapidity slices, and fit the observable distribution in each slice to a linear combination of topics.

For this analysis, we use the PFN R-fit method to determine the jet topics and then fit linear combinations of the constituent multiplicity distribution. The resulting quark and gluon fractions in rapidity slices of size $\Delta\eta = 0.2$ are shown in Fig. 16. Here, the uncertainties are dominated by the those of the overall quark fractions. The slight $\eta \to -\eta$ asymmetry is within the expectation from statistical uncertainties. The kink in the rapidity spectrum around $|\eta| = 1.4$ mirrors that of Fig. 4(c), so it would be interesting to investigate potential detector mismodeling near that transition region.

E. Correlation dimension

As our final application of the quark fractions in Eq. (55), we compute separate correlation dimensions for the quark and gluon samples. The correlation dimension is a scale-dependent measure of the number of effective degrees of freedom describing a dataset [15–17], and it can be computed knowing just pairwise distances between data points, without needing explicit coordinates. This analysis was presented for a mixed sample of quark and gluon jets in Ref. [12], which contains more details about the analysis procedure.

As our pairwise distance measure between jets, we use the energy mover's distance (EMD), which quantifies the amount of "work" to rearrange one energy distribution into another [18]. It is based on the well-known earth mover's distance from computer vision [81–85], also known as the Wasserstein metric [86,87]. The EMD has units of energy, and angular distances are normalized such that they are measured in units of the jet radius. By computing all pairwise distances,

$$D_{k\ell} \equiv \text{EMD}(\mathcal{J}_k, \mathcal{J}_\ell), \tag{60}$$

one can triangulate the "space" of jets and derive a variety of geometric quantities [88]. See Refs. [89–92] for related EMD studies in particle physics.

One such geometric quantity is the correlation dimension [15–17] defined as

$$\dim(Q) = \frac{\partial \ln N(Q)}{\partial \ln Q}, \tag{61}$$

where

$$N(Q) = \sum_{k\ell} w_k w_\ell \Theta[D_{k\ell} < Q] \tag{62}$$

is the "bicross section" of the jet pairs within an EMD of Q from each other, where we have accounted for jet weights w_i in the case of weighted distributions. For unweighted events (i.e., $w_i = 1$), the bicross section simply reduces to the number of jet pairs. For a uniform distribution in a compact domain of $\mathbb{R}^n$, we have $\dim(Q) = n$ as long as Q is small compared to the size of the domain.

To address the case of quark/gluon mixtures, we can modify Eq. (62) to handle the case of multiple samples:

$$N_{AB}(Q) = \sum_{k\in A,\ell\in B} w_k w_\ell \Theta[D_{k\ell} < Q]. \tag{63}$$

The correlation dimension derived from $N_{AB}(Q)$ has the interpretation of the dimensionality of the B manifold at a scale Q averaged over the A manifold, and is symmetric between A and B.

In practice, we choose a binning in Q and work with the number of jet pairs falling into each EMD bin. Concretely, let $\{Q_0, Q_1, \ldots, Q_S\}$ be the edges of S bins in Q, and

$$n_{AB}(s) = \sum_{k\in A,\ell\in B} w_k w_\ell \Theta[Q_{s-1} \le D_{k\ell} < Q_s] \tag{64}$$

be the number of jet pairs in bin $s = 1, \ldots, S$. Note that $n_{AB}(s) = n_{BA}(s)$ due to the symmetry of the EMD. We can relate this expression to Eq. (63) via

$$N_{AB}(Q_s) = \sum_{t=1}^{s} n_{AB}(t). \tag{65}$$

For two mixed samples M_1 and M_2 with known quark fractions f_1 and f_2, the three quantities:

$$n_{M_1M_1}(s), n_{M_1M_2}(s), n_{M_2M_2}(s), \tag{66}$$

can be inverted bin by bin to determine the quark/gluon values:

$$n_{qq}(s), n_{qg}(s), n_{gg}(s), \tag{67}$$

and subsequently the correlation dimension for the quark-quark, quark-gluon, and gluon-gluon samples, respectively, via Eqs. (65) and (61). The equations relating the quantities in Eq. (66) to those in Eq. (67) can be derived by ensuring bicross sectional consistency with the known fractions f_1 and f_2 and total cross section of M_1 and M_2:

$$\begin{bmatrix} \frac{n_{M_1M_1}(s)}{\sigma_1^2} \\ \frac{n_{M_1M_2}(s)}{\sigma_1\sigma_2} \\ \frac{n_{M_2M_2}(s)}{\sigma_2^2} \end{bmatrix} = \begin{pmatrix} f_1^2 & 2f_1(1-f_1) & (1-f_1)^2 \\ f_1f_2 & f_1(1-f_2)+f_2(1-f_1) & (1-f_1)(1-f_2) \\ f_2^2 & 2f_2(1-f_2) & (1-f_2)^2 \end{pmatrix} \begin{bmatrix} \frac{n_{qq}(s)}{(f_1\sigma_1+f_2\sigma_2)^2} \\ \frac{n_{qg}(s)}{(f_1\sigma_1+f_2\sigma_2)((1-f_1)\sigma_1+(1-f_2)\sigma_2)} \\ \frac{n_{gg}(s)}{((1-f_1)\sigma_1+(1-f_2)\sigma_2)^2} \end{bmatrix}, \tag{68}$$

where the total cross section of the mixed samples are

$$\sigma_1 = \sum_{i\in M_1} w_i, \qquad \sigma_2 = \sum_{i\in M_2} w_i. \tag{69}$$

One can easily solve Eq. (68) for $n_{qq}(s)$, $n_{gg}(s)$, and $n_{qg}(s)$ for each s given $n_{M_1M_1}(s)$, $n_{M_2M_2}(s)$, $n_{M_1M_2}(s)$, σ_1, σ_2, f_1, and f_2.

The results of this analysis are shown in Fig. 17(a), comparing the OMNIFOLD CVU results to those obtained from the parton-level labels in PYTHIA. Here, we have restricted our attention to a narrow jet range of $p_T \in [495, 505]$ GeV, and we rescale the jet constituents such that the jet p_T is always 500 GeV. At the scale $Q = 250$ GeV, the correlation dimensional is zero, since at this resolution scale ($Q \approx p_T/2$), a jet looks like a single parton. As the Q scale decreases, the dimensionality increases logarithmically as one resolves more features within the jet, with higher dimensionalities associated with the gluon samples compared to the quark ones.

We can gain additional insight into the correlation dimension by performing a first-principles calculation of this quantity in the strongly ordered limit. For simplicity of presentation, we ignore running of the strong coupling constant α_s. Consider the EMD between two jets $i \in I$ and $j \in J$ with color factors C_I and C_J. In the strongly ordered limit, we need only consider the case where one jet has emitted a soft-collinear gluon, whereas the other jet still looks like a single parton. In that case, the EMD reduces to the $\beta = 1$ jet angularity [93–95] between one of the jets and the jet axis. Considering an ensemble of such jet pairs, the cumulative distribution for the EMD is

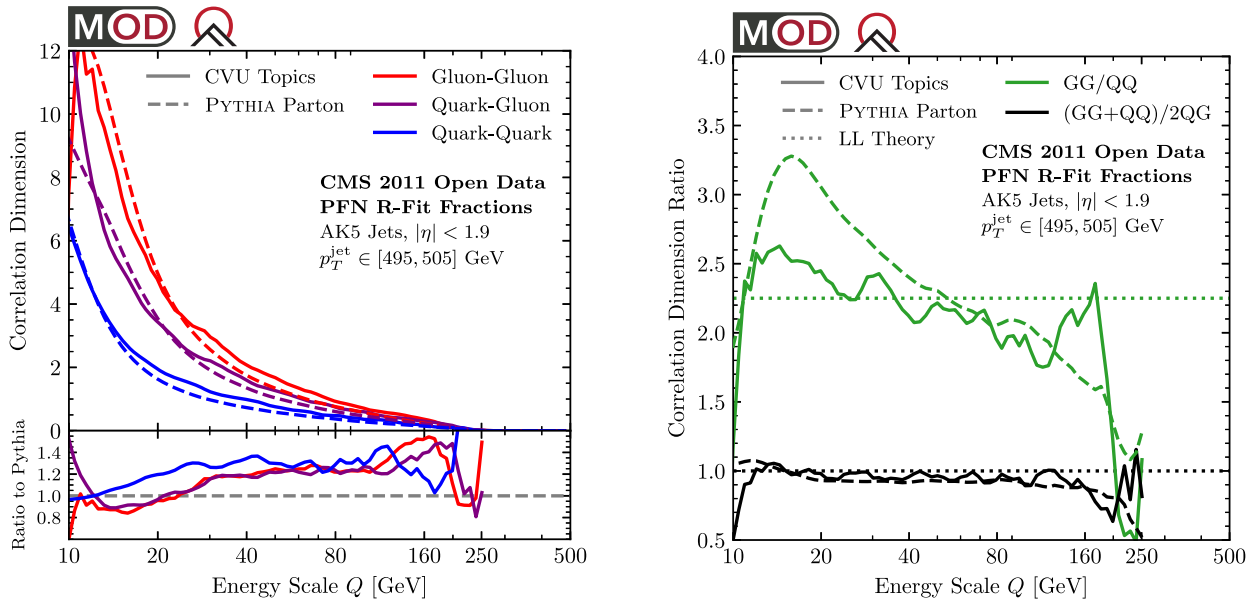

FIG. 17. (a) Correlation dimensions for three different types of jet pairs: gluon-gluon, quark-gluon, and quark-quark. The PYTHIA baseline uses parton-level labels. For the CMS open data result with CVU, the quark fractions are given by Eq. (55). These fractions are used in Eq. (68) to unmix the bicross sections of the central and forward jet combinations into quark and gluon combinations. (b) Ratios of the correlation dimension from Eqs. (72) and (73), compared to first-principles QCD calculations in the strongly ordered limit.

$$N_{IJ}(Q) = \exp\left[-\frac{\alpha_s(C_I + C_J)}{\pi}\log^2\frac{p_T/2}{Q}\right]. \quad (70)$$

Here, the exponential is the Sudakov factor for *neither* of the two jets to have an emission above the scale Q (i.e., the product of the Sudakov factors for the individual jets). The factor of $p_T/2$ in the logarithm is because this is approximately the maximum EMD achievable between two anti-k_t jets.

Taking the logarithmic derivative in Eq. (61), the (mixed) correlation dimension is

$$\dim_{IJ}(Q) = \frac{2\alpha_s(C_I + C_J)}{\pi}\log\frac{p_T/2}{Q}. \quad (71)$$

From this, we can predict that the ratio between the gluon-gluon and quark-quark correlation dimensions is

$$\frac{\dim_{gg}(Q)}{\dim_{qq}(Q)} = \frac{C_g}{C_q} = \frac{9}{4}, \quad (72)$$

where we have used $C_g = C_A = 3$ and $C_q = C_F = 4/3$. Intriguingly, this means that the correlation dimensions exhibit Casimir scaling. We can also predict that the correlation dimensions satisfy the following relation:

$$\frac{\dim_{gg}(Q) + \dim_{qq}(Q)}{2\dim_{qg}(Q)} = \frac{2C_g + 2C_q}{2(C_g + C_q)} = 1. \quad (73)$$

The two analytic predictions in Eqs. (72) and (73) are plotted in Fig. 17(b), along with the CMS open data results with CVU. We see quite good agreement, especially given the relatively simplicity of the strongly ordered analysis. Interestingly, the tilt in these predictions, which are minimally expected from running coupling effects and subleading terms in the splitting function, are smaller in the CMS open data than in PYTHIA. We hope that this encourages future precision calculations of the correlation dimensions of quark, gluon, and other types of jets.

VI. CONCLUSIONS

In this paper, we disentangled "quark" and "gluon" jet substructure distributions in the CMS 2011 open data using the operational approach advocated in Ref. [1]. We improved the statistical robustness of jet topics by introducing two new methods to extract reducibility factors: the L-fit method based on the log-likelihood ratio and the R-fit method based on the ROC curve. In order to mitigate sample dependence from inhomogeneities in the CMS detector response, we implemented central value unfolding with OMNIFOLD. We found noticeable differences in both the quark/gluon fractions and the quark/gluon distributions compared to the PYTHIA baseline, which would be interesting to investigate further in future work.

There are a number of directions to extend the proof-of-concept study presented here. At minimum, one could apply our improved jet topics procedures to the unfolded ATLAS measurements from Ref. [3] and the unfolded CMS measurements from Ref. [72]. Binned distributions from these studies are available at HEPDATA [96,97], which is

sufficient for the anchor bin method already implemented by ATLAS. Note that Ref. [97] and v2 of Ref. [96] have covariance matrices, which is essential for this kind of analysis. To make the most of the L-fit or R-fit methods, though, it would be beneficial to have results with finer binning (or unbinned [75]). Similarly, it would be exciting to go beyond central value unfolding and account for systematic uncertainties in the context of OMNIFOLD. This would require a more detailed understanding of the CMS detector response to make sure that the unfolding does not erroneously adjust the truth distributions to account for potential detector mismodeling. Finally, one could apply jet topics to observables designed to isolate pure quark and gluon samples [6].

Our analysis focused on a narrow jet p_T range, but with more p_T slices, one could study the evolution of the quark and gluon distributions as a function of energy scale. This would be particularly interesting in the context of the correlation dimension analysis, to test the strongly ordered prediction in Eq. (71) that the dimension depends on the ratio of Q/p_T. Similarly, it would be interesting to augment this two-sample analysis with more jet samples, either to test the prediction of sample independence (see Ref. [1] for a quantification strategy) or to try to identify additional jet categories (e.g. up-type versus down-type quarks). For the latter, one would want to apply machine learning techniques sensitive to the charge and flavor of the jet constituents, such as the PFN-ID approach [14]. We look forward to future developments of the jet topics method, to improve the (operational) connection between long-distance measurements and short-distance QCD dynamics.

ACKNOWLEDGMENTS

We thank Jasmine Brewer, Matt LeBlanc, Eric Metodiev, Benjamin Nachman, and Andrew Turner for helpful feedback. We thank Radha Mastandrea, Eric Metodiev, and Preksha Naik for collaboration in the initial stages of this project. We thank CERN, the CMS collaboration, and the CMS Data Preservation and Open Access (DPOA) team for making research-grade collider data available to the public. We thank the letters Δ and O for pointing out alternate meanings of N_{95} in the public health literature. This work was supported by the Office of Nuclear Physics of the U.S. Department of Energy (DOE) under Grant No. DE-SC0011090, by the DOE Office of High Energy Physics under Grants No. DE-SC0012567 and No. DE-SC0019128, and by the National Science Foundation under Cooperative Agreement Grant No. PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/).

APPENDIX: COMPARISON TO MONTE CARLO RESULTS

In this appendix, we repeat the analyses from Secs. IV and V, but applied to the PYTHIA Monte Carlo samples. This is a cross-check of our new jet topics procedures on a sample without any unfolding complications.

In Fig. 18, we show results for the extraction of the anchor values. Like in Fig. 11, we see that the better quark/gluon discriminants yield more separation in the anchor values. Unlike in Fig. 11, the smaller anchor value is shifted

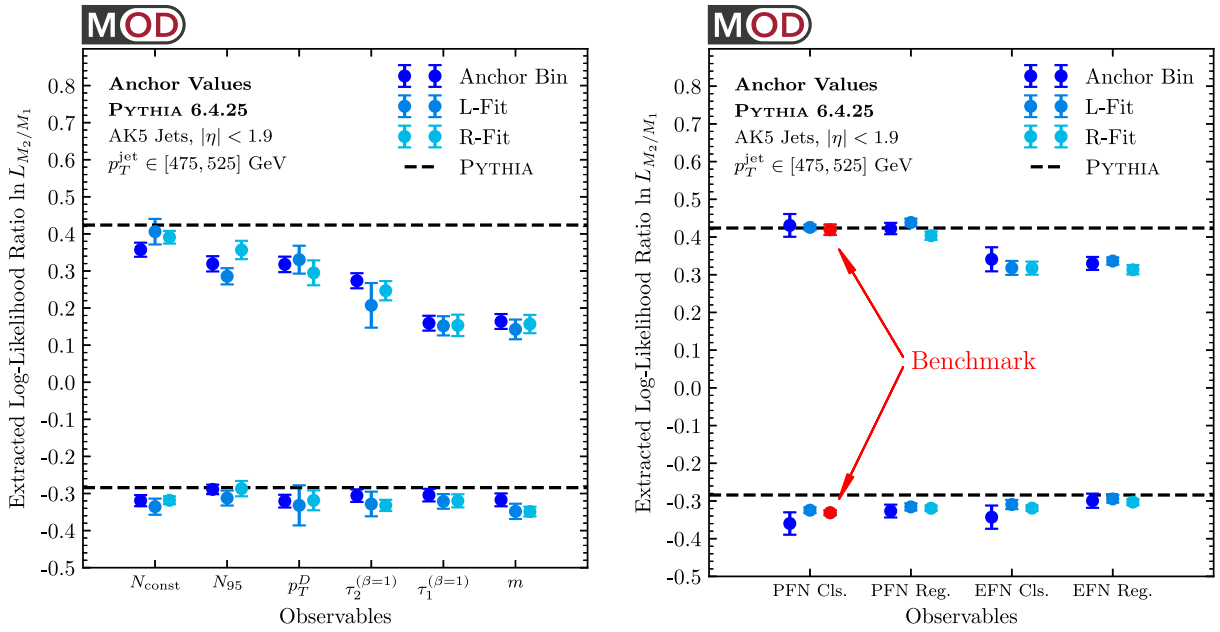

FIG. 18. Same as Fig. 11 but applied to the PYTHIA 6.4.25 dataset. Interestingly, the extracted anchor values for the benchmark method in red does not match the horizontal dashed PYTHIA parton-level expectation from Eq. (30).

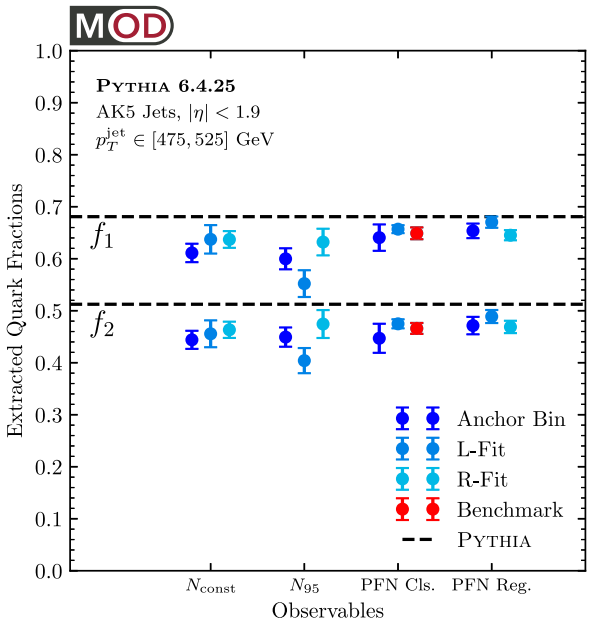

FIG. 19. Same as Fig. 12 but applied to the PYTHIA 6.4.25 dataset. Interestingly, the extracted jet topics are more gluon enriched compared to the expectation from Eq. (30).

systematically downward compared to the expectation from PYTHIA parton labels in Eq. (30).

This downward shift is reflected in Figs. 19 and 20. Here, the quark fractions extracted from constituent multiplicity, image activity, PFN classification, and PFN regression are consistent with each other but smaller than the fraction of quark-parton-labeled jets in PYTHIA. Because the jet topics procedure acts at the level of distributions, we cannot identify which parton-level quark jets are being categorized as topics-level gluon jets. One way this mismatch can happen is if a quark parton in PYTHIA arises from a gluon splitting via $g \to q\bar{q}$. Because the $g \to q\bar{q}$ splitting function is universal in the collinear limit, it would be "correct" to assign quarks from this splitting process to the gluon sample, though we were not able to find direct evidence that this is actually happening. Interestingly, this mismatch did not show up in the CMS open data result in Fig. 12, so it is also possible that there is a more subtle issue with how PYTHIA generates phase space configurations.

Like in the main text, we use the R-fit method with PFN classification as our default strategy to extract quark fractions:

$$\text{PYTHIA jet topics}: \begin{matrix} f_1 \simeq 0.649 \pm 0.012 \\ f_2 \simeq 0.466 \pm 0.011 \end{matrix}. \tag{A1}$$

In Fig. 21, we show the separate "quark" and "gluon" distributions using these fractions. The difference between jet topics and parton-truth-labeled samples is most apparent in the IRC-unsafe observables, where the extracted gluon distributions are more quarklike than the partonic expectation. Interestingly, the mismatch is much less evident, though still present, in the IRC-safe observables. This motivates further studies with additional samples taken from different regions of phase space, to see if these quark/gluon trends persist.

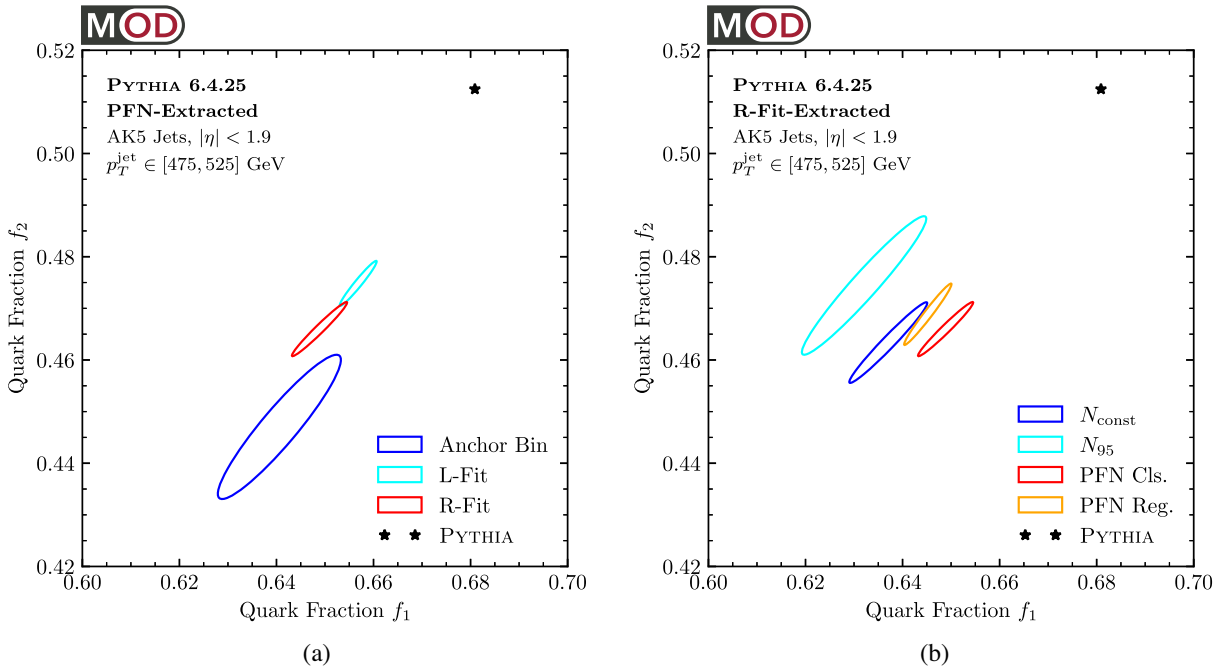

FIG. 20. Same as Fig. 13 but applied to the PYTHIA 6.4.25 dataset. Interestingly, the ellipses are offset from, but diagonally aligned with, the parton label expectation from Eq. (30).

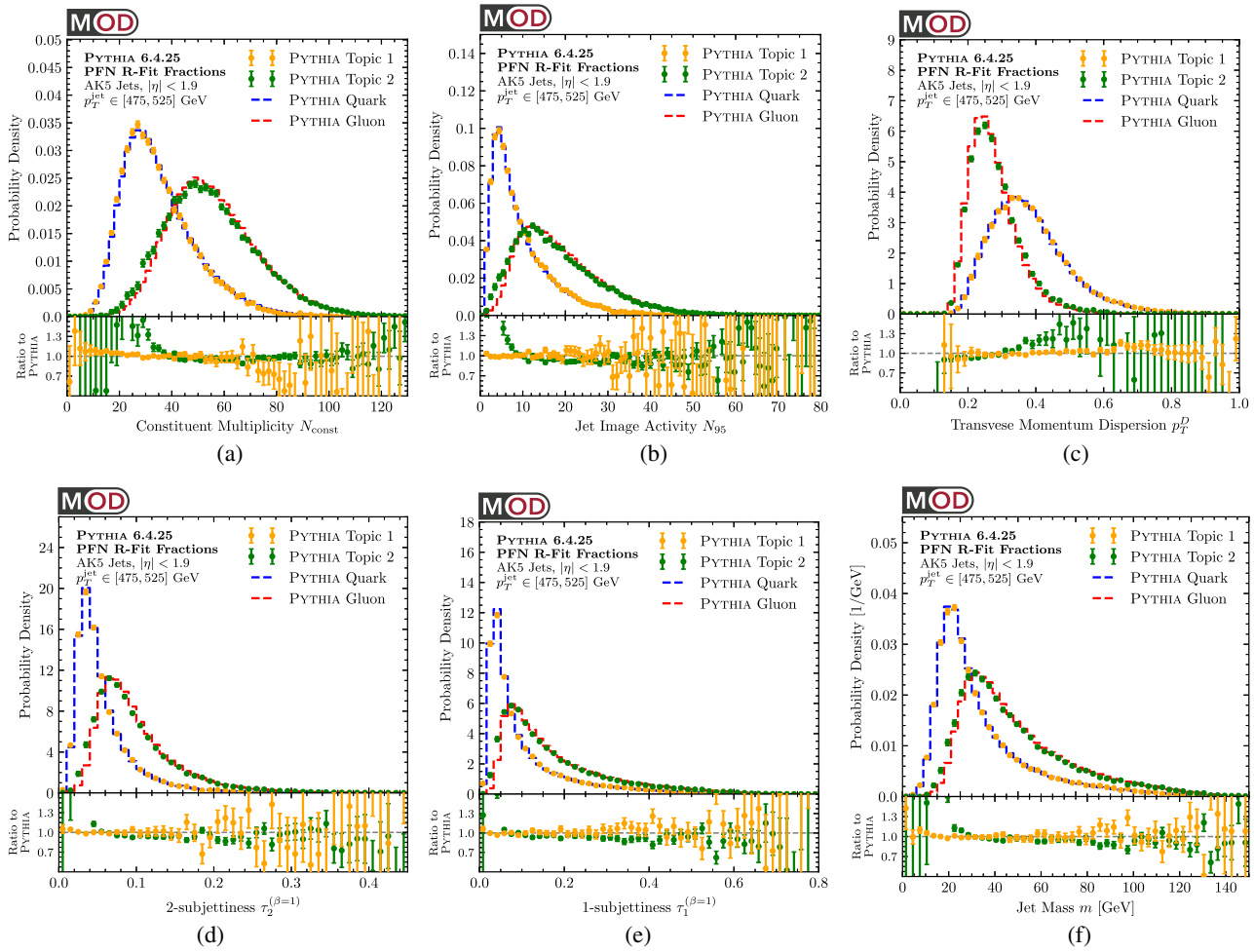

FIG. 21. Same as Fig. 14 but applied to the PYTHIA 6.4.25 dataset. Compared to the truth-parton-labeled samples, the extracted "gluon" distribution (green) is more quark like.

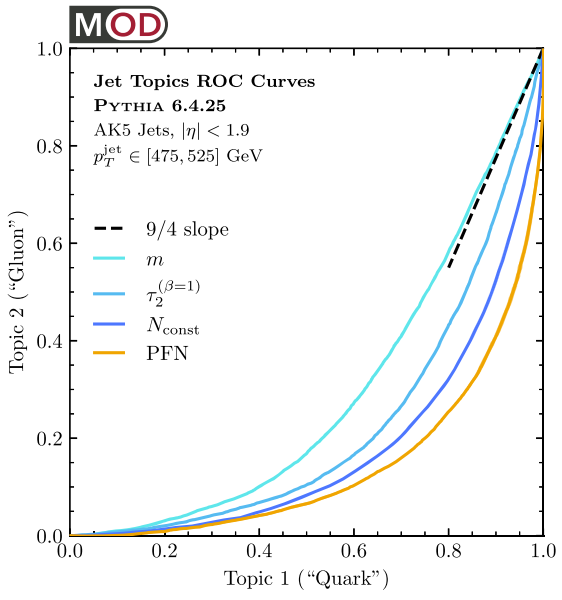

FIG. 22. Same as Fig. 15 but applied to the PYTHIA 6.4.25 dataset.

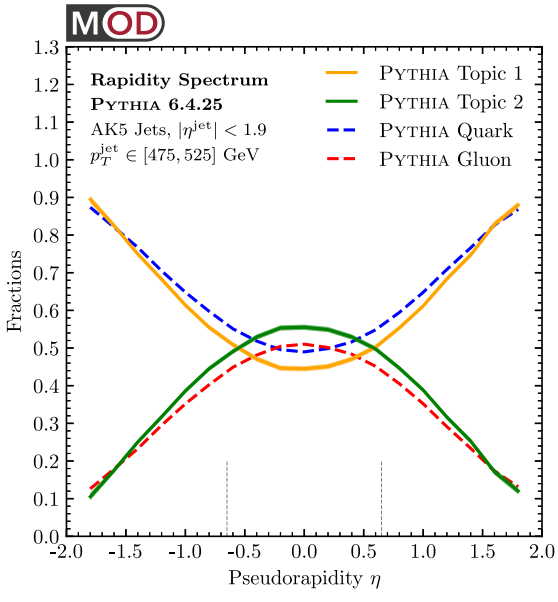

FIG. 23. Same as Fig. 16 but applied to the PYTHIA 6.4.25 dataset. Here, the topics extracted from PYTHIA are more gluon-like compared to the parton labels.

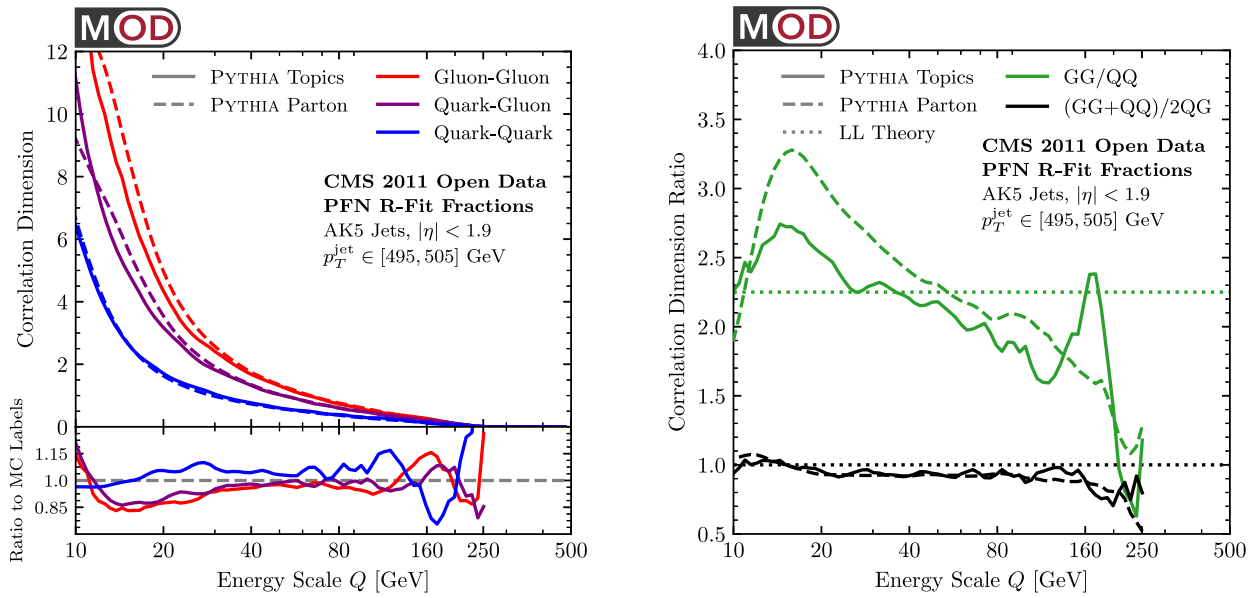

FIG. 24. Same as Fig. 17 but applied to the PYTHIA 6.4.25 dataset.

We show plots of the extracted ROC curve in Fig. 22, which agree qualitatively with the results from the CMS open data. We show plots of the rapidity spectrum and correlation dimension in Figs. 23 and 24, respectively. The difference between the topics-labeled curves and the parton-labeled curve highlights the ambiguity of quark and gluon labels beyond leading-logarithmic accuracy, motivating further studies using operational definitions that are well defined in data.

[1] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler, An operational definition of quark and gluon jets, J. High Energy Phys. 11 (2018) 059.

[2] Eric M. Metodiev and Jesse Thaler, Jet Topics: Disentangling Quarks and Gluons at Colliders, Phys. Rev. Lett. **120**, 241602 (2018).

[3] Georges Aad *et al.* (ATLAS Collaboration), Properties of jet fragmentation using charged particles measured with the ATLAS detector in pp collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. D **100**, 052011 (2019).

[4] Radha Mastandrea, Analyzing CMS open collider data through topic modeling, S. B. thesis, Massachusetts Institute of Technology, 2019, https://hdl.handle.net/1721.1/123389.

[5] Jasmine Brewer, Jesse Thaler, and Andrew P. Turner, Data-driven quark and gluon jet modification in heavy-ion collisions, Phys. Rev. C **103**, L021901 (2021).

[6] Iain W. Stewart and Xiaojun Yao, Pure quark and gluon observables in collinear drop, J. High Energy Phys. 09 (2022) 120.

[7] Yueyang Ying, Jasmine Brewer, Yi Chen, and Yen-Jie Lee, Data-driven extraction of the substructure of quark and gluon jets in proton-proton and heavy-ion collisions, arXiv:2204.00641.

[8] Barry M. Dillon, Darius A. Faroughy, and Jernej F. Kamenik, Uncovering latent jet substructure, Phys. Rev. D **100**, 056002 (2019).

[9] Ezequiel Alvarez, Federico Lamagna, and Manuel Szewc, Topic model for four-top at the LHC, J. High Energy Phys. 01 (2020) 049.

[10] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc, Learning the latent structure of collider events, J. High Energy Phys. 10 (2020) 206.

[11] CMS Collaboration, Jet primary dataset in AOD format from RunA of 2011 (/Jet/Run2011A-12Oct2013-v1/AOD), CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.UP77.P6PQ.

[12] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, Exploring the space of jets with CMS open data, Phys. Rev. D **101**, 034009 (2020).

[13] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Open Data | Jet Primary Dataset | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3340205.

[14] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler, Energy flow networks: Deep sets for particle jets, J. High Energy Phys. 01 (2019) 121.

[15] Peter Grassberger and Itamar Procaccia, Characterization of Strange Attractors, Phys. Rev. Lett. **50**, 346 (1983).
[16] Francesco Camastra, Data dimensionality estimation methods: A survey, Pattern Recognit. **36**, 2945 (2003).
[17] Balázs Kégl, Intrinsic dimension estimation using packing numbers, in *Advances in Neural Information Processing Systems*, edited by S. Becker, S. Thrun, and K. Obermayer (MIT Press, Cambridge, MA, 2002), Vol. 15.
[18] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler, Metric Space of Collider Events, Phys. Rev. Lett. **123**, 041801 (2019).
[19] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands, PYTHIA 6.4 Physics and Manual, J. High Energy Phys. 05 (2006) 026.
[20] S. Chatrchyan *et al.* (CMS Collaboration), The CMS Experiment at the CERN LHC, J. Instrum. **3**, S08004 (2008).
[21] Anders Andreassen, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler, OmniFold: A Method to Simultaneously Unfold All Observables, Phys. Rev. Lett. **124**, 182001 (2020).
[22] Anders Andreassen, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, Adi Suresh, and Jesse Thaler, Scaffolding Simulations with Deep Learning for High-dimensional Deconvolution, in *9th International Conference on Learning Representations* (2021), arXiv:2105.04448.
[23] Patrick T. Komiske, Serhii Kryhin, and Jesse Thaler, OmniFold Weights | CMS 2011A Open Data | Jet Primary Dataset | pT 375–700 GeV, Zenodo (2022), 10.5281/zenodo.6519307.
[24] Julian Katz-Samuels, Gilles Blanchard, and Clayton Scott, Decontamination of mutual contamination models, arXiv:1710.01167.
[25] Georges 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).
[26] Samuel Bright-Thonney and Benjamin Nachman, Investigating the topology dependence of quark and gluon jets, J. High Energy Phys. 03 (2019) 098.
[27] Andrew J. Larkoski and Eric M. Metodiev, A theory of quark vs. gluon discrimination, J. High Energy Phys. 10 (2019) 014.
[28] Philippe Gras, Stefan Höche, Deepak Kar, Andrew Larkoski, Leif Lönnblad, Simon Plätzer, Andrzej Siódmok, Peter Skands, Gregory Soyez, and Jesse Thaler, Systematics of quark/gluon tagging, J. High Energy Phys. 07 (2017) 091.
[29] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.
[30] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola, Deep sets, arXiv:1703.06114.
[31] CMS releases new batch of research data from LHC (2016), http://opendata.cern.ch/docs/cms-releases-new-batch-of-research-data-from-lhc.
[32] Vardan Khachatryan *et al.* (CMS Collaboration), The CMS trigger system, J. Instrum. **12**, P01020 (2017).
[33] Particle-flow event reconstruction in CMS and performance for jets, taus, and MET, Technical Report No. CMS-PAS-PFT-09-001, CERN, Geneva, 2009.
[34] Commissioning of the Particle-flow event reconstruction with the first LHC collisions recorded in the CMS detector, Technical Report No. CMS-PAS-PFT-10-001, 2010.
[35] A. M. Sirunyan *et al.* (CMS Collaboration), Particle-flow reconstruction and global event description with the CMS detector, J. Instrum. **12**, P10003 (2017).
[36] Patrick T. Komiske, Ian Moult, Jesse Thaler, and Hua Xing Zhu, Analyzing N-point energy correlators inside jets with CMS open data, arXiv:2201.07800.
[37] S. Agostinelli *et al.* (GEANT4 Collaboration), GEANT4: A simulation toolkit, Nucl. Instrum. Methods Phys. Res., Sect. A **506**, 250 (2003).
[38] Rick Field, Min-Bias and the underlying event at the LHC, Acta Phys. Pol. B **42**, 2631 (2011).
[39] CMS Collaboration, Simulated dataset QCD_Pt-170to300_TuneZ2_7 TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.WKRR.DCJP.
[40] CMS Collaboration, Simulated dataset QCD_Pt-300to470_TuneZ2_7 TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.X3XQ.USQR.
[41] CMS Collaboration, Simulated dataset QCD_Pt-470to600_TuneZ2_7 TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.BKTD.SGJX.
[42] CMS Collaboration, Simulated dataset QCD_Pt-600to800_TuneZ2_7 TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive)CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.EJT7.KSAY.
[43] CMS Collaboration, Simulated dataset QCD_Pt-800to1000_TuneZ2_7 TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.S3D5.KF2C.
[44] CMS Collaboration, Simulated dataset QCD_Pt-1000to1400_TuneZ2_7 TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.96U2.3YAH.
[45] CMS Collaboration, Simulated dataset QCD_Pt-1400to1800_TuneZ2_7 TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.RC9V.B5KX.
[46] CMS Collaboration, Simulated dataset QCD_Pt-1800_TuneZ2_7 TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal (2016), 10.7483/OPENDATA.CMS.CX2X.J3KW.
[47] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Simulation | Pythia 6 QCD 170-300 | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3341500.
[48] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Simulation | Pythia 6 QCD 300-470 | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3341498.
[49] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Simulation | Pythia 6 QCD 470-600 | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3341419.
[50] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Simulation |

Pythia 6 QCD 600-800 | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3364139.
[51] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Simulation | Pythia 6 QCD 800-1000 | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3341413.
[52] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Simulation | Pythia 6 QCD 1000-1400 | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3341502.
[53] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Simulation | Pythia 6 QCD 1400-1800 | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3341770.
[54] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler, CMS 2011A Simulation | Pythia 6 QCD 1800-inf | pT > 375 GeV | MOD HDF5 Format, Zenodo (2019), 10.5281/zenodo.3341772.
[55] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[56] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez, FastJet User Manual, Eur. Phys. J. C **72**, 1896 (2012).
[57] Vardan Khachatryan *et al.* (CMS Collaboration), Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV, J. Instrum. **12**, P02014 (2017).
[58] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez, The catchment area of jets, J. High Energy Phys. 04 (2008) 005.
[59] Jet performance in pp collisions at 7 TeV, Technical Report No. CMS-PAS-JME-10-003, 2010.
[60] CMS Collaboration, Determination of jet energy calibration and transverse momentum resolution in CMS, J. Instrum. **6**, 11002 (2011).
[61] Georges Aad *et al.* (ATLAS Collaboration), Measurement of jet charge in dijet events from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector, Phys. Rev. D **93**, 052003 (2016).
[62] Georges 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).
[63] Jon Pumplin, How to tell quark jets from gluon jets, Phys. Rev. D **44**, 2025 (1991).
[64] CMS Collaboration, Performance of quark/gluon discrimination in 8 TeV pp data, Technical Report No. CMS-PAS-JME-13-002, 2013.
[65] Jesse Thaler and Ken Van Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.
[66] Jesse Thaler and Ken Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness, J. High Energy Phys. 02 (2012) 093.
[67] Andrew J. Larkoski, Gavin P. Salam, and Jesse Thaler, Energy correlation functions for jet substructure, J. High Energy Phys. 06 (2013) 108.
[68] EnergyFlow Package (2017), https://energyflow.network/.
[69] Francois Chollet *et al.*, Keras (2015), https://github.com/fchollet/keras.
[70] Diederik P Kingma and Jimmy Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.
[71] Roman Kogler *et al.*, Jet substructure at the large hadron collider: Experimental review, Rev. Mod. Phys. **91**, 045003 (2019).
[72] Armen Tumasyan *et al.* (CMS Collaboration), Study of quark and gluon jet substructure in $Z +$ jet and dijet events from pp collisions, J. High Energy Phys. 01 (2022) 188.
[73] V. Andreev *et al.* (H1 Collaboration), Measurement of Lepton-Jet Correlation in Deep-Inelastic Scattering with the H1 Detector Using Machine Learning for Unfolding, Phys. Rev. Lett. **128**, 132002 (2022).
[74] H1 Collaboration, Multi-differential Jet Substructure Measurement in High Q^2 DIS Events with HERA-II Data, H1prelim-22-034 (2022), https://www-h1.desy.de/h1/www/publications/htmlsplit/H1prelim-22-034.long.html.
[75] Miguel Arratia *et al.*, Publishing unbinned differential cross section results, J. Instrum. **17**, P01024 (2022).
[76] G. D'Agostini, A multidimensional unfolding method based on Bayes' theorem, Nucl. Instrum. Methods Phys. Res., Sect. A **362**, 487 (1995).
[77] Pauli Virtanen *et al.* (SciPy 1.0 Contributors), SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).
[78] M. Williams, A novel approach to the bias-variance problem in bump hunting, J. Instrum. **12**, P09034 (2017).
[79] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control **19**, 716 (1974).
[80] Bradley Efron and Robert J. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability Vol. 57 (Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993).
[81] Shmuel Peleg, Michael Werman, and Hillel Rom, A unified approach to the change of resolution: Space and gray-level, IEEE Trans. Pattern Anal. Mach. Intell. **11**, 739 (1989).
[82] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, A metric for distributions with applications to image databases, in *Proceedings of the Sixth International Conference on Computer Vision, ICCV '98* (IEEE Computer Society, Washington, DC, USA, 1998), pp. 59–66.
[83] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vis. **40**, 99 (2000).
[84] Ofir Pele and Michael Werman, A linear time histogram metric for improved SIFT matching, in *Computer Vision—ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, 2008, Proceedings, Part III* (Springer Berlin, Heidelberg, 2008), pp. 495–508.
[85] Ofir Pele and Ben Taskar, The tangent earth mover's distance, in *Geometric Science of Information—First International Conference, GSI 2013, Paris, France, 2013. Proceedings* (Springer Berlin, Heidelberg, 2013), pp. 397–404.
[86] Leonid N Wasserstein, Markov processes over denumerable products of spaces describing large systems of automata, Probl. Inf. Transm. **5**, 47 (1969), http://mi.mathnet.ru/eng/ppi/v5/i3/p64.
[87] Roland L Dobrushin, Prescribing a system of random variables by conditional distributions, Theory Probab. Appl. **15**, 458 (1970).

[88] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler, The hidden geometry of particle collisions, J. High Energy Phys. 07 (2020) 006.

[89] Cari Cesarotti and Jesse Thaler, A robust measure of event isotropy at colliders, J. High Energy Phys. 08 (2020) 084.

[90] M. Crispim Romão, N. F. Castro, J. G. Milhano, R. Pedro, and T. Vale, Use of a generalized energy Mover's distance in the search for rare phenomena at colliders, Eur. Phys. J. C **81**, 192 (2021).

[91] Tianji Cai, Junyi Cheng, Nathaniel Craig, and Katy Craig, Linearized optimal transport for collider events, Phys. Rev. D **102**, 116019 (2020).

[92] Tianji Cai, Junyi Cheng, Katy Craig, and Nathaniel Craig, Which metric on the space of collider events?, Phys. Rev. D **105**, 076003 (2022).

[93] Stephen D. Ellis, Christopher K. Vermilion, Jonathan R. Walsh, Andrew Hornig, and Christopher Lee, Jet shapes and jet algorithms in SCET, J. High Energy Phys. 11 (2010) 101.

[94] Andrew J. Larkoski, Duff Neill, and Jesse Thaler, Jet shapes with the broadening axis, J. High Energy Phys. 04 (2014) 017.

[95] Andrew J. Larkoski, Jesse Thaler, and Wouter J. Waalewijn, Gaining (Mutual) information about quark/gluon discrimination, J. High Energy Phys. 11 (2014) 129.

[96] ATLAS Collaboration, Properties of jet fragmentation using charged particles measured with the ATLAS detector in pp collisions at $\sqrt{s} = 13$ TeV, HEPData (collection) (2021), 10.17182/hepdata.89321.

[97] CMS Collaboration, Study of quark and gluon jet substructure in $Z +$ jet and dijet events from pp collisions, HEPData (collection) (2021), 10.17182/hepdata.111308.