MITHIG-TH-22-001, CERN-TH-2022-057

Data-driven extraction of the substructure of quark and gluon jets in proton-proton and heavy-ion collisions

Yueyang Ying,[1] Jasmine Brewer,[2] Yi Chen,[1] and Yen-Jie Lee[1]

[1] *Relativistic Heavy Ion Group, MIT, Cambridge, Massachusetts, USA*
[2] *Theoretical Physics Department, CERN, CH-1211 Genève 23, Switzerland*

The different modifications of quark- and gluon-initiated jets in the quark-gluon plasma (QGP) produced in heavy-ion collisions is a long-standing question that has not yet received a definitive answer from experiments. In particular, the relative sizes of the modification of quark and gluon jets differ between theoretical models. Therefore, a fully data-driven technique is crucial for an unbiased extraction of the quark and gluon jet spectra and substructure. We perform a proof-of-concept study based on proton-proton and heavy-ion collision events from the Pyquen generator with statistics accessible in Run 4 of the Large Hadron Collider. We use a statistical technique called topic modeling to separate quark and gluon contributions to jet observables. We demonstrate that jet substructure observables, such as the jet shape and jet fragmentation function, can be extracted using this data-driven method. These values can then be used to obtain the modification of quark and gluon jet substructures in the QGP. We also perform the topic separation on smeared input data to demonstrate that the approach is robust to fluctuations arising from a QGP background. These results suggest the potential for an experimental determination of quark and gluon jet spectra and their substructure.

During the first millionth of a second after the Big Bang, the universe comprised hot and dense primordial matter of deconfined quarks and gluons before cooling down and forming ordinary matter. This deconfined phase of matter, the quark-gluon plasma (QGP), only exists at extremely high temperatures and pressures and is recreated on earth in high-energy heavy-ion collisions (see [1] for a review).

High-energy collisions between protons or nuclei occasionally produce very high-energy quarks or gluons that successively fragment and hadronize into collimated sprays of particles called jets. Jets are a ubiquitous tool for studying Quantum Chromodynamics (QCD) and have been widely used in the studies of both proton-proton [2] and heavy-ion collisions [3–6] at the Large Hadron Collider (LHC), the Relativistic Heavy Ion Collider (RHIC), and recently in electron-positron annihilation [7] with archived ALEPH data at the Large Electron-Position Collider [8].

The hot quark-gluon plasma produced in heavy-ion collisions modifies the properties of jets. High-energy partons propagating through the QGP lose energy due to multiple elastic scatterings and medium-induced gluon radiation [9–22], often referred to as jet quenching [23–25]. The resulting suppression of the yield of jets in heavy-ion collisions compared to an equivalent number of proton-proton collisions has been observed [26–32]. The structure of jets is also modified [33–39] in heavy-ion collisions and is an important tool for studying the properties of the quark-gluon plasma.

Quarks and gluons interact with the QCD medium proportional to their color charge, meaning that gluons interact more with the medium than quarks by a factor $C_A/C_F = 9/4$. However, a jet initiated by a quark or gluon quickly fragments into both quarks and gluons. Understanding the modification of quark- and gluon-initiated jets in the quark-gluon plasma may shed light on how the quark-gluon plasma resolves color structure within a jet (see e.g. [40–43]).

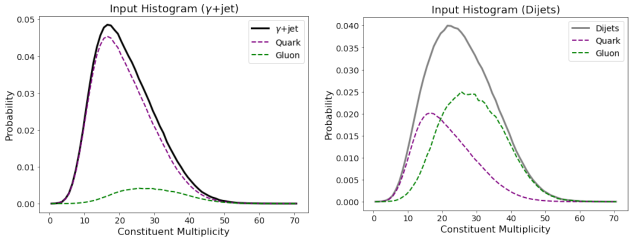
Experimentally, distinguishing quark and gluon jets is challenging since jet measurements are a combination of jets initiated by both. There has been extensive work on data-driven techniques for distinguishing quark and gluon jets in proton-proton [44–54] and heavy-ion collisions [40, 55, 56]. Especially in heavy-ion collisions with substantial theoretical uncertainties in Monte Carlo event generators, it is highly advantageous to distinguish quark and gluon jets in a way that does not rely on Monte Carlo labeling. In this case, we wish to identify two physics-motivated categories (quark- and gluon-initiated jets) underlying unlabeled jet measurements. Prior work has demonstrated success in using a statistical technique called topic modeling to distinguish quark- and gluon-initiated jets in both proton-proton [44, 45] and heavy-ion [55] collisions, using two measurable jet samples that differ in their quark- and gluon-initiated jet fractions.

In this work, we apply topic modeling to dijet and photon-jet (γ+jet) samples from Pyquen [57], which is a Monte Carlo event generator that simulates medium-induced energy loss of partons in heavy-ion collisions. In Section I we discuss the topic modeling approach, which relies on the assumption that dijet and γ+jet samples are mixtures of the same underlying quark- and gluon-initiated jet distributions, except with different quark and gluon fractions. In Section II we discuss the samples we use, and in Section III we use jet constituent multiplicity distributions to extract the quark and gluon fractions in these jet samples, separately in simulations of proton-proton and heavy-ion collisions. In Section IV we use these fractions to extract quark and gluon jet substructure observables in proton-proton and heavy-ion collisions. The quark and gluon jet substructure accessed using this data-driven method agree qualitatively with the substructure of quark- and gluon-initiated jets as defined from Monte Carlo-level information. This suggests the potential for an entirely data-driven procedure to experimentally extract quark and gluon jet substructure and their modification. In Section V we show that these results are robust to Gaussian smearing of the multiplicity distribution, suggesting the possibility of using this technique in the large background present in heavy-ion collisions.

I. TOPIC MODELING

To distinguish quark- and gluon-initiated jets experimentally, jet samples collected from events at RHIC or the LHC can be thought of as unlabeled mixtures of quark and gluon jets. Distinguishing quark and gluon jets from these unlabelled mixtures falls in the broad category of unsupervised learning, where the goal is to infer some structure or pattern in a set of unlabeled data. Following previous work [44, 55], we apply the "topic modeling" technique to solve this problem. Topic modeling is traditionally used to discover abstract "topics" that occur in a collection of text documents. In the context of jet topics, the two categories ("topics") underlying jet measurements are quark- and gluon-like jets [44].

To illustrate the concept of topic modeling, we present an example which we will refer to throughout the section. Suppose we have two input distributions, input A (pp γ+jet sample), and input B (pp dijet sample), as shown below in Fig. 1. In the case of quark/gluon topic modeling, we assume that these two input distributions are both a combination of the same two unknown base distributions, or "topics" (one quark-like and one gluon-like). Heuristically, this assumption is based on the intuition that jets can be classified as being initiated by a quark or gluon. These contributions to the example input distributions are shown as dashed curves in Fig. 1. Each input sample has a different fraction of each topic. The goal of the algorithm is to derive these underlying distributions from measurements.



(a) Input A (γ+jet sample)

(b) Input B (dijet sample)

FIG. 1: Example of two input distributions that might be used for topic modeling, demonstrating that each input is a superposition of the same underlying base distribution shapes

Jet samples collected from colliders are assumed to be mixtures of these jet topics, so that each jet observable histogram is a mixture of the two underlying quark/gluon base distributions. Mathematically, we can then represent the input histograms as

$$p^{(s)}(x) = f^{(s)} b_1(x) + (1 - f^{(s)}) b_2(x) \tag{1}$$

where $p^{(s)}(x)$ is the probability density of some observable x in sample s, $b_i(x)$ are the base distributions (topics), and $f^{(s)}$ and $1 - f^{(s)}$ are the fractions of topic 1 and 2 in sample s, respectively. However, Eq. 1 is ambiguous because there are infinitely many ways to define b_1 and b_2 and modify $f^{(s)}$ accordingly such that the equation remains true. In order to resolve this, we follow [44] and use the `DEMIX` algorithm [58], which breaks this ambiguity by choosing unique base distributions b_1 and b_2 that satisfy an additional requirement called mutual irreducibility.

`DEMIX` results in the *mutually irreducible* [59] underlying distributions, b_1 and b_2, that satisfy the requirement that neither contains any contribution from the other. In other words, we cannot write $b_1(x) = cb_2(x) + (1-c)F$, or vice versa, for any probability distribution F and $0 < c \leq 1$. This also implies that $\lim_{x \to x_{\max}} b_1(x)/b_2(x) = 0$ and $\lim_{x \to x_{\min}} b_2(x)/b_1(x) = 0$, where the probability distributions b_1, b_2 are defined on $x \in (x_{\min}, x_{\max})$ [1].

It is worth noting that `DEMIX` requires the input distributions to have different purities of the same underlying base distributions and guarantees that the two resulting base distributions are mutually irreducible. Since the Monte Carlo definition of quark and gluon jets is not well-defined, Ref. [45] defines the *operational definition* quark and gluon categories as the mutually irreducible underlying distributions in a jet substructure feature space, given two mixed QCD jet samples at a fixed p_T. For this work, we choose to use constituent multiplicity (number of constituent hadrons in a given jet) as the jet observable because the multiplicities of quark and gluon jets are mutually irreducible in the high-energy limit [52] and it exhibits good performance in proton–proton [44] and heavy-ion [55] studies.

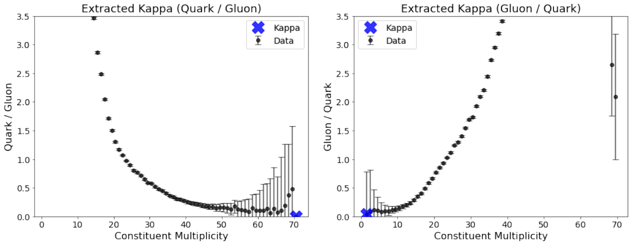
Extracting the base distributions requires finding the reducibility factor κ, which is the largest amount one distribution that can be subtracted from the other such that all bins remain non-negative,

$$\kappa_{ij} = \inf_x \frac{p^{(i)}(x)}{p^{(j)}(x)}\,, \tag{2}$$

where i, j index the samples.

It is worth noting that the mutual irreducibility of the base distributions is the assumption that κ_{qg} and κ_{gq} are zero. The extracted κ_{qg} and κ_{gq} are shown in Fig. 2, along with the ratio of each bin in the MC-level quark and gluon jet distributions of our example. The proximity to

[1] Depending on how we define $b_1(x)$ and $b_2(x)$, the limits may be reversed. That is, we may find $\lim_{x \to x_{\min}} b_1(x)/b_2(x) = 0$ and $\lim_{x \to x_{\max}} b_2(x)/b_1(x) = 0$



(a) Quark / Gluon (pp)

(b) Gluon / Quark (pp)

FIG. 2: The ratio of the quark and gluon truth histograms in our example. The extracted κ_{qg} and κ_{gq} are marked along the tail of the plot, which demonstrates that these base distributions are approximately mutually irreducible.

zero demonstrates the mutual irreducibility of the quark and gluon distributions as defined by the Monte Carlo.

With the mixture distributions p_A and p_B and the reducibility factors, the base distributions are given by

$$\begin{aligned} b_1(x) &= \frac{p_A(x) - \kappa_{AB} p_B(x)}{1 - \kappa_{AB}}, \\ b_2(x) &= \frac{p_B(x) - \kappa_{BA} p_A(x)}{1 - \kappa_{BA}} \end{aligned} \tag{3}$$

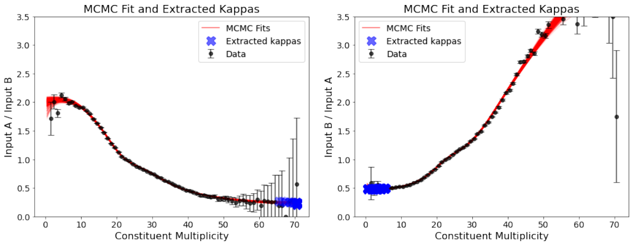
Following Ref. [55], we extract κ by fitting the sampled histograms p_A and p_B to a sum of skew-normal distributions, expressed as

$$f_N(x; \alpha_i, \theta) = \sum_{k=1}^{N} \alpha_{i,k} \mathrm{SN}(x; \mu_k, \sigma_k, s_k) \tag{4}$$

where $\mathrm{SN}(x; \mu_k, \sigma_k, s_k)$ represents a skew-normal distribution with parameters μ_k, σ_k, and s_k. While the mixture fractions α_i are unique to the input histograms, μ_k, σ_k, s_k are shared between the two. We use $N = 4$, such that we have 18 fit parameters[2], represented by α_A, α_B, and θ [55].

Assuming that the counts in the histograms follow a Poisson distribution, the best-fit parameters and the corresponding uncertainties can thus be captured by the Poisson-likelihood chi-square function [55, 60, 61]. In order to extract the parameter values and uncertainties from the likelihood function, we use Markov chain Monte Carlo (MCMC) [62]. We obtain initial estimates of the parameter values by running a simultaneous least-squares fit. For the results shown in this paper, we use 100 MCMC walkers, initialized using the least-squares parameters, and run for 35,000 samples using a burn-in of 30,000 samples.

Once we have the MCMC fits for the input distributions, the reducibility factors in Eq. 2 can be extracted from the ratios of these fits. Fig. 3 displays the ratios of our input distributions, along with the results from the MCMC. Each fit is an element of the posterior distribution of the parameters that is sampled by the MCMC, each with a different minimum, or κ value. In order to extract the topics and calculate the corresponding uncertainty using Eq. 3, we sample κ_{AB} and κ_{BA} from the fits and calculate the mean and standard deviation of the distribution. Here, the reducibility factor for input A/input B, κ_{AB}, is extracted from the right tail of Fig. 3a, and similarly, κ_{BA} is extracted from the left tail of Fig. 3b, as that is where the minimum of the curve is located. The sampled reducibility factors are shown in Fig. 3.



(a) γ+jet / Dijet (pp)

(b) Dijet / γ+jet (pp)

FIG. 3: The ratio of input A and input B histograms, shown along with the MCMC fit and correspondingly sampled κ_{AB} and κ_{BA} values along the tails.

II. SIMULATED COLLISION EVENTS

This proof-of-concept study is based on proton-proton and heavy-ion collision events from the PYQUEN generator [57] with statistics accessible in Run 4 of the Large Hadron Collider. The PYQUEN event generator simulates radiative and collisional energy loss of partons in the QGP in heavy-ion collisions [57]. The input distributions are photon-jet (γ+jet) and dijet samples. We choose these because at Large Hadron Collider energies, γ+jet and dijets have different quark and gluon jet fractions. In Section IV, we consider modified jets in PYQUEN that are not embedded in thermal background. We will explore the consequences of thermal smearing in Section V.

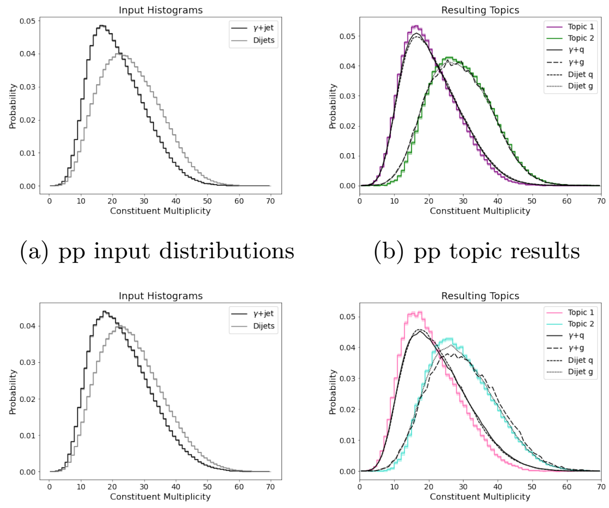
We generate proton-proton (pp) and heavy-ion (PbPb) events at $\sqrt{s} = 5.02$ TeV using $\hat{p_T} > 80$ GeV, where $\hat{p_T}$ is the hard scattering scale. The impact parameter range for PbPb events corresponds to 0-10% centrality. We use FASTJET 3.3.0 [63, 64] to reconstruct anti-k_t jets with radius $R = 0.4$ [65]. In the γ+jet samples, we select the leading jet in the opposite direction to the high-momentum photon ($|\Delta\phi| > \pi/2$), and in the dijet sam-

[2] The 18 fit parameters are composed of 3 parameters in α_A, 3 in α_B, and 12 in θ. The fractions represented by α_A and α_B, each contain 3 parameters since the last fraction can be calculated: $1 - (\alpha_1 + \alpha_2 + \alpha_3)$. Each skew normal distribution is defined by parameters μ_k, σ_k, s_k, which gives 12 parameters in θ when $N = 4$.

ples, we select the two jets in the event with the largest transverse momenta.

We only include jets with $80 < p_T < 100$ GeV and we impose cuts of $|\eta_{\text{jet}}| < 1$ and $|\eta_\gamma| < 1.442$. We additionally remove any jets for which a photon carries more than 80% of the jet p_T. This removes a low multiplicity peak in the γ+jet sample that is due to the clustering algorithm incorrectly recognizing a photon with high energy as the leading jet in the opposite azimuthal direction (compared to the high-momentum photon) [3].

Throughout this work, we will assess the performance of the topic modeling algorithm by presenting comparisons to distributions of quark- and gluon-initiated jets as defined at the Monte Carlo (MC) level. These labels are only defined at leading order and are therefore not strictly well-defined, but are nonetheless a useful proxy. To determine such labels, we compare the angular distance between the selected jet and the two outgoing matrix elements in the simulation. For γ+jet, we simply label the jet by the outgoing matrix element that is not the photon. For dijets, we match the jet to the outgoing matrix element with the smallest angular distance $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ from the jet. In all MC labels, we only include jets for which $\Delta R < 0.4$ between the matrix element and the jet axis. In pp, 97% of the jets in the γ+jet sample and 92% of those in the dijet sample satisfy the criterion to be given a quark/gluon truth label. In the PbPb sample, we utilize 95% and 89% for γ+jet and dijet quark/gluon truth labels, respectively. Ultimately, the MC labeling in any of the results should not be taken as the absolute truth, but rather as an approximation. Unless otherwise stated, we use the γ+jet sample for MC quark and gluon labels.



(a) pp input distributions

(b) pp topic results

(c) PbPb input distributions

(d) PbPb topic results

FIG. 4: Pyquen-generated proton-proton and heavy-ion normalized constituent multiplicity input distributions for γ+jet and dijets, along with topic extraction results, displaying the resulting topics in comparison to the gluon and quark MC truth labels.

III. TOPIC MODELING RESULTS

In this section, we show the results of the topic modeling algorithm on the Pyquen data. Since previous work [44] has demonstrated that constituent multiplicity approximately satisfies quark-gluon mutual irreducibility, we use this observable as input to the topic modeling. The input distributions and the resulting topics for both the proton-proton and heavy-ion data are shown in Fig. 4. We also show the MC truth-labeled quark and gluon distributions from γ+jet and dijet samples. The topic analysis is performed separately for simulated proton-proton and heavy-ion jets, meaning that the topics extracted from the two systems are fully independent.

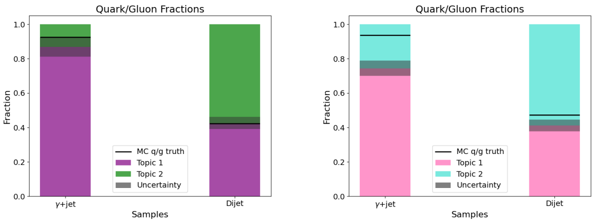
In general, the extracted topics correspond fairly well to the multiplicity distributions for quark- and gluon-initiated jets as defined from the MC level, with topics 1 and 2 corresponding to quark-like and gluon-like jets, respectively. The agreement between topics and the MC definition is slightly better for proton-proton than for heavy-ion events. The quark-like topics tend to be narrower in both proton-proton and heavy-ion samples, though the effect is exacerbated in heavy-ions. A similar discrepancy was also found in Ref. [55], based on results from a different Monte Carlo generator, Jewel. We note that the MC-level quark distribution is quite sensitive to the definition of "quark-initiated" jets, which is fundamentally ambiguous. A similar discrepancy was also found in Ref. [55] and was discussed in some detail in an Appendix therein. We additionally find a small discrepancy between the MC-level gluon distributions depending on whether they are estimated from the γ+jet or dijet sample, which could indicate mild sample-dependence of quark and gluon jets or biases due to the MC-level definition of the initiating parton. We note that more jets fail the criterion for MC quark and gluon labeling in heavy-ion compared to proton-proton samples, so the MC-level labeling is more uncertain in this case. On the other hand, the topics are defined on jet samples that include all jets, not just those that are close to one of the leading-order hard matrix elements.

Fig. 5 shows the quark and gluon fractions of each sample, extracted from the topic modeling algorithm. The results are compared to the corresponding quark and gluon fractions derived from the MC labels. While the topic and MC fractions are within uncertainties for the proton-proton sample, there are substantial differences in the heavy-ion sample, consistent with differences in the extracted topics. Moreover, the topic modeling algorithm more accurately reproduces the MC-level quark

[3] This can be interpreted as a back-to-back photon event, where a photon (plus noise) is incorrectly deemed a jet.

and gluon fractions for dijets than for γ+jets in both pp and PbPb collisions; this is because the quark-like topic deviates from the MC quark distribution and γ+jet has a higher quark fraction. These fractions are consistent with those found in JEWEL in [55].

Compared to [55] we also note that the constituent multiplicity distributions themselves are substantially different (primarily due to decays of π^0 which are allowed in this work). Though this impacts the extracted quark and gluon jet multiplicities, it does not impact the topic modeling algorithm itself as long as those distributions are (approximately) mutually irreducible and the features of quark and gluon jets do not depend on whether they are produced in γ+jet or dijet events. Excellent agreement between topic modeling results and the MC definition in proton-proton collisions suggests that quark and gluon multiplicities are approximately mutually irreducible and that γ+jet and dijet samples provide independent fractions of the same underlying quark and gluon jet distributions. Beyond additional ambiguities in the MC labeling, the worse performance of topic modeling in heavy-ion collisions could indicate either lower mutual irreducibility of quark and gluon constituent multiplicity or sample dependence, both of which could potentially arise from interactions with the medium. It is therefore non-trivial that we have found consistent results in PYQUEN to those in JEWEL, which have completely different descriptions of the medium interaction. This provides hope that topic modeling performance may be comparable in measurements to that seen in these two independent models.



(a) pp fractions

(b) PbPb fractions

FIG. 5: Quark and gluon fractions in each sample extracted via topic modeling, compared to MC truth quark and gluon fractions.

IV. JET SUBSTRUCTURE EXTRACTION

While the data-driven determination of quark- and gluon-like jet fractions in a sample is significant, applying these results to extract quark and gluon jet substructure allows for deeper insight into the modification of quark and gluon jets in the quark-gluon plasma. In this section, we will use the quark and gluon topic fractions extracted from constituent multiplicity distributions in the previous section to extract the quark- and gluon-like distributions for other jet observables. We will consider the jet shape, jet fragmentation, jet mass, and jet splitting fraction and compare the results to MC-labeled quark and gluon jet distributions. We also utilize these results to determine the modification of quark and gluon jet observables by taking the ratio of the jet observable between heavy-ion and the proton-proton samples, separately for quark- and gluon-like jets. While the MC-level definition of quark- and gluon-initiated jets is convenient as a qualitative benchmark for the success of this approach, we emphasize that it is both ill-defined and not measurable. The topic modeling procedure, therefore, provides novel access to the separate modification of quark and gluon jet substructure in heavy-ion collisions.

A. Jet Shape

Jet shape describes the jet transverse momentum distribution as a function of radial distance from the jet axis, and can be described by the following equation

$$\rho(r) = \frac{1}{r_b - r_a} \frac{1}{N_{\text{jet}}} \sum_{\text{jets}} \frac{\sum_{\text{tracks} \in [r_a, r_b)} p_T^{\text{track}}}{p_T^{\text{jet}}} . \quad (5)$$

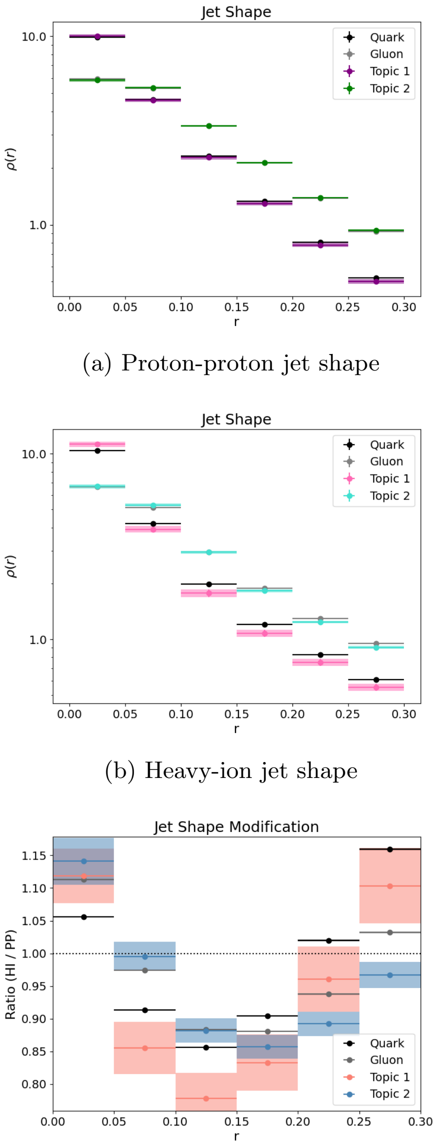
Here, r is the radial distance from the jet axis, and r_a, and r_b are the inner and outer radii of the given annulus [66]. Each annulus corresponds to a bin in the jet shape plot, where r_a is the left edge of the bin and r_b is the right edge of the bin.

In order to obtain the jet shape using our topic modeling results, we can simply perform a linear combination using the extracted κ values for each bin in the jet shape:

$$\begin{aligned} \rho_1(r) &= \frac{\rho_{\gamma+\text{jet}}(r) - \kappa_{AB}\rho_{\text{dijets}}(r)}{1 - \kappa_{AB}}, \\ \rho_2(r) &= \frac{\rho_{\text{dijets}}(r) - \kappa_{BA}\rho_{\gamma+\text{jet}}(r)}{1 - \kappa_{BA}} . \end{aligned} \quad (6)$$

Here, $\rho_{\gamma+\text{jet}}(r)$ and $\rho_{\text{dijets}}(r)$ are the jet shapes for the $\gamma + \text{jet}$ and dijets, respectively.

In Fig. 6 we show the jet shapes extracted from the topic modeling procedure compared to the MC-level distributions of the shape of quark- and gluon-initiated jets. In both proton-proton and heavy-ion collisions, the quark-like topic has a much narrower shape than the gluon-like one, consistent with the MC-level expectation. In proton-proton collisions, the topics are in excellent agreement with the MC definition of quark- and gluon-initiated jets. In heavy-ion collisions, the agreement is qualitative, with the topics being slightly narrower than the MC definition, consistent with the slightly lower multiplicity of topics compared to MC in Fig. 4d. The ratio between proton-proton and heavy-ion quark and gluon jet shapes are shown in Fig. 6c. The qualitative trend of the topic modification and the MC-level modification are the same, with quark jets having a larger deviation from one at small r, presumably due to their steeper jet



(a) Proton-proton jet shape

(b) Heavy-ion jet shape

(c) Jet shape modification ratio

FIG. 6: Proton-proton (a) and heavy-ion (b) jet shape extraction using topic modeling results from Fig. 4. The jet shape results for the two topics are shown in comparison to the jet shapes constructed from the quark and gluon MC truth labels. The modification of jet shape in the quark-gluon plasma, according to the extracted topics as well as the MC truth labels, is shown as the ratio between heavy-ion and proton-proton jet shape (c).

shape in pp. This feature is enhanced in the topic ratio for quark-like jets since the topic modeling result for the quark-like jet shape is steeper than the MC definition.

B. Jet Fragmentation Function

The topic modeling results from Section III can also be used to extract the quark and gluon jet fragmentation function. The jet fragmentation function gives the longitudinal momentum distribution of the tracks inside a jet,

$$D(\xi) = \frac{1}{N_{\rm jet}} \frac{dN_{\rm track}}{d\xi} \,. \tag{7}$$

Here, $N_{\rm jet}$ is the total number of jets, and $N_{\rm track}$ is the number of tracks in a jet. $\xi = \ln(1/z)$, where z is the longitudinal momentum fraction, is defined as

$$z = \frac{p_T \cos \Delta R}{p_T^{\rm jet}} = \frac{p_T}{p_T^{\rm jet}} \cos \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \,. \tag{8}$$

Here, $p_T^{\rm jet}$ is the transverse momentum of the jet relative to the beam direction, p_T is the transverse momentum of a charged particle in the jet, and $\Delta\eta$ and $\Delta\phi$ are measures of distance between the particle and E-scheme jet axis in pseudorapidity and azimuth [67].

In the jet fragmentation function, we can also compute each bin of the topics using a linear combination as shown below.

$$\begin{aligned} D_1(\xi) &= \frac{D_{\gamma+\rm jet}(\xi) - \kappa_{AB} D_{\rm dijets}(\xi)}{1-\kappa_{AB}}, \\ D_2(\xi) &= \frac{D_{\rm dijets}(\xi) - \kappa_{BA} D_{\gamma+\rm jet}(\xi)}{1-\kappa_{BA}} \end{aligned} \tag{9}$$

While the jet shape is self-normalized, the jet fragmentation function is normalized by the total number of jets, such that the integral of the histogram over ξ represents the average number of charged particles per jet. Therefore, rather than normalize to get a probability density, we take the direct combination of the per-jet quantities in each bin because we want the output to be a per-jet quantity. By definition,

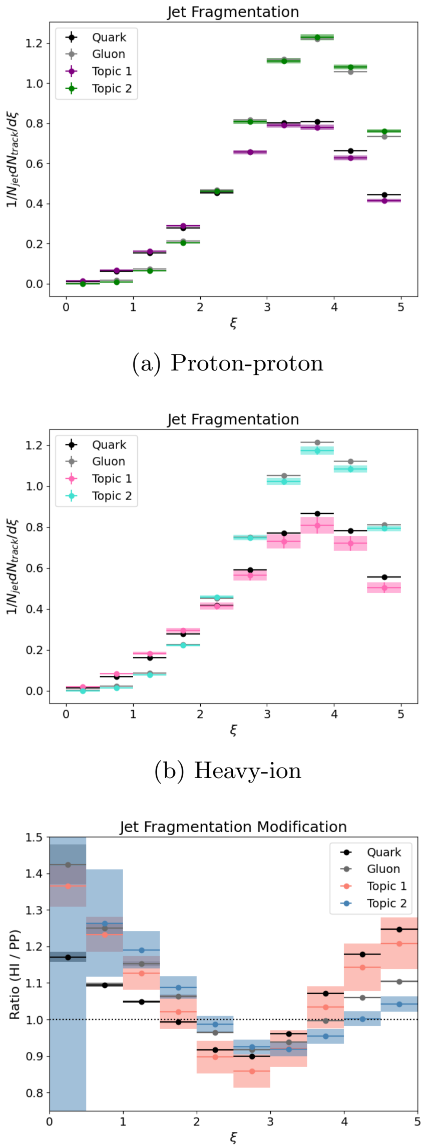
$$\begin{aligned} D_{\rm dijets}(\xi) &= f_d D_1(\xi) + (1-f_d) D_2(\xi) \\ D_{\gamma+\rm jet}(\xi) &= f_\gamma D_1(\xi) + (1-f_\gamma) D_2(\xi) \end{aligned} \tag{10}$$

After we integrate, we arrive at the following set of equations:

$$\begin{aligned} N_{\rm tracks}^{\rm (dijets)} &= f_d N_{\rm tracks}^{\rm (topic\ 1)} + (1-f_d) N_{\rm tracks}^{\rm (topic\ 2)} \\ N_{\rm tracks}^{(\gamma+\rm jet)} &= f_\gamma N_{\rm tracks}^{\rm (topic\ 1)} + (1-f_\gamma) N_{\rm tracks}^{\rm (topic\ 2)} \end{aligned} \tag{11}$$

which shows that the average number of tracks in dijets (or γ+jets) is equal to the weighted average of the average number of tracks of each topic. Therefore, rather than include any normalization, we directly apply κ to the dijet and γ+jet jet fragmentation values in order to solve for the jet fragmentation of the topics.

Fig. 7 shows the extracted jet fragmentation function using topic modeling compared to the fragmentation of MC-defined quark- and gluon-initiated jets. Overall



(a) Proton-proton

(b) Heavy-ion

(c) Jet fragmentation modification ratio

FIG. 7: Extracting jet fragmentation for proton-proton (a) and heavy-ion (b) collision using topic modeling results from Fig. 4. The jet fragmentation modification, represented by the ratio between PbPb and pp jet fragmentation, is also shown (c) for the topics, as well as quark and gluon.

there is a qualitative agreement between the topics and the MC definition over the full range of ξ. Fig. 7c shows the ratio between proton-proton and heavy-ion fragmentation functions for quark and gluon topics compared to MC results, which are in qualitative agreement. The quantitative agreement between quark (gluon) ratios at high (low) ξ is apparently accidental since there are differences in the fragmentation in both numerator and denominator.

C. Jet Mass

In addition to the jet shape and jet fragmentation, which measure the distribution of energy in specified areas of the jet cone, we also demonstrate topic modeling as applied to two additional per-jet substructure observables: jet mass and jet splitting fraction z_g.

The jet mass is calculated from the total four-momentum of all the constituents in the jet and is expressed as $m = \sqrt{E^2 - |\vec{p}|^2}$, where E is the jet energy and $\vec{p}$ is the momentum of the jet. To extract the jet mass histograms using our topic modeling results, we take a linear combination of the normalized jet mass input histograms, $H_{\gamma+\text{jet}}(m)$ and $H_{\text{dijets}}(m)$, using the extracted κ values:

$$\begin{aligned} H_1(m) &= \frac{H_{\gamma+\text{jet}}(m) - \kappa_{AB} H_{\text{dijets}}(m)}{1-\kappa_{AB}}, \\ H_2(m) &= \frac{H_{\text{dijets}}(m) - \kappa_{BA} H_{\gamma+\text{jet}}(m)}{1-\kappa_{BA}} \end{aligned} \tag{12}$$

The resulting jet mass histograms for quark and gluon topics in pp and PbPb are shown in Fig. 8, along with those for the MC-labelled quark and gluon samples. The ratio between the PbPb and pp jet mass histogram bins is shown in Fig. 8c. As before, we find qualitative agreement between the topics and MC-level quark- and gluon-initiated jet distributions, with the quark topic having a slightly lower mass consistent with the lower multiplicity of the quark topic in Fig. 4d. Larger deviations from the MC definition in the mass modification ratio at low and high masses are due to the slightly larger deviation of the quark topic in heavy-ions than in proton-proton.

D. Jet Splitting Fraction

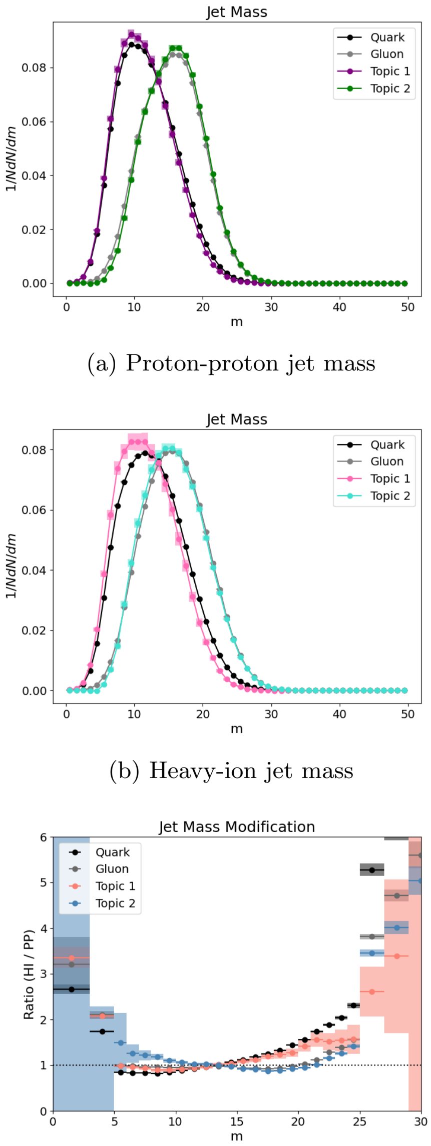
The jet momentum splitting fraction z_g describes the momentum sharing of the first hard splitting inside a jet and is related to the underlying QCD splitting functions. Technically, z_g is the momentum ratio of the leading and subleading subjets for the first splitting in a jet that passes the soft drop condition [68].

In order to find the leading and subleading subjets, we use SoftDrop [69] / mMDT [70] to decluster the jet's branching history, until the transverse momenta of the subjets fulfill the SoftDrop condition:

$$\frac{\min\left(p_{T,i}, p_{T,j}\right)}{p_{T,i} + p_{T,j}} > z_{cut}\theta^{\beta} \tag{13}$$

where θ represents the relative distance in the jet resolution parameter between the two subjets. The settings of SoftDrop used for this analysis were $z_{\text{cut}} = 0.1$ and $\beta = 0$ [68, 71, 72].

The procedure for extracting the topics is the same as for jet mass. The quark and gluon splitting functions extracted from topics and from the MC definition are shown in Fig. 9 for both proton-proton and heavy-ion samples.



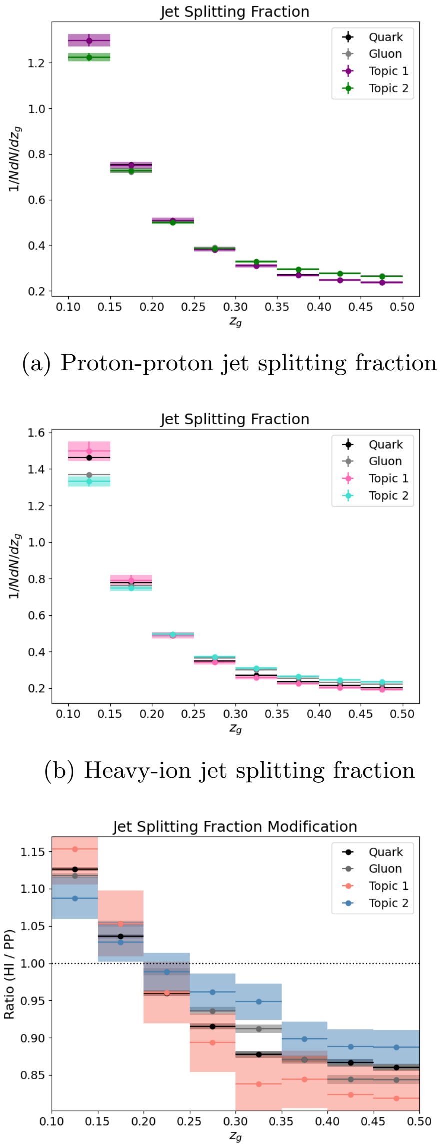
(a) Proton-proton jet mass

(b) Heavy-ion jet mass

(c) Jet mass modification ratio

FIG. 8: Topic modeling results for proton-proton (a) and heavy-ion (b) jet mass, compared to MC truth. Modification of the jet mass in the QGP, as defined by the ratio of the heavy-ion and proton-proton jet mass spectra is shown as well (c).

In this case, the topics agree semi-quantitatively with the MC definition in both proton-proton and heavy-ion samples, and in their ratio Fig. 9c. The better agreement between the topics and MC definition in this observable compared to others shown in this Section may be due to the comparatively weak dependence on the quark and gluon fractions.



(a) Proton-proton jet splitting fraction

(b) Heavy-ion jet splitting fraction

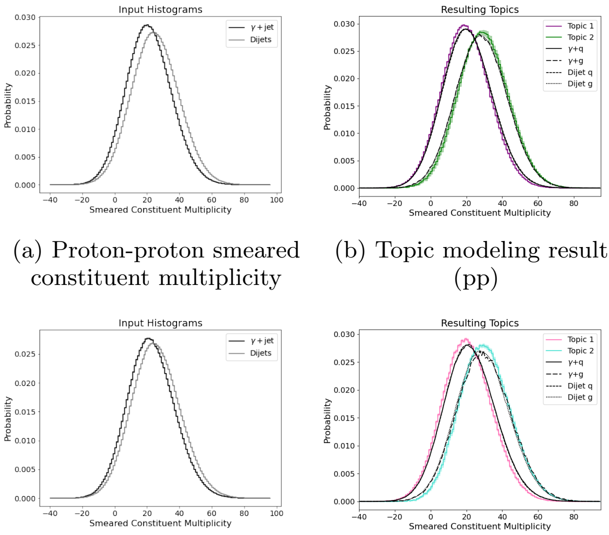
(c) Jet splitting fraction modification ratio

FIG. 9: Topic modeling results for proton-proton (a) and heavy-ion (b) jet mass, compared to MC truth. Modification of the jet mass in the QGP, as defined by the ratio of the heavy-ion and proton-proton jet mass spectra is shown as well (c).

V. SIMULATING THERMAL BACKGROUND

In heavy-ion collisions, particles from the jets are accompanied by a large background of particles from the quark-gluon plasma that adds additional particles in the jet radius. Even with effective background subtraction, this still contributes to non-negligible smearing of the properties of jets in experimental data. These effects contribute in addition to the Pyquen-generated PbPb distributions shown in this paper, which only include the jets themselves without effects from the background. Since the topic separation ultimately depends on multiplicity distributions which may be especially sensitive to these effects, in this section we estimate the effectiveness of the topic modeling in the presence of such smearing.

We consider smearing both the proton-proton and heavy-ion multiplicity distributions with this background and performing the topic separation on those smeared distributions. In proton-proton collisions, the smearing would be done for example through embedding in minimum-bias heavy-ion events. We use $dN/d\eta\, d\phi \sim 1600/(2\pi)$ [73] to estimate that the number of particles from the background inside of a cone of $R = 0.4$ is $\sim dN/d\eta\, d\phi(\pi R^2) \sim 11^2$. Assuming that background subtraction can eliminate the average, we estimate the expected fluctuations of particles within the jet cone as a Gaussian distribution $\mathcal{N}(0, 11)$. The resulting smeared histograms and smeared MC-labeled quark and gluon distributions are shown in Fig. 10. To extract reducibility factors we fit these distributions with 2 skew-normal and 2 normal underlying distributions, rather than with 4 skew-normal distributions as in the original method.



(a) Proton-proton smeared constituent multiplicity

(b) Topic modeling results (pp)

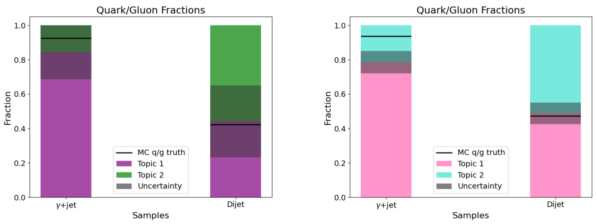
(c) Heavy-ion smeared constituent multiplicity

(d) Topic modeling results (PbPb)

FIG. 10: Smeared constituent multiplicity inputs (left) with corresponding topic separation results (right)

With smearing, the input multiplicity distributions for γ+jets and dijets are less distinguishable. However, the topic separation is still capable of performing well, which demonstrates the robustness of the algorithm to changes in the input distributions.

The quark-like fractions for γ+jets and dijets in each sample are shown in Fig. 11. The smeared quark and gluon fraction values are similar to those of the unsmeared dataset in Fig. 5. The uncertainty is much larger when background fluctuations are included because the absolute tails of the distributions (where the topics should be extracted) are at the tail of both the "true" multiplicity distributions and the Gaussian smearing distribution. Nonetheless, the consistency of the results indicates that the topic separation is resistant to constituent multiplicity fluctuations in the jet cone due to large backgrounds.



(a) Jet fractions in pp sample

(b) Jet fractions in PbPb sample

FIG. 11: Topic separation jet fractions displayed against MC truth quark and gluon truth values

The resulting quark/gluon substructure extraction using the smeared pp and heavy-ion datasets are in Appendix A. As in the main text, we find qualitative agreement between the extracted jet substructure and MC-level quark- and gluon-initiated jet distributions. The uncertainty on the pp results is significantly larger, while the increase in uncertainty in PbPb is not as drastic. Despite the precision of the topic separation deteriorating, the accuracy of calculated substructure values remains similar to the unsmeared dataset. This is also reflected in the modification plots comparing the ratio of the smeared heavy-ion substructure values to the unsmeared proton-proton substructure values.

VI. DISCUSSION AND CONCLUSION

In summary, our results from Pyquen-generated Monte Carlo samples corroborate previous proof-of-concept studies performed using Jewel, and demonstrate that a fully data-driven technique can potentially be used to extract separate quark and gluon jet distributions from experimental samples, without additional knowledge or templates. We extend the previous study by demonstrating that the resulting fractions can be used to extract jet substructure observables for quark- and gluon-like jets from γ+jet and dijet substructure measurements. As a proof-of-principle, we showed results for the jet shape, fragmentation function, jet mass, and splitting function separately for quark- and gluon-like jet topics. We additionally found that these results are robust to smearing the multiplicity distributions used to extract the topics by a large background as in heavy-ion collisions. These results suggest potential for an experimental determination of quark and gluon jet spectra and their substructure using this technique.

The resulting topics and their modification are in good qualitative agreement with the MC-level definition for quark- and gluon-initiated jets. There are quantitative discrepancies, which could result from ambiguities in the MC-level definition of quark and gluon jets. We define the MC labels to only include jets that have a quark or gluon outgoing matrix element within the jet radius, which does not include all jets in the sample. This implies that the input γ+jet and dijet samples, from which we extract topics, are not pure mixtures of Monte Carlo-labelled quark and gluon jets. Discrepancies could also arise from minor violations of the assumptions of the topic modeling algorithm. For example, if constituent multiplicities of quark and gluon jets are not fully mutually irreducible in heavy-ion collisions, different observables (for example, derived from machine learning) may be required to yield better results [45] and could be an interesting avenue for future work.

We have found that the substructure observables obtained from the topics provide a robust estimate of the quark and gluon jet substructure modification in the quark-gluon plasma. This provides a powerful technique to study quark and gluon jet modification in the quark-gluon plasma. If measured, quark and gluon jet substructure would provide strong additional constraints on the theory and modeling of jet interactions with the quark-gluon plasma.

CODE AVAILABILITY

The code for the topic modeling algorithm and the subsequent substructure observable extraction can be found at https://github.com/kying18/jet-topics.

ACKNOWLEDGEMENTS

We are grateful to Jesse Thaler and Nima Zardoshti for their valuable discussions on this topic and for their detailed comments on the manuscript. This work has been supported by the Department of Energy, Office of Science, under Grant No. DE-SC0011088 (to Y.Y., Y.C., and Y.L.).

Appendix A: Smeared Substructures and Modification Plots

The quark and gluon jet substructures extracted from the smeared constituent multiplicity are shown in Fig. 12 (pp) and Fig. 13 (PbPb). While the extracted substructures are quite similar between the smeared data and unsmeared data, the uncertainty is notably larger. In addition, the topic 2 jet mass calculation falls slightly negative in both the pp and PbPb samples, which is not physically attainable. This is due to high κ values on the left tail of the extraction.

We also display the modifications in the QGP in Fig. 14, which are determined using the ratio between the smeared heavy-ion jet substructure and the unsmeared proton-proton jet substructure.

[1] Wit Busza, Krishna Rajagopal, and Wilke van der Schee. Heavy ion collisions: The big picture and the big questions. *Annual Review of Nuclear and Particle Science*, 68(1):339–376, Oct 2018.

[2] Roman Kogler et al. Jet Substructure at the Large Hadron Collider: Experimental Review. *Rev. Mod. Phys.*, 91(4):045003, 2019.

[3] Guang-You Qin and Xin-Nian Wang. Jet quenching in high-energy heavy-ion collisions. *Int. J. Mod. Phys. E*, 24(11):1530014, 2015.

[4] Megan Connors, Christine Nattrass, Rosi Reed, and Sevil Salur. Jet measurements in heavy ion physics. *Rev. Mod. Phys.*, 90:025005, 2018.

[5] Leticia Cunqueiro and Anne M. Sickles. Studying the QGP with Jets at the LHC and RHIC. *Prog. Part. Nucl. Phys.*, 124:103940, 2022.

[6] Liliana Apolinário, Yen-Jie Lee, and Michael Winn. Heavy quarks and jets as probes of the QGP. *Prog. Part. Nucl. Phys.*, 127:103990, 2022.

[7] Yi Chen et al. Jet energy spectrum and substructure in e^+e^- collisions at 91.2 GeV with ALEPH Archived Data. 11 2021.

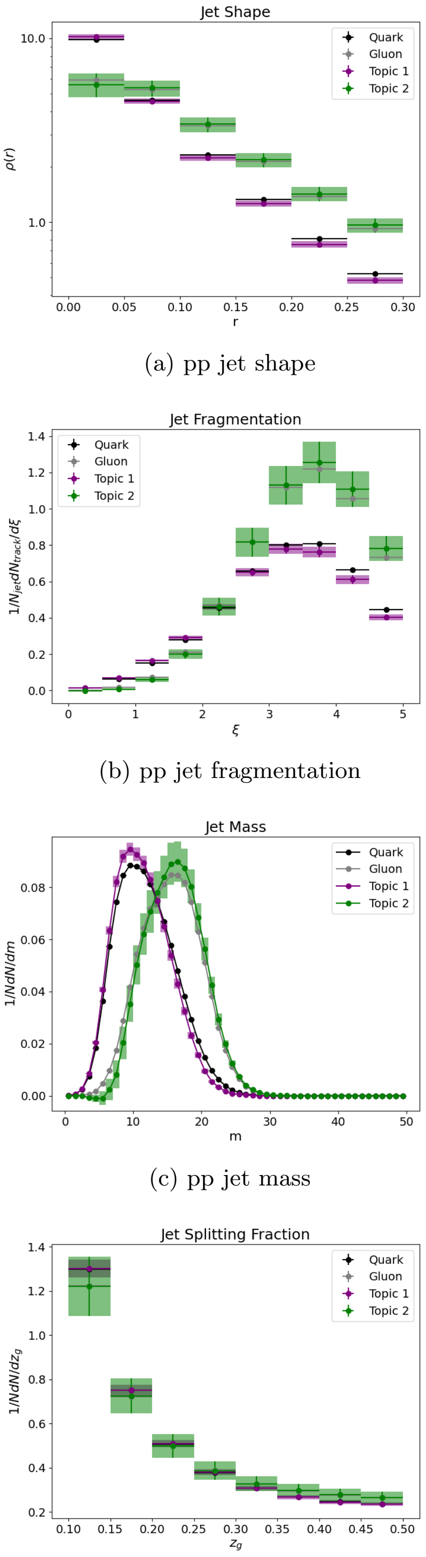
[8] Anthony Badea, Austin Baty, Paoti Chang, Gian Michele Innocenti, Marcello Maggi, Christopher Mcginn, Michael Peters, Tzu-An Sheng, Jesse Thaler, and Yen-Jie Lee. Measurements of two-particle correlations in e^+e^- collisions at 91 GeV with ALEPH archived data. *Phys. Rev. Lett.*, 123(21):212002, 2019.

[9] R. Baier, Yuri L. Dokshitzer, Alfred H. Mueller, S. Peigne, and D. Schiff. Radiative energy loss and p(T) broadening of high-energy partons in nuclei. *Nucl. Phys. B*, 484:265–282, 1997.

[10] R. Baier, Yuri L. Dokshitzer, Alfred H. Mueller, S. Peigne, and D. Schiff. Radiative energy loss of high-energy quarks and gluons in a finite volume quark - gluon plasma. *Nucl. Phys. B*, 483:291–320, 1997.

[11] B. G. Zakharov. Fully quantum treatment of the Landau-Pomeranchuk-Migdal effect in QED and QCD. *JETP Lett.*, 63:952–957, 1996.

[12] M. Gyulassy, P. Levai, and I. Vitev. NonAbelian energy loss at finite opacity. *Phys. Rev. Lett.*, 85:5535–5538, 2000.



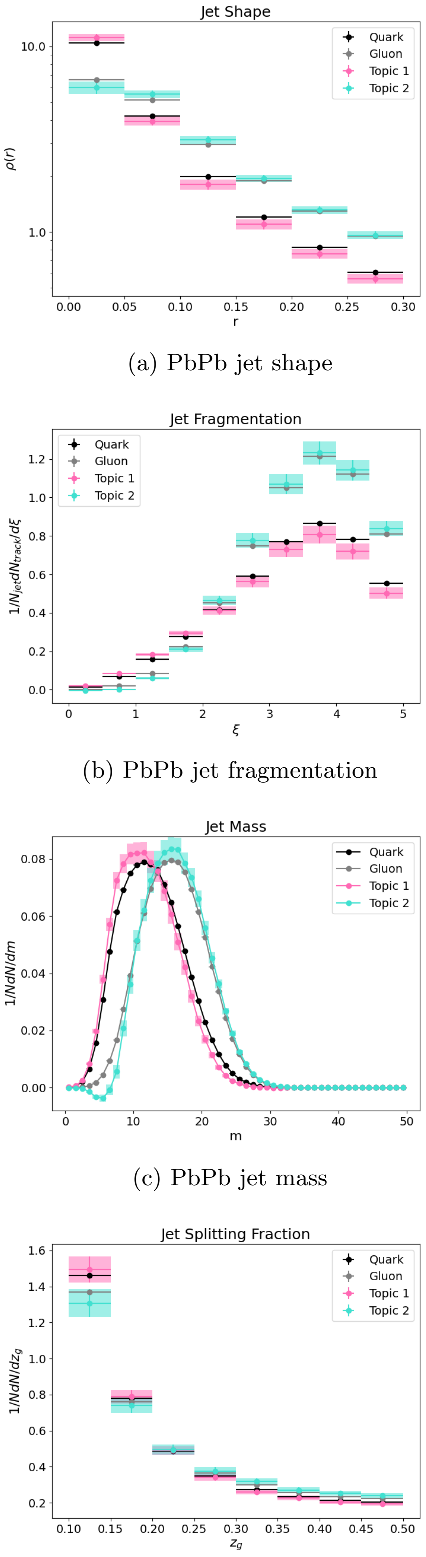
(a) pp jet shape

(b) pp jet fragmentation

(c) pp jet mass

(d) pp jet splitting fraction

FIG. 12: Proton-proton quark/gluon jet substructures extracted from smeared data



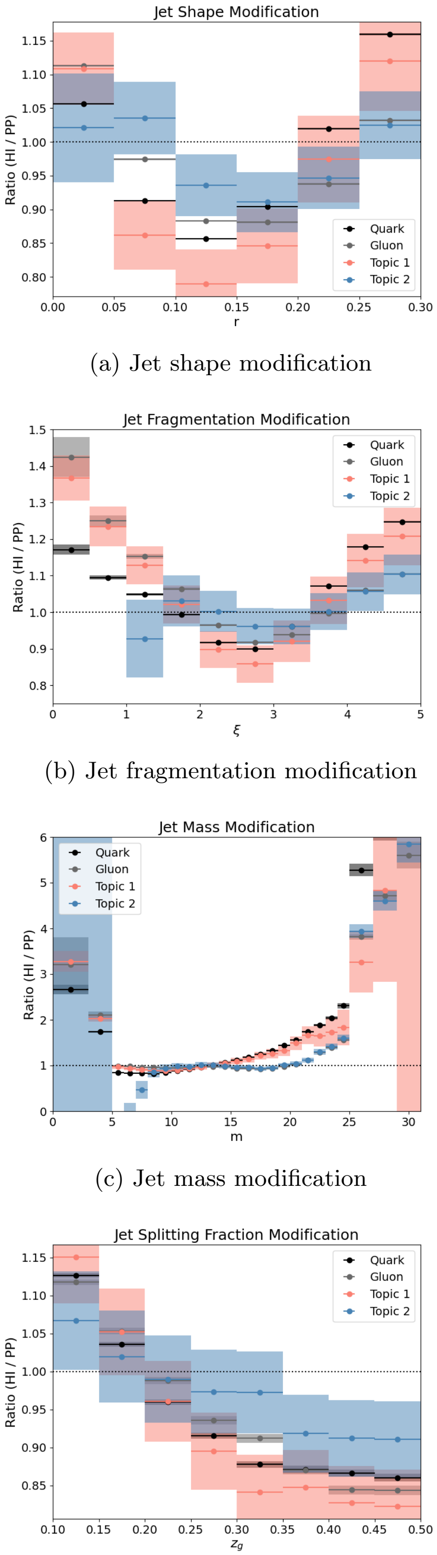
(a) PbPb jet shape

(b) PbPb jet fragmentation

(c) PbPb jet mass

(d) PbPb jet splitting fraction

FIG. 13: Heavy-ion quark/gluon jet substructures extracted from smeared data



(a) Jet shape modification

(b) Jet fragmentation modification

(c) Jet mass modification

(d) Jet splitting fraction modification

FIG. 14: QGP modification comparing smeared heavy-ion substructure results to unsmeared proton-proton substructure results

[13] M. Gyulassy, P. Levai, and I. Vitev. Reaction operator approach to nonAbelian energy loss. *Nucl. Phys. B*, 594:371–419, 2001.

[14] Urs Achim Wiedemann. Gluon radiation off hard quarks in a nuclear environment: Opacity expansion. *Nucl. Phys. B*, 588:303–344, 2000.

[15] Carlos A. Salgado and Urs Achim Wiedemann. Calculating quenching weights. *Phys. Rev. D*, 68:014008, 2003.

[16] Abhijit Majumder. Hard collinear gluon radiation and multiple scattering in a medium. *Phys. Rev. D*, 85:014023, 2012.

[17] Xiao-feng Guo and Xin-Nian Wang. Multiple scattering, parton energy loss and modified fragmentation functions in deeply inelastic e A scattering. *Phys. Rev. Lett.*, 85:3591–3594, 2000.

[18] Xin-Nian Wang and Xiao-feng Guo. Multiple parton scattering in nuclei: Parton energy loss. *Nucl. Phys. A*, 696:788–832, 2001.

[19] Paul M. Chesler and Krishna Rajagopal. Jet quenching in strongly coupled plasma. *Phys. Rev. D*, 90(2):025033, 2014.

[20] Jorge Casalderrey-Solana, Doga Can Gulhan, José Guilherme Milhano, Daniel Pablos, and Krishna Rajagopal. A Hybrid Strong/Weak Coupling Approach to Jet Quenching. *JHEP*, 10:019, 2014. [Erratum: JHEP 09, 175 (2015)].

[21] Bjoern Schenke, Charles Gale, and Sangyong Jeon. MARTINI: An Event generator for relativistic heavy-ion collisions. *Phys. Rev. C*, 80:054913, 2009.

[22] Chanwook Park, Sangyong Jeon, and Charles Gale. Jet modification with medium recoil in quark-gluon plasma. *Nucl. Phys. A*, 982:643–646, 2019.

[23] J. D. Bjorken. Energy Loss of Energetic Partons in Quark - Gluon Plasma: Possible Extinction of High p(t) Jets in Hadron - Hadron Collisions. 8 1982.

[24] Georges Aad et al. Observation of a Centrality-Dependent Dijet Asymmetry in Lead-Lead Collisions at $\sqrt{s_{NN}} = 2.77$ TeV with the ATLAS Detector at the LHC. *Phys. Rev. Lett.*, 105:252303, 2010.

[25] Serguei Chatrchyan et al. Observation and studies of jet quenching in PbPb collisions at nucleon-nucleon center-of-mass energy = 2.76 TeV. *Phys. Rev. C*, 84:024906, 2011.

[26] Georges Aad et al. Measurement of the jet radius and transverse momentum dependence of inclusive jet suppression in lead-lead collisions at $\sqrt{s_{NN}}$= 2.76 TeV with the ATLAS detector. *Phys. Lett. B*, 719:220–241, 2013.

[27] Jaroslav Adam et al. Measurement of jet suppression in central Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 2.76$ TeV. *Phys. Lett. B*, 746:1–14, 2015.

[28] Vardan Khachatryan et al. Measurement of inclusive jet cross sections in *pp* and PbPb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. C*, 96(1):015202, 2017.

[29] Morad Aaboud et al. Measurement of the nuclear modification factor for inclusive jets in Pb+Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV with the ATLAS detector. *Phys. Lett. B*, 790:108–128, 2019.

[30] Shreyasi Acharya et al. Measurements of inclusive jet spectra in pp and central Pb-Pb collisions at $\sqrt{s_{\rm NN}} =$ 5.02 TeV. *Phys. Rev. C*, 101(3):034911, 2020.

[31] Albert M Sirunyan et al. First measurement of large area jet transverse momentum spectra in heavy-ion collisions. *JHEP*, 05:284, 2021.

[32] L. Adamczyk et al. Measurements of jet quenching with

semi-inclusive hadron+jet distributions in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV. *Phys. Rev. C*, 96(2):024905, 2017.

[33] Serguei Chatrchyan et al. Modification of Jet Shapes in PbPb Collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Lett. B*, 730:243–263, 2014.

[34] Serguei Chatrchyan et al. Measurement of Jet Fragmentation in PbPb and pp Collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. C*, 90(2):024908, 2014.

[35] Georges Aad et al. Measurement of inclusive jet charged-particle fragmentation functions in Pb+Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV with the ATLAS detector. *Phys. Lett. B*, 739:320–342, 2014.

[36] Morad Aaboud et al. Measurement of jet fragmentation in Pb+Pb and *pp* collisions at $\sqrt{s_{\rm NN}} = 2.76$ TeV with the ATLAS detector at the LHC. *Eur. Phys. J. C*, 77(6):379, 2017.

[37] Albert M Sirunyan et al. Observation of Medium-Induced Modifications of Jet Fragmentation in Pb-Pb Collisions at $\sqrt{s_{NN}} = 5.02$ TeV Using Isolated Photon-Tagged Jets. *Phys. Rev. Lett.*, 121(24):242301, 2018.

[38] Albert M Sirunyan et al. Jet Shapes of Isolated Photon-Tagged Jets in Pb-Pb and pp Collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV. *Phys. Rev. Lett.*, 122(15):152001, 2019.

[39] Morad Aaboud et al. Comparison of Fragmentation Functions for Jets Dominated by Light Quarks and Gluons from *pp* and Pb+Pb Collisions in ATLAS. *Phys. Rev. Lett.*, 123(4):042001, 2019.

[40] Yang-Ting Chien and Raghav Kunnawalkam Elayavalli. Probing heavy ion collisions using quark and gluon jet substructure, 2018.

[41] Yacine Mehtar-Tani and Soeren Schlichting. Universal quark to gluon ratio in medium-induced parton cascade. *JHEP*, 09:144, 2018.

[42] Jian-Wei Qiu, Felix Ringer, Nobuo Sato, and Pia Zurita. Factorization of jet cross sections in heavy-ion collisions. *Phys. Rev. Lett.*, 122(25):252301, 2019.

[43] Liliana Apolinário, João Barata, and Guilherme Milhano. On the breaking of Casimir scaling in jet quenching. *Eur. Phys. J. C*, 80(6):586, 2020.

[44] Eric M. Metodiev and Jesse Thaler. Jet topics: Disentangling quarks and gluons at colliders. *Physical Review Letters*, 120(24), Jun 2018.

[45] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. An operational definition of quark and gluon jets. *Journal of High Energy Physics*, 2018(11), Nov 2018.

[46] G. Aad, B. Abbott, J. Abdallah, S. Abdel Khalek, O. Abdinov, R. Aben, B. Abi, M. Abolins, O. S. AbouZeid, and et al. Light-quark and gluon jet discrimination in *pp* collisions at $\sqrt{s} = 7$ TeV with the atlas detector. *The European Physical Journal C*, 74(8), Aug 2014.

[47] Lorella M. Jones. Tests for Determining the Parton Ancestor of a Hadron Jet. *Phys. Rev. D*, 39:2550, 1989.

[48] Z. Fodor. How to See the Differences Between Quark and Gluon Jets. *Phys. Rev. D*, 41:1726, 1990.

[49] Jason Gallicchio and Matthew D. Schwartz. Quark and gluon jet substructure. *Journal of High Energy Physics*, 2013(4), Apr 2013.

[50] Jason Gallicchio and Matthew D. Schwartz. Quark and gluon tagging at the lhc. *Physical Review Letters*, 107(17), Oct 2011.

[51] 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. *JHEP*, 07:091, 2017.

[52] Christopher Frye, Andrew J. Larkoski, Jesse Thaler, and Kevin Zhou. Casimir Meets Poisson: Improved Quark/Gluon Discrimination with Counting Observables. *JHEP*, 09:083, 2017.

[53] Andrew J. Larkoski and Eric M. Metodiev. A Theory of Quark vs. Gluon Discrimination. *JHEP*, 10:014, 2019.

[54] Frederic Dreyer, Gregory Soyez, and Adam Takacs. Quarks and gluons in the Lund plane. 12 2021.

[55] Jasmine Brewer, Jesse Thaler, and Andrew P. Turner. Data-driven quark- and gluon-jet modification in heavy-ion collisions. *Physical Review C*, 103(2), Feb 2021.

[56] A. M. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, T. Bergauer, M. Dragicevic, J. Erö, A. Escalante Del Valle, M. Flechl, and et al. Measurement of quark- and gluon-like jet fractions using jet charge in PbPb and pp collisions at 5.02 TeV. *Journal of High Energy Physics*, 2020(7), Jul 2020.

[57] PYQUEN event generator. http://lokhtin.web.cern.ch/lokhtin/pyquen/.

[58] Julian Katz-Samuels, Gilles Blanchard, and Clayton Scott. Decontamination of mutual contamination models, 2019.

[59] Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. Classification with asymmetric label noise: Consistency and maximal denoising, 2016.

[60] Steve Baker and Robert D. Cousins. Clarification of the use of chi-square and likelihood functions in fits to histograms. *Nuclear Instruments and Methods in Physics Research*, 221(2):437–442, 1984.

[61] Xiangpan Ji, Wenqiang Gu, Xin Qian, Hanyu Wei, and Chao Zhang. Combined neyman–pearson chi-square: An improved approximation to the poisson-likelihood chi-square. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 961:163677, May 2020.

[62] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The mcmc hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312, Mar 2013.

[63] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. Fastjet user manual. *The European Physical Journal C*, 72(3), Mar 2012.

[64] Matteo Cacciari and Gavin P. Salam. Dispelling the n^3 myth for the Kt jet-finder. *Physics Letters B*, 641(1):57–61, Sep 2006.

[65] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008.

[66] S. Chatrchyan, V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, M. Friedl, and et al. Modification of jet shapes in pbpb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Physics Letters B*, 730:243–263, Mar 2014.

[67] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, M. Friedl, and et al. Measurement of jet fragmentation in pbpb and pp collisions at sqrt(s[NN]) = 2.76 TeV. *Physical Review C*, 90(2), Aug 2014.

[68] A. M. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al. Measurement of the splitting function in pp and pb-pb collisions at $\sqrt{s_{NN}} = 5.02$ tev. *Physical Review Letters*, 120(14), Apr 2018.

[69] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, and Jesse Thaler. Soft drop. *Journal of High Energy Physics*, 2014(5), May 2014.

[70] Mrinal Dasgupta, Alessandro Fregoso, Simone Marzani, and Gavin P. Salam. Towards an understanding of jet substructure. *Journal of High Energy Physics*, 2013(9), Sep 2013.

[71] K. Kauder. Measurement of the shared momentum fraction zg using jet reconstruction in p+p and au+au collisions with star. *Nuclear and Particle Physics Proceedings*, 289-290:137–140, 2017. 8th International Conference on Hard and Electromagnetic Probes of High Energy Nuclear Collisions.

[72] Simone Marzani, Gregory Soyez, and Michael Spannowsky. Looking inside jets. *Lecture Notes in Physics*, 2019.

[73] Serguei Chatrchyan et al. Dependence on pseudorapidity and centrality of charged hadron production in PbPb collisions at a nucleon-nucleon centre-of-mass energy of 2.76 TeV. *JHEP*, 08:141, 2011.