Quark-gluon tagging with shower deconstruction: Unearthing dark matter and Higgs couplings

Danilo Ferreira de Lima, Petar Petrov, Davison Soper, and Michael Spannowsky

Phys. Rev. D **95**, 034001 — Published 1 February 2017

DOI: 10.1103/PhysRevD.95.034001

Prepared for submission to JHEP   DCPT/16/140, IPPP/16/70

Quark-Gluon tagging with Shower Deconstruction: Unearthing dark matter and Higgs couplings

Danilo Ferreira de Lima,[a] **Petar Petrov,**[b] **Davison Soper**[c] **and Michael Spannowsky**[b]

[a] *Physikalisches Institut, Ruprechts-Karls-Universität Heidelberg, 69120 Heidelberg, Germany*

[b] *Institute for Particle Physics Phenomenology, Department of Physics, Durham University, DH1 3LE, United Kingdom*

[c] *Institute of Theoretical Science University of Oregon Eugene, OR 97403-5203, USA*

E-mail: dferreir@cern.ch, p.m.petrov@durham.ac.uk, soper@uoregon.edu, michael.spannowsky@durham.ac.uk

Abstract: The separation of quark and gluon initiated jets can be an important way to improve the sensitivity in searches for new physics or in measurements of Higgs boson properties. We present a simplified version of the shower deconstruction approach as a novel observable for quark-gluon tagging. Assuming topocluster-like objects as input, we compare our observable with energy correlation functions and find a favorable performance for a large variety of jet definitions. We address the issue of infrared sensitivity of quark-gluon discrimination. When this approach is applied to dark matter searches in mono-jet final states, limitations from small signal-to-background ratios can be overcome. We also show that quark-gluon tagging is an alternative way of separating weak boson from gluon-fusion production in the process $p + p \to H + \text{jet} + \text{jet} + X$.

1 Introduction

Quark-gluon tagging of jets can be an important tool to separate signal from backgrounds. For instance, it is of interest to search for dark matter production by using the process in which produced dark matter particles recoil against a single jet, as described in [1]. Particularly when the mediator between the dark matter and the Standard Model particles is a scalar that couples preferably to the third-generation fermions, the associated jet is likely to be gluon initiated. One of the dominant Standard Model backgrounds however is the production of a jet plus a Z boson, in which the Z boson decays to $\nu\bar{\nu}$. In the tree level diagram for the background, the jet can be either quark initiated or a gluon initated. Thus if we can preferentially reject quark jets and keep gluon jets, we can improve the ratio of signal events retained to background events retained.

Conversely, many measurements of Higgs boson properties and couplings rely on the weak-boson-fusion production process $qq \to Hqq$ [2–5]. In particular, if one wants to measure the Higgs boson coupling to gauge bosons, one wants to look at this process and not the dominating gluon-fusion process $gg \to Hgg$ [6]. In $qq \to Hqq$, there are two quark jets, while in $gg \to Hgg$, there are two gluon jets. Hence, here we would prefer to reject gluon jets and keep quark jets to improve the precision of the measurement.

A third example would be the decays of squarks into jets and the lightest supersymmetric particle. Heavy squarks of the first and second generation decay almost exclusively into quarks and gauginos, while jets and missing transverse energy (MET) backgrounds have a larger gluon-jet component.

In all examples above, exploiting the different admixture of gluon and quark initiated jets can help to improve the signal-to-background ratio. Consequently, several observables have been proposed to exploit the differences in the radiation profiles of quarks and gluons [7–13] and have been studied in data by ATLAS [14] and CMS [15].

Suppose that we want to accept quark jets and reject gluon jets. Typically, one can adjust the parameters of the algorithm we use so as to obtain a desired fraction ε_s of quark jets accepted. Then ε_b^{-1}, the inverse of the fraction of gluon jets accepted, will depend on ε_s. In this paper, we present "ROC" curves showing $\varepsilon_b^{-1}(\varepsilon_s)$ versus ε_s. We want ε_b^{-1} to be as large as possible for any given ε_s. However, this performance metric is not the only issue that we need to address. We also need to know with reasonable accuracy the value of $\varepsilon_b^{-1}(\varepsilon_s)$ for a given ε_s. This information can come from experiment if the function $\varepsilon_b^{-1}(\varepsilon_s)$ is characteristic of quark-initiated versus gluon-initiated jets independently of how the jets are produced. We will investigate whether this is so in section 4. Information on $\varepsilon_b^{-1}(\varepsilon_s)$ for a given tagging method can also come from perturbation theory and simulation using parton shower event generators. Here, the findings of [14] indicate the need for the inclusion of certain detector effects in phenomenological analyses and the benefit of observables that are largely insensitive to non-perturbative effects. In this paper, we try to avoid sensitivity to parton splitting processes at very small momentum scales. For instance, we use observables that are technically infrared safe. However, we will discover that it is precisely parton splitting processes at quite small momentum scales that best distinguish the substructure of a quark jet from that of a gluon jet. Thus we cannot avoid a certain degree of infrared sensitivity. We return to this issue in section 4.

In this paper, we explore the use of several methods to distinguish between quark in gluon jets in $p + p \to Z + \mathrm{jet} + X$ and $p + p \to \mathrm{jet} + \mathrm{jet} + X$ events. We evaluate the performance and simulation uncertainties of the shower deconstruction method [16–18] and compare it to the use of energy correlation functions [10].

The structure of the paper is as follows: In section 2 we describe our analysis setup and the algorithms applied for quark/gluon tagging, emphasizing a method based on shower deconstruction. In section 3, we discuss their performance and uncertainties of these algorithms. We apply quark/gluon tagging based on shower deconstruction to dark matter searches and $p + p \to H + \mathrm{jet} + \mathrm{jet} + X$ production and evaluate by how much the signal-to-background ratio can be improved in section 5. In section 6 we offer a summary and our conclusions.

2 Jet substructure for quark-gluon tagging

In this section, we first describe the analysis setup for the paper. Then we discuss the input objects that we use for quark-jet versus gluon-jet discrimination. Next, we turn to the observables that we use.

2.1 The analysis setup

Our aim in this paper is to test the performance of algorithms designed to discriminate between quark-initiated jets and gluon-initiated jets. For this, we use two types of of events generated using Pythia 8 [19] with initial state radiation and underlying event switched on. The first type, and the one on which we will focus most, is a single jet with an associated invisible Z boson - $qg \to qZ(\nu\bar{\nu})$, $q\bar{q} \to gZ(\nu\bar{\nu})$. The other, which we use to show how much the tagging efficiency is affected by the event color flow, is dijet production $qq/gg \to qq$, $q\bar{q}/gg \to gg$. We generate four sets of each type in order to compare the performance at different limits for the transverse momentum in the hard scattering: $p_T > 200, 400, 600, 1000$ GeV.

For each event, we begin with input objects. The input objects can be hadrons, tracks, or certain calorimeter based objects, as described in the following subsection. We cluster the input objects into jets and select the leading jet: the one with the greatest transverse momentum. This is the “fat jet” that we wish to tag as being a probable quark jet or a probable gluon jet. To proceed, there should be at least one jet in the rapidity range $|y| < 5$ for $Z + \mathrm{jet}$ events or two such jets for dijet events. For the clustering into jets, we use the C/A algorithm with a standard radius $R_{\mathrm{fj}} = 0.4$ and a transverse momentum that reflects the event generation limit $p_{T\mathrm{fj}} > p_{T\mathrm{limit}}$. With $R = 0.4$, the fat jet is not so fat. This choice follows from the fact that we are analyzing the QCD radiation in the jet rather than looking for the decay of a heavy particle as is the case in many jet substructure studies. We also use a larger radius jet definition at $R_{\mathrm{fj}} = 0.8$ for some analyses.

2.2 Input objects

The observable quantities that we analyze for their ability to distinguish quark jets from gluon jets are built from certain input objects. We study four different classes of input objects: hadrons, tracks, and two sorts of calorimeter based objects.

While hadrons as input objects provide the most detailed information in the substructure of a jet, they are unlikely to be accessible in an experimental environment.

Using tracks allows very good angular resolution, but only for charged particles, while being blind to neutral particles. For tracks, we do not include a detector simulation, so that we do not take into account track efficiencies or energy smearing of tracks. Thus we likely overestimate the performance of the observables with track inputs.

Most of the analyses that we present are based on input objects built from idealized calorimeter cells. In general purpose experiments such as ATLAS [20] and CMS [21], often the calorimeter cells are not directly used to make jets. Instead, a combination of cells is used.

ATLAS uses "topoclusters" [22–24]. A topocluster is a group of topologically connected calorimeter cells, which are chosen based on an algorithm to suppress calorimeter noise. The algorithm starts by choosing a "seed" calorimeter cell, which has a signal over noise ratio over a specific threshold. It then combines it with neighbour cells that satisfy a minimum signal-to-noise ratio criterion iteratively. This method improves the jet algorithm inputs signal-to-noise ratio. Although it has the positive effect of improving the calorimeter's signal-to-noise ratio [24], it imposes a limitation in the angular resolution of the experiments. While the algorithm used to create topoclusters is clearly defined, the angular resolution limitation is not explicit in the algorithm. It depends on the calorimeter's noise average and cell sizes, which vary in both ATLAS and CMS, depending on the jet position.

Following a somewhat different approach, CMS uses so-called particle-flow (PF) objects [25]. PF objects consist of all visible particles in an event, i.e. muons, electrons, photons, charged hadrons, and neutral hadrons. Charged hadrons, electrons and muons are predominantly reconstructed from tracks in the tracker, while photons and neutral hadrons are reconstructed from energy deposits in topoclusters. Combining the topocluster and tracking system information, CMS can greatly improve the PF jets' spatial resolution with respect to calorimeter jets, e.g. by exploiting tracking information [26–29]. However, the jet-energy-resolution deteriorates quickly for jets with $R \leq 0.2$ [30]. Hence, the way CMS uses its PF objects currently results in a lower limit on the spatial resolution of jets, just as the angular resolution is limited by the size of topoclusters in ATLAS.

We conclude that jet substructure methods must take into account the finite angular resolution of calorimeter objects used as substructure inputs. In this phenomenological study, we approximate this resolution limitation by using Cambridge-Aachen (CA) [31, 32] jets with an R parameter of 0.1 and $p_T > 1$ GeV as input to the algorithms. We use two sorts of calorimetric input objects, which we call "massive topoclusters" and "massless topoclusters."

ATLAS topoclusters are forced to be massless. That is, after measuring the energy, pseudorapidity and azimuthal angle of the topocluster, its three-momentum is scaled to create a vector with $p^2 = 0$. We create massless topoclusters with this rescaling. However, we mostly use massive topoclusters, in which the topocluster momentum p is the sum of the momenta of the constituent particles, so that $p^2 > 0$.

We mostly use massive instead of massless input objects because we find that neglecting their masses leads to a deterioration in quark-gluon discrimination. One could imagine using a similar procedure to that described in [33] to calibrate the masses of small jets, analogous to our "massive topoclusters."

2.3 Observables for quark-gluon tagging

We will use two classes of jet substructure observables in order to distinguish quark jets from gluon jets. One is based on shower deconstruction, the other is based on energy correlations. We begin with shower deconstruction.

2.3.1 Shower deconstruction

Shower deconstruction [16–18] is a general method for distinguishing events created by a sought signal process from events created by other, less interesting, processes. In this case, the "signal" process creates a quark-initiated jet and we wish to distinguish this quark jet from "background" gluon jets. (Of course, we could reverse the roles of signal and background here.) We start with a list of the momenta $\{p\}_m = \{p_1, p_2, \ldots, p_m\}$ of m microjets – small radius jets – constructed from the contents of the larger fat jet. We calculate an approximation $P(\{p\}_m|q)$ that the observed microjets could be the result of a parton shower that starts with a quark parton and ends with m partons with momenta $\{p\}_m$. We similarly calculate an approximate probability $P(\{p\}_m|g)$ to obtain the observed microjets starting from a quark. Then we form the likelihood ratio

$$\chi(q, g) = \frac{P(\{p\}_m|q)}{P(\{p\}_m|g)} \ , \tag{2.1}$$

where the first argument indicates the signal hypothesis, *i.e.* quarks, and the second argument the background hypothesis, *i.e.* gluons. Note that $\chi(g, q) = 1/\chi(q, g)$. A large value of $\chi(q, g)$ indicates a likely quark jet, while a small value of $\chi(q, g)$ indicates a likely gluon jet. Thus imposing a cut $\chi(q, g) > \chi_{\rm cut}$ tags quark jets and imposing a cut $\chi(g, q) > \chi_{\rm cut}$ tags gluon jets.

The idea of the shower deconstruction method here is to distinguish the radiation pattern created by an initial quark from the radiation pattern of a gluon. This is rather different from our previous applications of shower deconstruction, in which the aim is to distinguish the pattern of partons produced by the decay of a heavy particle, such as a top quark, from the pattern of partons produced by normal QCD radiation. Distinguishing quark jets from gluon jets is harder. We have normal QCD radiation in either case, but gluon jets have, on average, more radiation because gluons have a larger color charge. We expect to see two differences between quark and gluon jets. First, gluon jets ought to be more likely to contain more microjets than quark jets. Second, the virtuality p_i^2 of the highest p_T microjet is likely to be larger in the gluon case than in the quark case because the microjet contains more radiation inside it even though the radiation is clustered into a single microjet.

To see how this works, we apply shower deconstruction for $qg \to qZ(\nu\bar{\nu})$, $q\bar{q} \to gZ(\nu\bar{\nu})$ events, taking massive topoclusters as the the input objects and using them to define a fat jet using a jet radius $R = 0.8$. The massive topoclulsters in the original fat jet are grouped into microjets using the k_T algorithm with radius $R_\mathrm{mj} = 0.3$ and a minimum transverse momentum $p_{T\mathrm{mj}}^{\mathrm{min}} = 10$ GeV. Then the likelihood variable χ from eq. (2.1) is calculated for each event. Different events have different numbers of microjets. In the right hand plot of figure 1, we plot the number of microjets in the gZ sample (blue) and in the qZ sample (green). Not surprisingly, quark jets are more likely than gluon jets to produce just one microjet, while gluon jets produce more microjets. This feature can help distinguish quark jets from gluon jets. However, when we look at the distribution of χ for those events with exactly one microjet, we find better quark-gluon discrimination than when we look for χ for those events with exactly two microjets, as illustrated in the left-hand plot of figure 1. This suggests that there is a lot of discriminating power in the shower-deconstruction χ for the simple case of one microjet. In fact, we find that when we simply calculate χ for the fat jet as a whole, without decomposing it into microjets, we get quark-gluon discriminating power that is often better than when the fat jet is decomposed into several microjets. This behavior is in sharp contrast to applications in which one wants to distinguish ordinary QCD jets from jets arising from the decay of a heavy particle like a top quark: it is important that a top quark decays into at least three jets.

Because using shower deconstruction with just one microjet works quite well, it is of interest to understand what shower deconstruction does in this case. The formula for χ for just one microjet is simply a ratio of Sudakov factors:

$$\chi = \frac{P(\{p\}_m|q)}{P(\{p\}_m|g)} = \frac{e^{-\mathcal{S}_q}}{e^{-\mathcal{S}_g}} = e^{-\left(\mathcal{S}_\mathrm{qqg}\Theta(\mathcal{S}_\mathrm{qqg}>0)-\mathcal{S}_\mathrm{ggg}\Theta(\mathcal{S}_\mathrm{ggg}>0)-n_f\mathcal{S}_\mathrm{gqq}\right)} \,. \tag{2.2}$$

where

$$\begin{aligned}
\mathcal{S}_\mathrm{qqg} &= \frac{C_\mathrm{F}}{\pi b_0^2}\left\{\ln\left(\frac{\alpha_\mathrm{S}(\mu_J^2)}{\alpha_\mathrm{S}(k_J^2)}\right)\left[\frac{1}{\alpha_\mathrm{S}(R_\mathrm{fj}^2 k_J^2)} - \frac{3b_0}{4}\right] + \frac{1}{\alpha_\mathrm{S}(\mu_J^2)} - \frac{1}{\alpha_\mathrm{S}(k_J^2)}\right\} \,, \\
\mathcal{S}_\mathrm{ggg} &= \frac{C_\mathrm{A}}{\pi b_0^2}\left\{\ln\left(\frac{\alpha_\mathrm{S}(\mu_J^2)}{\alpha_\mathrm{S}(k_J^2)}\right)\left[\frac{1}{\alpha_\mathrm{S}(R_\mathrm{fj}^2 k_J^2)} - \frac{11b_0}{12}\right] + \frac{1}{\alpha_\mathrm{S}(\mu_J^2)} - \frac{1}{\alpha_\mathrm{S}(k_J^2)}\right\} \,, \\
\mathcal{S}_\mathrm{gqq} &= \frac{T_\mathrm{R}}{3\pi b_0}\ln\left(\frac{\alpha_\mathrm{S}(\mu_J^2)}{\alpha_\mathrm{S}(k_J^2)}\right) \,.
\end{aligned} \tag{2.3}$$

Here μ_J is the jet mass, k_J is the jet transverse momentum, and $b_0 = (33 - 2n_\mathrm{f})/(12\pi)$.

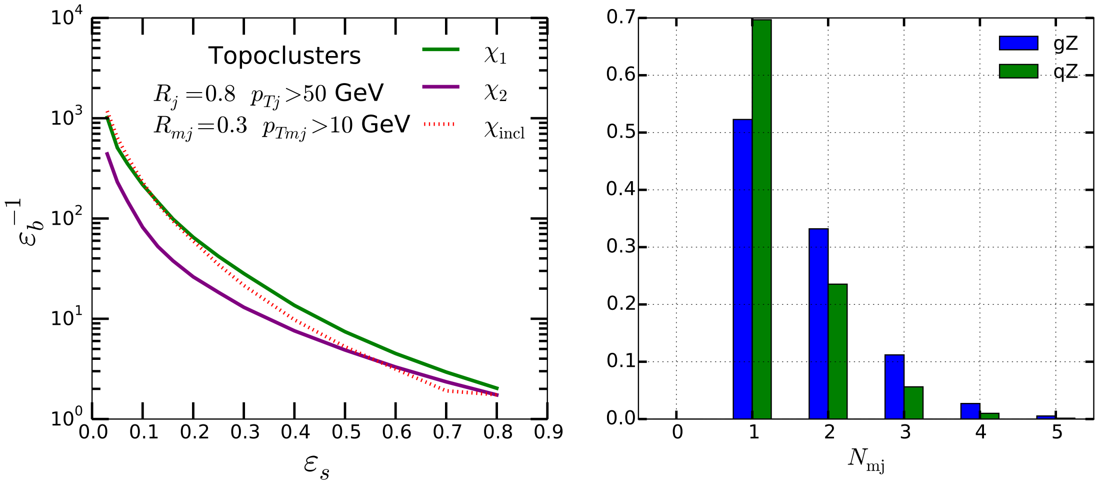

Figure 1. Left: quark (signal) vs gluon (background) ROC curves for χ with exactly one or exactly two microjets. Right: microjet multiplicity distribution.

In the case that we evaluate χ with simply the whole jet as the single microjet, we see that χ is a function of only two variables, the jet mass μ_J and the jet transverse momentum k_J. The function $\ln \chi$ is an approximation to the likelihood ratio

$$\ln L(q,g) = \ln P_{\rm MC}(\mu_J^2, k_J^2|q) - \ln P_{\rm MC}(\mu_J^2, k_J^2|g) \ .$$

If we use only the two variables μ_J^2 and k_J^2 to describe fat jets in each event, then $\ln L(q,g)$ provides the optimum way to distinguish quark jets from gluon jets as long as $P_{\rm MC}(\mu_J^2, k_J^2|q)$ and $P_{\rm MC}(\mu_J^2, k_J^2|g)$ provide accurate representations of nature. Thus one way to test whether the shower deconstruction variable χ is doing a good job is to construct the $\ln L(q,g)$ and compare $\ln \chi$ to $\ln L(q,g)$.

To build the likelihood function $L(q,g)$, we use the normalized (μ_J^2, k_J^2) histogram for the leading jets in $Z+q$ and $Z+g$ events. Then the likelihood in each bin is the ratio of the probability between the quark and gluon samples for that bin. However, the latter are strongly influenced by statistical fluctuations. We attempt to ameliorate this by "spreading" the probability of each bin. We use the gaussian kernel-density estimator [34] to smear the probability contained in each bin into a 2-dimensional gaussian distribution with the same normalization. The volume and mean of the gaussian kernel is fixed by the data, but the standard deviation is a free parameter that determines the "smoothing" effect. Even though the best way to determine this bandwidth parameter is through a cross-validation metric, we choose the parameter by visual comparison with the histograms. This leads to the distributions and contours in figure 2. The axes represent our two variables, μ_J^2 and k_J^2. In the bottom figure, we overlay three plots. The first is a scatter plot for the events in $Z+q$ jets and in $Z+g$ jets. The second, in yellow, is plot of contour lines of $\ln L(q,g)$ (after smoothing

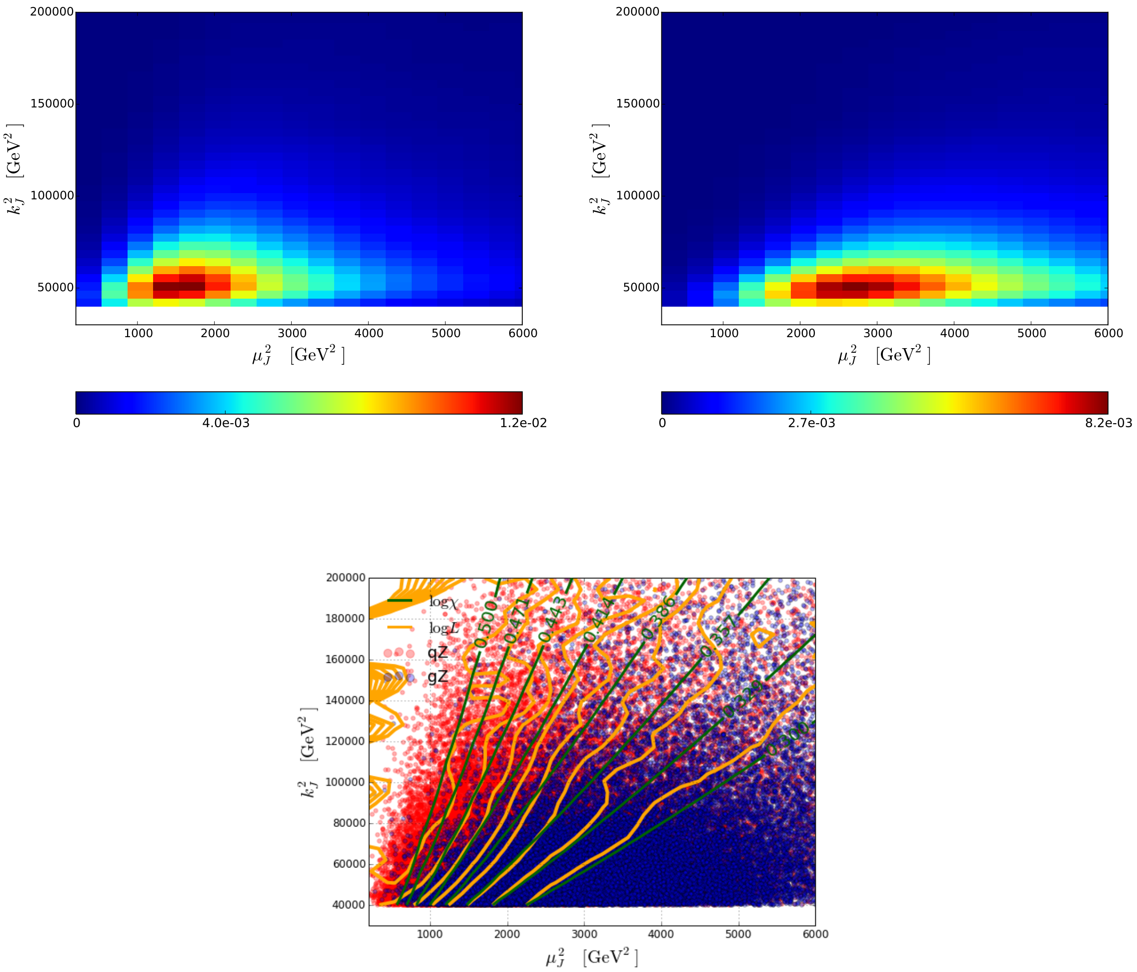

Figure 2. Gaussian kernel-density estimate of the $R = 0.4$ leading jets' mass and transverse momentum distribution in $Z + q$ (left) and $Z + g$ (right) events. In the bottom plot we overlay a scatter plot of the two distributions, contours of the likelihood derived from the gaussian kernel-density estimator and another contour plot of the shower deconstruction variable χ.

as described above). The third, in green, is a plot of contour lines of $\ln\chi$. We conclude that $\ln\chi$ is a reasonably good approximation to $\ln L(q, g)$.

In the analyses that follow, we mostly apply shower deconstruction to smaller, $R = 0.4$, fat jets, taking massive topoclusters as the input and using just one microjet, which is then equal to the whole fat jet.

2.3.2 Energy correlation functions

We now turn to an established family of observables with the potential to distinguish between quark and gluon jets: energy correlation functions and ratios derived from these functions

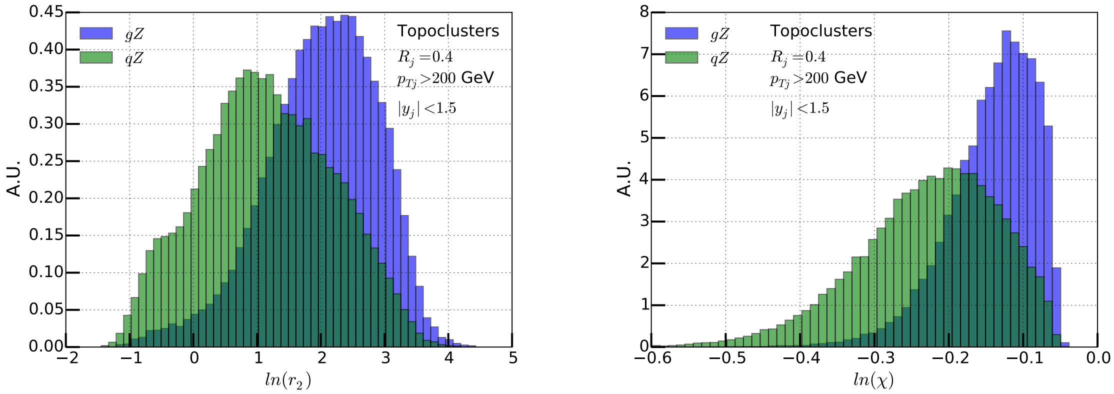

Figure 3. Distributions of r_2 (left) and $\ln(\chi)$ (right) in $Z+$jet events. The leading jet with $|y_j| < 1.5$ is reconstructed from massive topoclusters.

[10] [35]. The energy correlation functions are defined by

$$\begin{aligned}
&ECF(0,\beta) = 1, \\
&ECF(1,\beta) = \sum_{i\in J} p_{T,i}, \\
&ECF(2,\beta) = \sum_{i<j\in J} p_{T,i}p_{T,j}\left(R_{ij}\right)^{\beta}, \\
&ECF(N,\beta) = \sum_{i_1<i_2<..<i_n\in J}\left(\prod_{a=1}^{N} p_{T,i_a}\right)\left(\prod_{b=1}^{N-1}\prod_{c=b+1}^{N} R_{i_b i_c}\right)^{\beta},
\end{aligned} \tag{2.4}$$

From these, we can define the ratios

$$\begin{aligned}
r_N^{(\beta)} &= \frac{ECF(N+1,\beta)}{ECF(N,\beta)}, \\
C_N^{(\beta)} &= \frac{r_N^{(\beta)}}{r_{N-1}^{(\beta)}} = \frac{ECF(N+1,\beta)ECF(N-1,\beta)}{ECF(N,\beta)^2}.
\end{aligned} \tag{2.5}$$

The sums run over the constituents i of the jet J. We tested several jet shapes from this family (r_0, r_1, r_2, C_1, C_2). We also examined the variable D_2, defined in [35], and N-subjettiness variables [8] (τ_1, τ_2, τ_2/τ_1, τ_3/τ_2) with the angular exponent in all cases set to $\beta = 0.2$ for quark/gluon tagging, as suggested by the authors. Of those, C_1, r_1, and r_2 provided the best background rejection. If we express C_1, and r_2 explicitly using equations 2.4 and 2.5 we find

$$C_1 = \frac{\sum\limits_{i<j\in J} p_{T,i} p_{T,j} \left(R_{ij}\right)^{0.2}}{\sum\limits_{i,j\in J} p_{T,i} p_{T,j}},$$

$$r_2 = \frac{\sum\limits_{i<j<k\in J} p_{T,i} p_{T,j} p_{T,k} \left(R_{ij} R_{ik} R_{kj}\right)^{0.2}}{\sum\limits_{i<j\in J} p_{T,i} p_{T,j} \left(R_{ij}\right)^{0.2}}. \tag{2.6}$$

It is evident that the numerator of C_1 is larger if the radiation within the jet is split evenly between two or more distinct directions than if most of the energy is clustered within a small angular area. Therefore, C_1 is differentiates between 1-prong and 2-prong jets. The variable r_2 is larger if the radiation is localised in three directions and smaller for 2-prong and 1-prong jets.

The justification for the relatively small angular exponent comes from eq. (3.22) in [10]. The authors find a power law relation between the cumulative distributions of the C_1 variable for gluon and quark jets. A small β increases the magnitude of the power that relates the two distributions, thereby directly contributing to a better ROC curve. Note, however, that perturbative splitting probabilities have singularities at $R_{ij} = 0$. Thus the positive powers of R_{ij} are needed to keep the observables from being infrared unsafe against collinear splittings. With a power $\beta = 0.2$, our observables are technically infrared safe, but they are quite sensitive to infrared effects.

As a result of the asymmetry in the quark and gluon-jet distributions in figure 3, we find a different ROC curve for quark compared to gluon tagging[*]. For example, if we want to tag a quark and impose a cut on $\ln(\chi(q,g)) > 0.3$, we achieve $\varepsilon_s \simeq 0.21$ and $\varepsilon_b \simeq 0.017$. If we instead tag a gluon by requiring $\ln(\chi(g,q))$ to be bigger than a specific value, for $\varepsilon_s \simeq 0.21$ we find only $\varepsilon_b \simeq 0.05$.

A preliminary study of quark tagging with energy correlation variables uncovers some trends. As expected from the discussion in [10], we find that the variable C_1 is favored over r_2 over a large variety of jet parameters as long as the jets are reconstructed from hadrons. This can be seen in the bottom rows of figures 4 and 5, where its background fake rate is about

[*] According to eq. (3.7) in [10], if we were to perform quark tagging using C_1, the background fake rate as a function of the signal efficiency would be given by

$$\varepsilon_b(\varepsilon_s) = \varepsilon_s^{C_A/C_F} = \varepsilon_s^{2.25} \,. \tag{2.7}$$

Thus the gluon fake rate at 50% quark efficiency is $\varepsilon_b(0.5) \approx 0.21$. If we were to do the opposite and tag gluon jets at the expense of quark jets, then we would have to make the cut in the opposite direction of the C_1 distribution. Using the same relation between quark and gluon acceptances, we conclude that, when we retain 50% of the gluon jets in a sample, the fake rate from quark jets is $1 - (1-0.5)^{\frac{1}{2.25}} \approx 0.27$. Therefore, the same discriminating variable can perform differently depending on the type of tagging we would like to do. This asymmetry is strongly in favour of quark tagging for all of the variables that we study, as will become evident in the following sections.

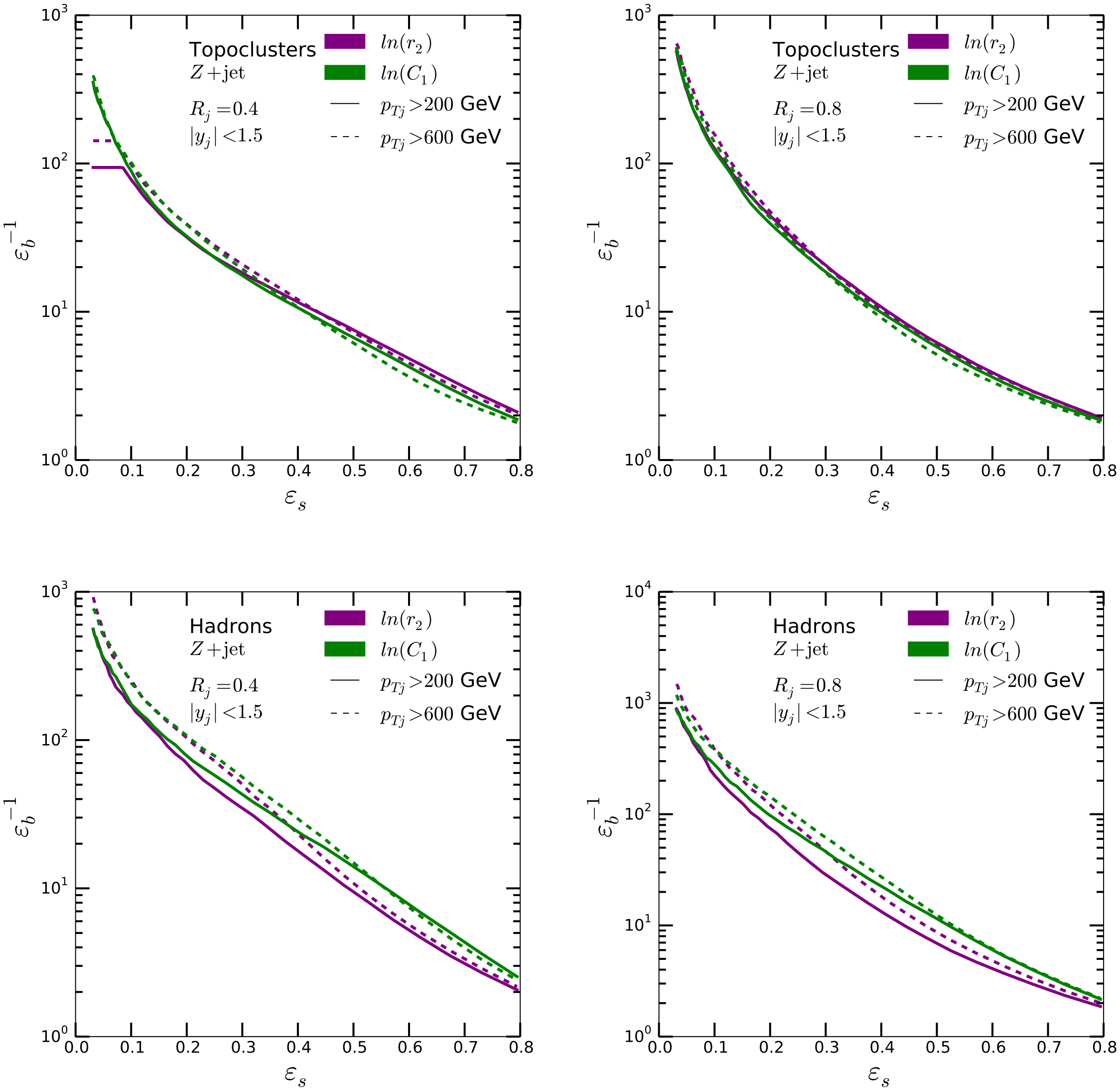

Figure 4. ROC plots comparing r_2 and C_1 performance at different jet p_T. The top row uses massive topoclusters as inputs and the bottom uses hadrons. The left (right) column uses jets with small (large) radius.

70% to 60% of that obtained with r_2 at moderate signal efficiency. This difference diminishes at small signal efficiency. A common trend among the energy correlation variables is that increasing the radius of the jet reduces the performance at moderate and large ε_s, but leads to improvement at low signal efficiency. This effect is true for any jet type as can be seen in the four plots of figure 5. Another trend in figure 4 is that for jets built from hadron inputs, a larger p_T limit increasingly improves background rejection as the signal cut becomes more stringent. This effect does not translate to topocluster inputs where the discrimination of the

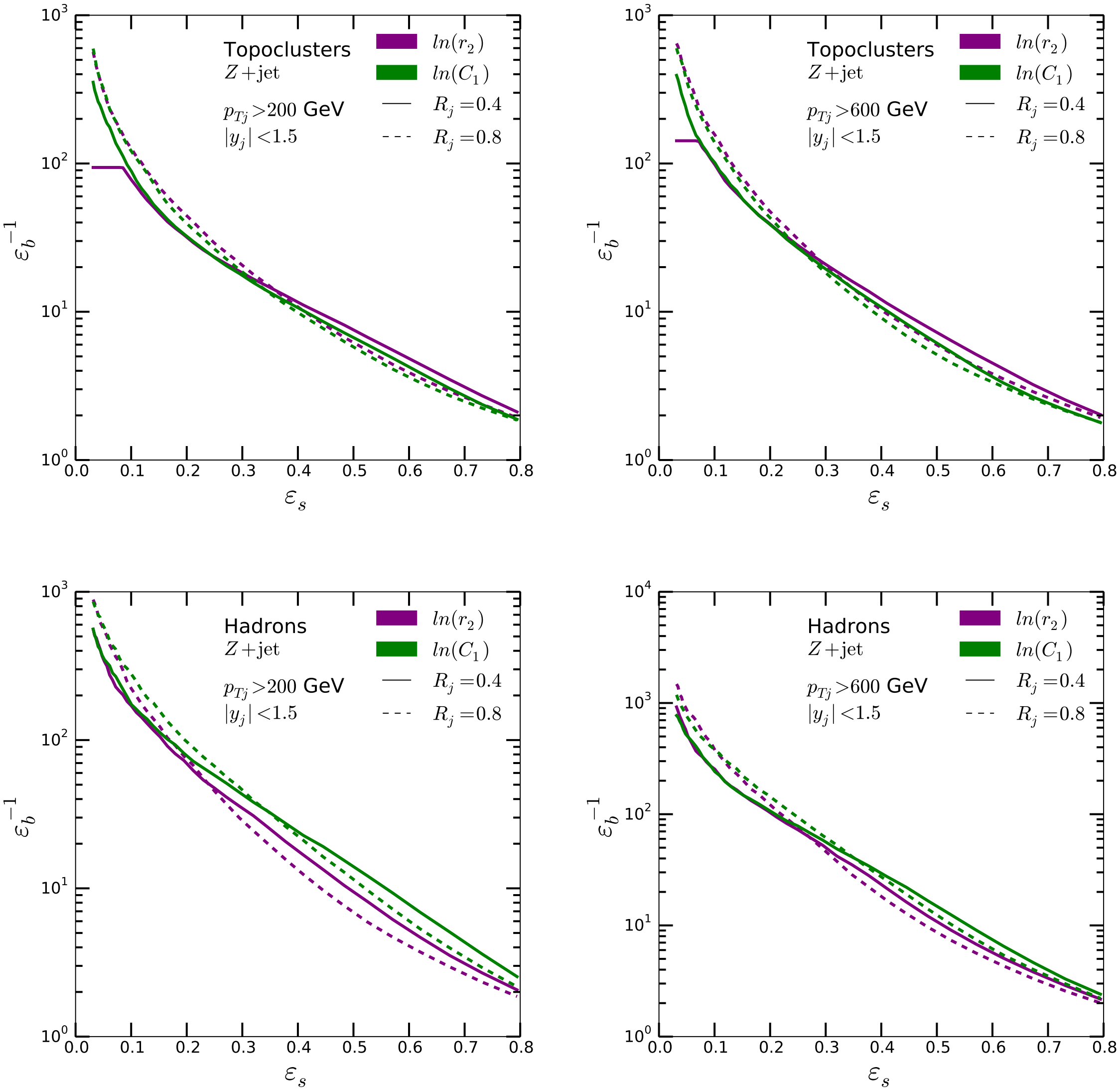

Figure 5. ROC plots comparing r_2 and C_1 performance at different jet radii. The top row uses massive topoclusters as inputs and the bottom uses hadrons. The left (right) column uses jets with small (large) boost.

energy correlation variables remains largely independent of the jet's transverse momentum.

3 Comparisons of tagging results

In this section, we compare methods for distinguishing quark jets from gluon jets.

We begin in figure 6 with a study of the dependence of four observables on the choice of input objects: hadrons, tracks, massless topoclusters, and massive topoclusters. In each

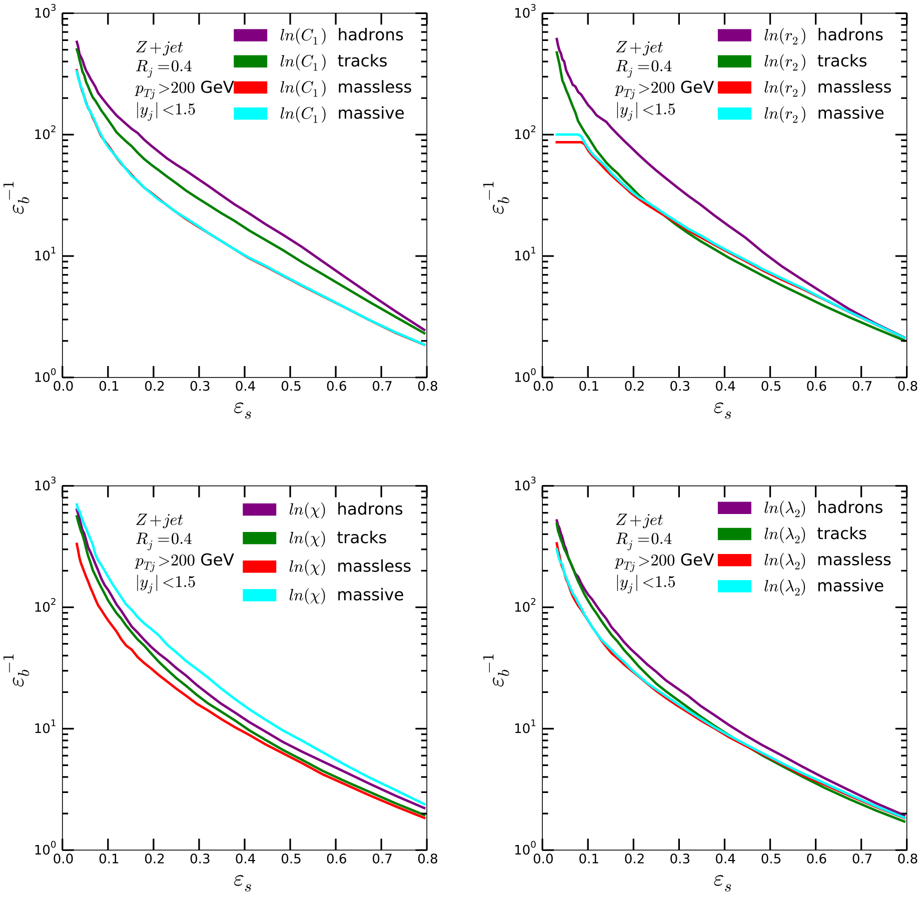

Figure 6. ROC curves of the leading jet with $|y| < 1.5$ for C_1 (upper left), r_2 (upper right), χ (lower left), λ_2 (lower right) and using hadrons, charged tracks, massless and massive topoclusters as inputs.

panel of figure 6, we show the dependence on input objects for one observable, C_1, r_2, χ from shower deconstruction with a single microjet, and the angularity variable λ_2 [11] defined by

$$\lambda_2 = \sum_{i \in J} p_{T,i}\, \theta_i^2 \Bigg/ \left(\sum_{i \in J} p_{T,i} \right) . \tag{3.1}$$

If the input constituents i are massless, λ_2 is approximately $2M_J^2/p_{T,J}^2$, where M_J is the jet mass. We show λ_2 because it is rather similar to χ if the input objects are all massless. However, χ is sensitive to the masses of the input objects while λ_2 is not. The ROC curves we show are obtained from distributions like the ones in figure 3 by swiping a cut from one end to the other.

All variables show some dependence on the input objects. Hadrons give the best results for C_1 and r_2, although detecting neutral as well as charged hadrons is not as realistic as the other input choices. After that, C_1 does best with tracks, while all of the other input choices work equally well for r_2. The variable λ_2 gives results that are rather insensitive to the choice of inputs, and not sensitive at all to the choice between massive and massless topoclusters. In contrast, the results for χ are significantly better with massive topocluster inputs than with massless topocluster inputs. This is to be expected because the topocluster mass μ_J is one of the variables used in the calculation of χ in eq. (2.3). With massless topoclusters as input, we are forced to set μ_J to a minimum value, $\mu_J = 1$ GeV, but this loses information. Perhaps surprisingly, χ works better with massive topocluster inputs than with all hadrons as inputs. This is because our definition of massive topoclusters drops topoclusters with $p_\mathrm{T} < 1$ GeV, on the grounds that such topoclusters would be experimentally unobservable. Dropping these low p_T topoclusters also helps to suppress unwanted contributions from initial state radiation, making χ more sensitive to the distinguishing features of quark jets compared to gluon jets.

We compare directly λ_2 to χ in figure 7. It is evident that shower deconstruction with massive topoclusters is better than the angularity variable. The latter is equivalent to the squared ratio between the jet mass and p_T as long as the input objects are massless and nearly collinear. The former condition is not satisfied in our case; therefore, we add the explicit ratio as a separate variable in the plot. Although much better than λ_2 it still performs worse than shower deconstruction.

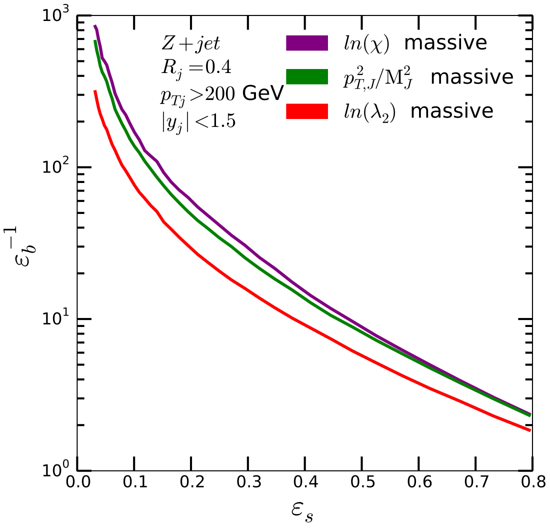

Figure 7. ROC curves of the leading jet with $|y| < 1.5$. We compare χ to λ_2 and a simple squared ratio of the jet transverse momentum and mass using massive topoclusters as inputs.

We turn next to a comparison of several observables that can be used for quark-gluon discrimination. Here, and in the studies that follow, we use massive topocluster inputs. The ROC curves for the observables are shown in figure 8. For shower deconstruction, we use just one microjet equal to the whole fat jet. Shower deconstruction χ has the best ROC curve. However, there is no dominant jet-shape or energy correlation function variable. Instead, there is a tier of closely spaced ROC curves. The top tier contains [r_2, r_1, C_1, τ_1, τ_2] and spreads within a band of about $\Delta\varepsilon_b \approx 20\%$ across the entire ε_s range. The ratio r_2 consistently performs better at moderate and large signal efficiency and remains competitive at small efficiency. Therefore, to the benefit of clarity of the results we are going to present, we believe it is acceptable to compare our choice of χ with r_2.

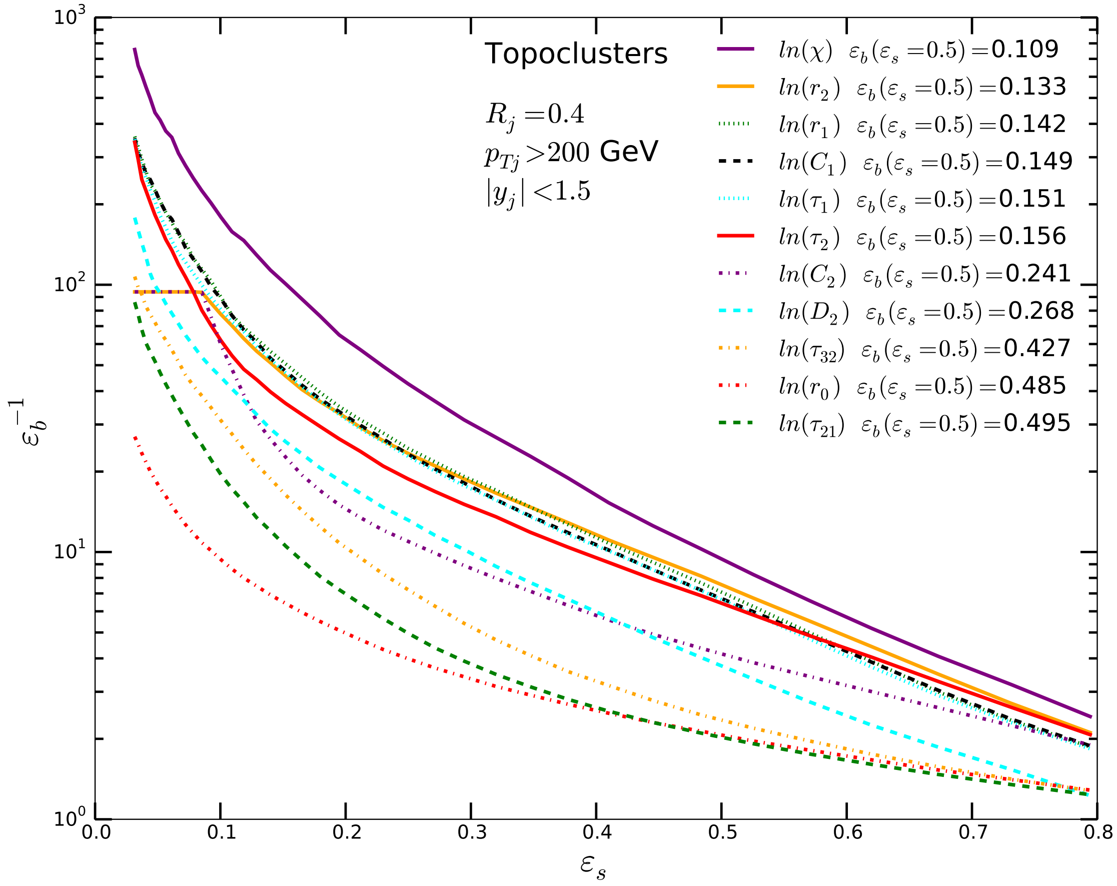

Figure 8. ROC curves for all distributions for quark tagging of Z + jet events. Leading jet with $|y| < 1.5$ reconstructed from massive topoclusters.

In figure 9 we show the ROC curves for the observables χ and r_2 for quark tagging (left) and gluon tagging (right) respectively. It is immediately apparent that quark tagging performs much better than gluon tagging, as already suggested by the analytic approximation of [10] and the discussion in section 2.3.2. At small efficiencies the gluon rejection in the left plot is four times better than the quark rejection on the right for shower deconstruction and two times better for r_2. One might anticipate this trend by looking at the probability

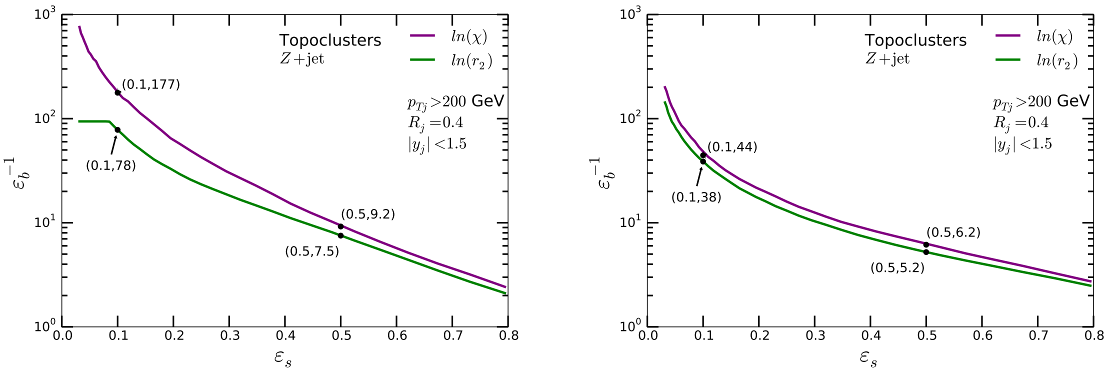

Figure 9. Left: ROC curves for quark tagging and gluon rejection from $Z+$jet events. Right: ROC curves for gluon tagging and quark rejection from $Z+$jet events. The leading jet with $|y| < 1.5$ is reconstructed from massive topoclusters.

densities of the variables. It is true for both observables, although more obvious for χ, that the quark distribution drops off slower at the gluon-like region end (large values) than the gluon distribution at the quark-like end (low values). This asymmetry allows for the substantial gluon rejection at small quark efficiency. Another feature is that the single-branch χ performs better than r_2 across the entire signal efficiency range in both quark and gluon tagging. For quark tagging it is about 20% better at moderate efficiencies and about a factor of two better at low efficiency. The difference is notably smaller when we attempt gluon tagging and almost disappears at low efficiency if we replace r_2 with a better performing energy correlation variable at that efficiency region. An obvious feature, although in a region that we do not explore, in the r_2 ROC curve is the plateau at $\varepsilon_s < 0.1$. It is an artefact from binning of jets on which the variable cannot be defined. The ratio r_2 needs at least 3 jet constituents. The condition is not always met with $R = 0.4$ jets reconstructed from topoclusters. More careful treatment of this bin can remove the plateau. It has to be noted that the energy correlation and N-subjettiness variables are used without optimisation with the recommended value $\beta = 0.2$ for quark and gluon tagging. Hence, there might be room for further improvements.

The results in figure 9 are obtained from jets with $p_T > 200\,\text{GeV}$. Collisions at the LHC can provide sufficient energy for much more boosted jets, either from a heavy particle decay or from a recoil in a high p_T event. In figure 10 we see the effect on quark tagging from increasing the jet transverse momentum. While we saw in figure 4 that increasing the jet p_T beyond 200 GeV has little or no effect on energy correlation variables, there is a distinct improvement in quark tagging with shower deconstruction as the jet gets more boosted. Moreover, the improvement is significant at 50% signal efficiency (40% better background rejection) and it steadily widens the difference between the χ and r_2 performance, leading to a factor of three better gluon rejection by χ than r_2 at $\varepsilon_s = 0.1$.

In the comparisons presented so far, we focused on central jets with rapidity $|y_j| < 1.5$.

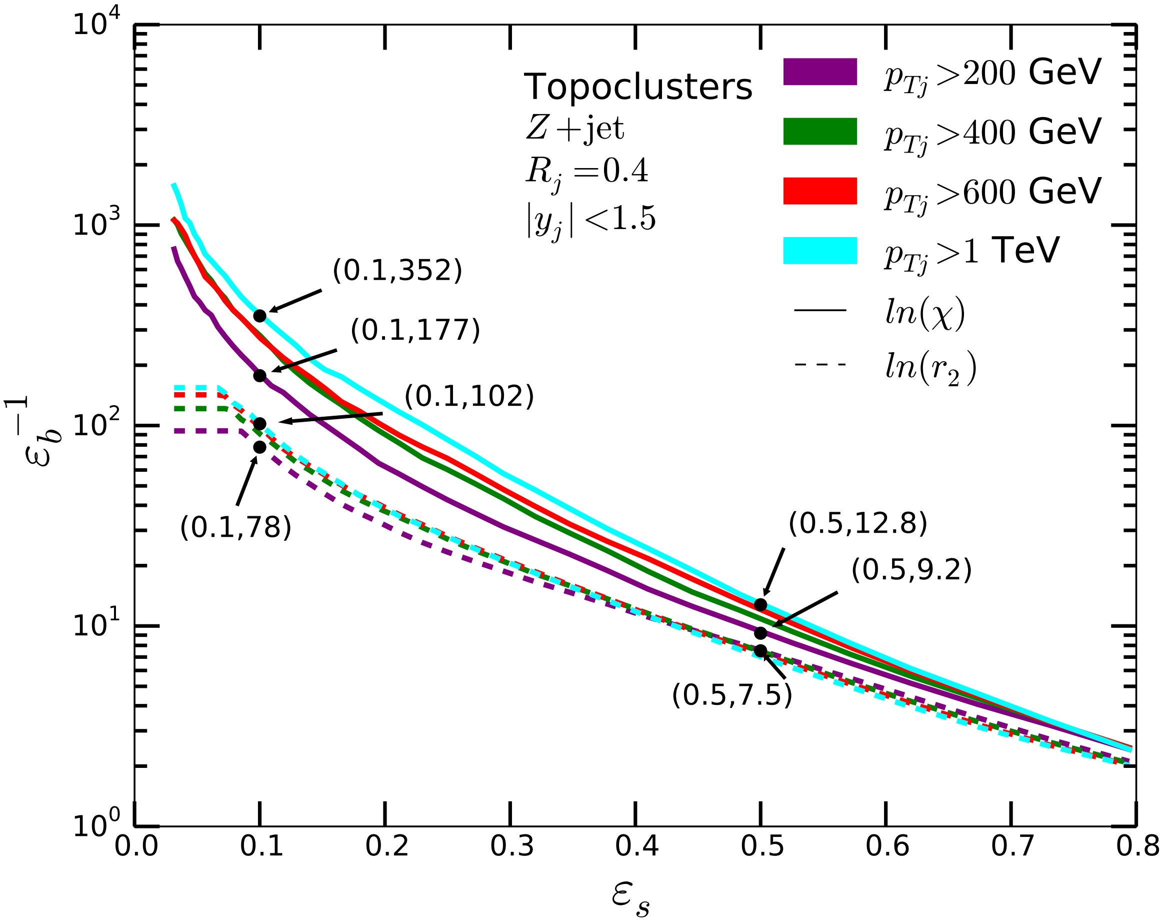

Figure 10. ROC curves for all p_T bins for quark tagging of Z + jet events with χ and r_2. The leading jet with $|y| < 1.5$ is reconstructed from massive topoclusters. The solid lines correspond to $\ln(\chi)$ of shower deconstruction and the dashed lines to the energy correlation function $\ln(r_2)$.

We can ask what happens when we extend the range of jet rapidity to $|y| < 2.5$. The results are shown in figure 11. For jets with $p_T > 200$ GeV, the ROC curve for quark tagging using r_2 is changed very little when the jet rapidity window is widened. However, ROC curve for quark tagging using χ becomes worse. This behavior warrants further investigation. If we look at the same question for jets with $p_T > 1$ TeV, then the effect of widening the rapidity window goes away. This may be because there are not many jets with $p_T > 1$ TeV and high rapidity.

We next study the effect on quark-gluon discrimination when we increase the radius of the fat jet from $R_{\rm fj} = 0.4$ to $R_{\rm fj} = 0.8$. For the larger fat jet size, we try two versions of shower deconstruction. In the first version, we construct χ using only one microjet, equal

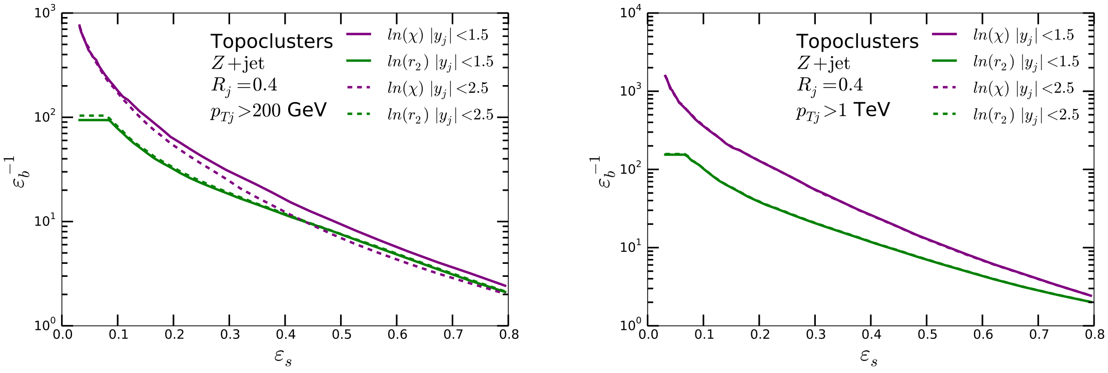

Figure 11. Effect of changing the rapidity window. The left panel shows ROC curves for quark tagging and gluon rejection from Z + jet events for massive topocluster jets with $p_T > 200$ GeV for two choices of the rapidity window. The right panel shows the same comparison for $p_T > 1$ TeV.

to the fat jet, as we have done in the previous studies with the smaller fat jet size. In the second version, we use the complete shower deconstruction algorithm [16–18] as described in section 2.3.1. The microjets are Cambridge-Aachen jets with $R_{mj} = 0.1$ and $p_{Tmj} > 5$ GeV. We denote the corresponding likelihood ratio by χ^*.

We compare ROC curves for r_2 and χ in in the left plot of figure 12. We see that the ROC curve for r_2 improves in the lower half of the ε_s range and diminishes somewhat in the upper half of the range as the fat jet radius increases. However, for most of the ε_s range, the ROC curve for the one-microjet version of χ becomes worse with a fatter fat jet. For $R_{\rm fj} = 0.8$, we compare ROC curves for r_2 and χ^* in right plot of figure 12. We find that full shower deconstruction performs better than r_2 across the whole range of signal efficiencies.

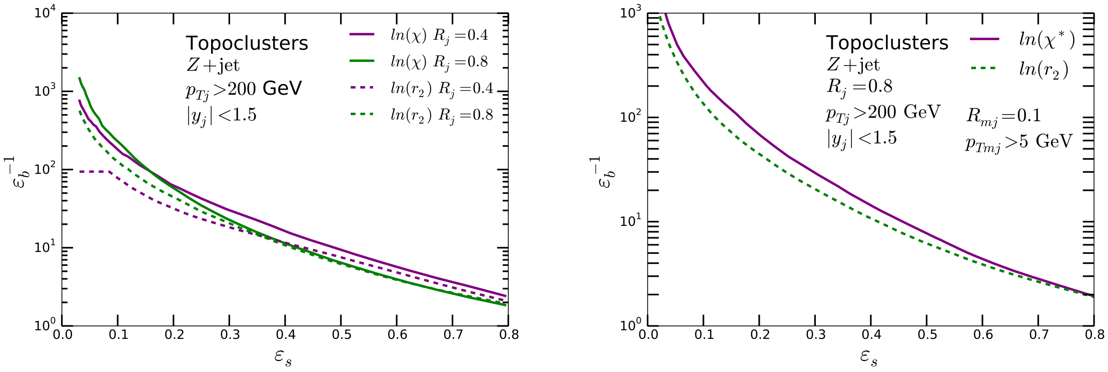

Figure 12. Effect of changing the fat jet radius. The left panel shows ROC curves for $\ln(\chi)$ and $\ln(r_2)$ from $R = 0.4$ and $R = 0.8$ Cambridge-Aachen jets built from massive topoclusters. The right panel shows ROC curves from $R = 0.8$ jets for $\ln(r_2)$ and full shower deconstruction ($\ln(\chi^*)$). The microjets for χ^* are Cambridge-Aachen jets with $R_{mj} = 0.1$ and $p_{Tmj} > 5$ GeV.

4 Sensitivity to the underlying process and parton shower

If we want to use quark-gluon discrimination in a search for new physics or a measurement of Higgs properties, we need to know the ROC curves for the observables we use as accurately as possible. Otherwise, the measurements will suffer from substantial systematic uncertainty. We can imagine calibrating the ROC curves by comparing experiment to results from event generators for known Standard Model processes. For this to work, we need to be sure that the performance of the observables we use does not depend on the underlying hard process. However, it was shown in [36, 37] that jet observables may depend on the event's colour flow. Such a conclusion was reached in [13, 14] also for quark and gluon tagging specifically. Thus we need to check whether this is the case for the observables that we have studied.

In figure 13, using Pythia 8 events, we compare the χ ROC curve for tagging quark jets in Z + jet events to that for dijet events. There is hardly any difference. We do the same for r_2 and again find hardly any difference. When compared to the difference between the χ and r_2 methods, it becomes evident that quark tagging with either is reliable for jets from different hard processes. Even though we only show the results with a single jet definition, we have confirmed it for jets with larger transverse momentum as well as larger radius parameter.

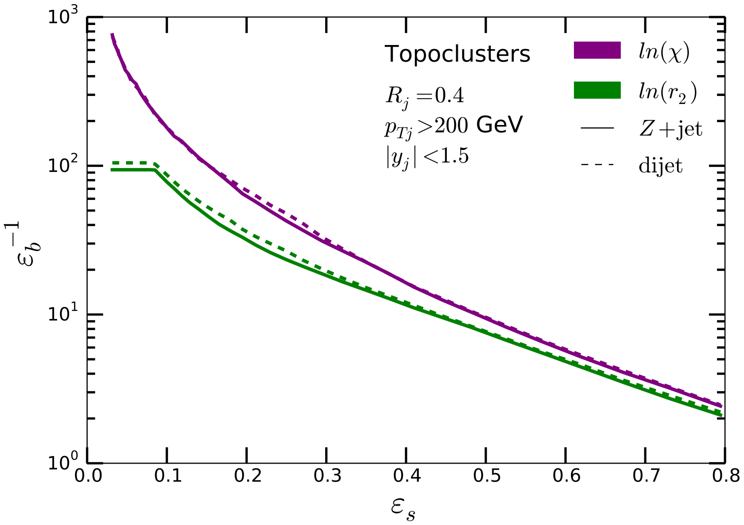

Figure 13. ROC curves for χ and r_2 applied to the leading jet of Z + jet and dijet events.

We can also ask whether existing parton shower Monte Carlos (with their default tunes) are sufficiently accurate to predict the ROC curves for χ and r_2. To answer this question, in figure 14 we compare the performance of these observables for Z + jet events generated by two different parton showers, Pythia 8 [19] and Sherpa [38]. For χ, we see that there is a rather substantial difference over much of the ϵ_s range. For r_2, the difference is not quite as large, but still not negligible.

What accounts for this difference? We can look at Section IV.5 of Ref.[13] for some insight. The authors of this study looked at quark-gluon discrimination in electron-positron annihilation using generalized angularity observables that are perturbatively infrared safe (and some that are not infrared safe, which we do not discuss here). An infrared safe observable is, by definition, not sensitive to parton splittings that are infinitesimally close to the soft or collinear singularities of perturbation theory. Nevertheless, such an observable can be sensitive to splittings that are at numerically small momentum scales. The study [13] examined quark gluon discrimination using several parton shower programs. When hadronization was turned off, there were very substantial differences in quark-gluon discrimination among the programs. It is not clear, at least to us, what characteristics of the parton shower programs led to greater or less quark-gluon discrimination. When hadronization was turned on, quark-gluon discrimination generally increased, suggesting quark jets hadronize quite differently from gluon jets and that this difference affects even nominally infrared safe observables. There were again very substantial differences in quark-gluon discrimination among the programs, but the differences now appeared to depend heavily on the hadronization model that the programs used.

Evidently, if parton shower event generators are to be useful in the analysis of quark-gluon discrimination, they need to better reflect the differences between quark jets and gluon jets, so that the parton shower dependence seen in figure 14 is reduced. We believe that this goal is achievable. It seems clear that hadronization has an important effect on variables that are sensitive to the difference between quark and gluon jets. The hadronization models in the shower program, as well as certain other parameters in the programs, can be tuned to match data. We note that the mixture of quark and gluon jets inevitably differ between jets in $\mathrm{p} + \mathrm{p} \to \mathrm{jet} + \mathrm{jet}$ and $\mathrm{p} + \mathrm{p} \to Z + \mathrm{jet}$. Thus, if the data used for Monte Carlo tuning include quark-gluon sensitive observables applied to jets in these two processes, then it seems at least plausible that the tuned shower programs would do better in describing both quark jets and gluon jets.

5 Application of quark-gluon tagging

5.1 Dark matter mono-jet

Searches for dark matter at the LHC have become a vibrant field of research in recent years [39–42]. If the dark matter particle communicates via a mediator with the Standard Model (SM) sector, given a small enough mass of the dark matter candidate, it can be produced at the LHC. While the dark matter particle is only weakly interacting with the detector material, its presence can be inferred indirectly by measuring its associated production with SM particles that carry large transverse momentum, e.g. jets. As shown in [1], the dominating backgrounds to high-p_T mono-jet searches are $Z+\mathrm{jet}$ and $W+\mathrm{jet}$. Due to the large invariant-mass final state and the structure of parton distribution functions, both of the gauge bosons are likely to be produced in association with a quark rather than a gluon, see table 1.

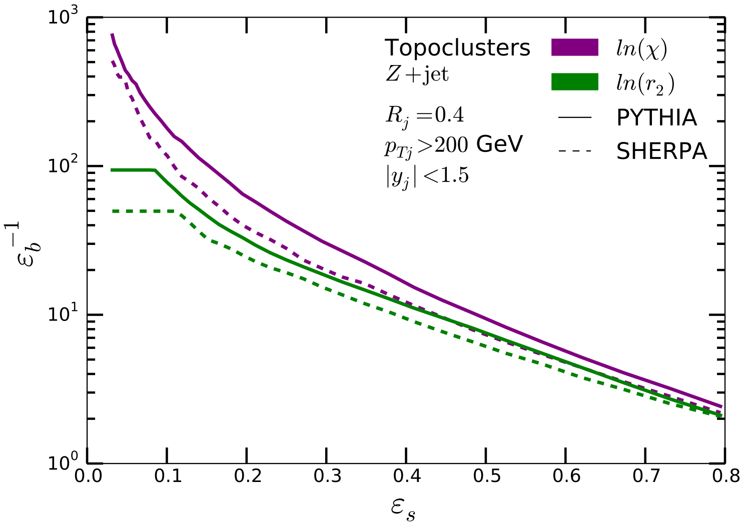

Figure 14. ROC curves for χ and r_2 applied to the leading jet of Z+jet events generated with Pythia and Sherpa.

Suppose the mediator is a scalar particle that couples to SM particles in agreement with the paradigm of minimal flavor violation, e.g. according to the Lagrangian [43, 44]

$$\mathcal{L}_{\text{scalar}} \supset -\frac{1}{2}m_{\text{MED}}^2 S^2 - g_{\text{DM}} S\,\bar{x}x - \sum_q g_{SM}^q S\,\bar{q}q - m_{\text{DM}}\bar{x}x\,. \tag{5.1}$$

The coupling constant g_{DM} denotes the interaction of the messengers with the dark sector particles. For simplicity we take the dark matter candidate to be a Dirac fermion x. The messenger's couplings to quarks are taken to be proportional to the corresponding Higgs Yukawa couplings $y_q = m_q/v$. As a reference and for definiteness we take $g_{\text{DM}} = y_{\text{DM}}$ and $g_{\text{SM}}^q = y_q$. Hence, the mediator couples preferentially to the top quark and decays for large g_{DM} to dark matter particles. In this case most of the jets produced in association with the dark matter particles are gluon-induced and the signal strength corresponds to the one of the SM Higgs boson with $m_H = 200$ GeV and BR$(h \to \bar{x}x) \simeq 1$, see table 1.

We use Pythia 8 to calculate signal S + jet and background Z + jet event rates. We assume the dark matter and mediator masses to be $m_{\text{DM}} = 20$ GeV and $m_{\text{MED}} = 200$ GeV respectively.

Even for such an optimistic scenario, the signal-to-background ratio S/B is small, i.e. $S/B \lesssim 0.07$, and systematic uncertainties on measurements , and systematic uncertainties on measurements with missing transverse energy are generically large [45]. The combined set of uncertainties in this channel, as shown in table 1 of [1], amounts to $5 - 10\%$. Hence, a signal-to-background ratio of less than 10% can render this search for cross sections we consider insensitive. Therefore, due to the lack of useful kinematic observables in this simple $2 \to 2$

$\sigma(\text{jet}+\text{MET})$ [fb]				
13 TeV LHC				
	$p_{T,j} > 250\text{GeV}$	$\lvert y\rvert < 1.5$	$\epsilon(\chi(g,q)) \simeq 50\%$	$\epsilon(\chi(g,q)) \simeq 10\%$
$pp \to (S \to \bar{x}x)j$	190	139	46.5	8.17
$pp \to (S \to \bar{x}x)g$	96.5	78.6	36.7	6.77
$pp \to (S \to \bar{x}x)q$	93.3	60	9.27	1.14
$pp \to (Z \to \bar{\nu}\nu)j$	2830	2170	430	62.2
$pp \to (Z \to \bar{\nu}\nu)g$	334	245	122	24.6
$pp \to (Z \to \bar{\nu}\nu)q$	2460	1890	299	40.3
S/B	0.067	0.064	0.11	0.13

Table 1. Production cross sections for a top-philic scalar mediator of mass $m_S = 200$ GeV that decays predominantly into dark matter, see eq. 5.1, and the dominant Standard Model background $Z+$jet at $\sqrt{s} = 13$ TeV.

process, applying a quark/gluon tagger can be vital to improve S/B beyond a necessary, signal cross-section dependent, threshold. After applying cuts on $\chi(g,q)$ corresponding to 50% and 10% we find $S/B \simeq 0.11$ and $S/B \simeq 0.13$ respectively. To transform this gain in S/B in a sensitivity improvement for dark matter searches, the systematic uncertainties from quark-gluon tagging should be small. This requires to address points raised in section 4 and, more specifically, the design of q/g-tagging approaches that show a stable performance for a wide class of processes.

5.2 Separation of gluon- and weak boson fusion in Hjj

Several ways have been proposed to separate the gluon-fusion from the weak boson-fusion process in dijet associated Higgs production $pp \to Hjj$. Among the methods proposed are rapidity gaps [2, 6], mini-jet vetos [46, 47], the matrix element method [48] and event shapes [49]. We add another arrow to the quiver by applying quark-gluon tagging.

To show the benefit of our approach we calculate the weak boson and the loop-induced gluon-fusion contributions to $pp \to Hjj$. The former allows to measure Higgs-gauge boson couplings and shows very small theoretical uncertainties [50–52].

The number of signal events depends on the sum of production processes p and Higgs decay channel $H \to YY$:

$$\sigma(H) \times \text{BR}(YY) \sim \left(\sum_p g_p^2\right) \frac{g_{HYY}^2}{\sum_{\text{modes}} g_i^2}, \tag{5.2}$$

assuming no interference between the different production mechanisms, where g denotes the Higgs couplings involved. The sum in the denominator runs over all kinematically accessible decay modes. Hence, the precision in measuring any Higgs boson coupling benefits from separating the production mechanisms.

$\sigma(pp \to Hjj)$ [fb]			
13 TeV LHC			
	$p_{T,j} > 50$ GeV, $\Delta R_{jj} > 2.0$	$\epsilon(\text{WBF}) \simeq 50\%$	$\epsilon(\text{WBF}) \simeq 10\%$
WBF $pp \to Hjj$	880	440	91
GF $pp \to Hjj$	900	180	15
GF $pp \to Hqq$	22	11	2.2
GF $pp \to Hgg$	450	61	1.8
GF $pp \to Hqg$	360	90	8
S/B	0.98	2.5	6.1

Table 2. LO production cross sections for gluon- and weak boson fusion of a Higgs boson with mass $m_H = 125$ GeV, separated into the respective partonic subprocesses. The two columns on the right show the results after applying a double quark tag with a combined efficiency of 50% and 10% respectively.

We generate the events using Sherpa, including the full top loop dependence and require at least two C/A $R = 0.4$ jets with $p_{T,j} > 50$ GeV, $|y_j| < 4.5$ and $\Delta R_{jj} \geq 2.0$. After the initial event selection cuts we already find a cross section ratio between gluon and weak boson fusion of ~ 1. For this analysis we do not decay the Higgs boson, as this approach can be applied irrespective of the decay mode of interest. Hence, we abstain from considering other Standard Model backgrounds which would depend strongly on the Higgs decay.

In table 2 we show by how much this ratio can be improved after applying a double quark tag on the two hardest jets of the event. We find that the gluon fusion contribution can be confidently reduced and even be rendered irrelevant if the WBF rates allow for tight quark tagging.

To give an example how quark-gluon tagging can improve Higgs coupling measurements, we can consider the process $pp \to jj(H \to ZZ^* \to 4l)$. In general this process is not necessarily considered a prime channel to measure the Higgs boson coupling to massive gauge bosons. Although the process is almost free from reducible backgrounds [53], due to efficient cuts on the four and two-lepton systems, the total rate after hard WBF cuts is quite small ($\ll 0.1$ fb). Using quark-gluon tagging allows us to retain a larger cross section while keeping at the same time gluon-fusion induced Higgs production under control. For the branching ratios of the Higgs and Z bosons we assume $\text{Br}(H \to ZZ^*) \simeq 2.62 \cdot 10^{-2}$ and $\text{Br}(Z \to l^+l^-) \simeq 0.06$, where l represents electrons and muons. The number of measured events is calculated as

$$N(\text{WBF}) \equiv \epsilon(\text{WBF}) \cdot \sigma(\text{WBF}) \cdot \text{Br}(H \to 4l) \cdot \mathcal{L}, \tag{5.3}$$

and

$$N(\text{GF}) \equiv \epsilon(\text{GF}) \cdot \sigma(\text{GF}) \cdot \text{Br}(H \to 4l) \cdot \mathcal{L}, \tag{5.4}$$

resulting for an integrated luminosity $\mathcal{L} = 1000\ \text{fb}^{-1}$ in $N(\text{WBF}) \simeq 83$ and $N(\text{GF}) \simeq 85$ before applying quark gluon tags on the accompanying jets. After applying quark-gluon tagging, for the working point $\epsilon(\text{WBF}) \simeq 50\%$ (10%) of table 2, we find $N(\text{WBF}) \simeq 42$ (9) and $N(\text{GF}) \simeq 17$ (1). While the application of quark-gluon tags do not improve on $S/\sqrt{S+B}$, for which we find $S/\sqrt{S+B} \simeq 6.4$ before and $S/\sqrt{S+B} \simeq 5.4$ after quark-gluon tagging with $\epsilon(\text{WBF}) = 50\%$ respectively. However, the combination of measurements including quark gluon tagging at different working points allows to improve the limit setting on deviations from Standard Model Higgs couplings.

The analytic dependence of the number of observed events on the coupling modifications can be parametrised as

$$N_{\text{tot}} = \Delta g_{hgg}^2 \Delta g_{hVV}^2 N(\text{GF}) + \Delta g_{hVV}^4 N(\text{WBF}), \tag{5.5}$$

where $\Delta g_i \equiv g_{i,\text{mod}}/g_{i,\text{SM}}$ and we assumed for simplicity that all Higgs-gauge boson couplings are modified the same way, i.e. $\Delta g_{hWW} = \Delta g_{hZZ} = \Delta g_{hVV}$. Note that interference between WBF and GF is highly suppressed [54].

In figure 15 we show the couplings that can be excluded to roughly 95% C.L. by requiring $|N_{\text{tot}} - N_{SM}|/\sqrt{N_{SM}} \lesssim 2$. While the sensitivity bands widen for smaller $\epsilon(\text{WBF})$, smaller gluon fusion contributions change the cross section dependence on g_{hgg}, thus, increasing sensitivity along otherwise blind directions of coupling combinations. That is, assuming that the experimental results obtained by using three different working points are all consistent with the Standard Model, one can exclude every combination of couplings that is outside of the intersection of the three bands in figure 15.

6 Conclusions

As illustrated in section 5, tagging jets as being likely quark initiated or likely gluon initiated can be used for separating signal from background in LHC events. In the earlier sections of this paper, we studied issues related to how such quark-gluon tagging can be performed.

Our studies suggest that, at least for the methods investigated, quark-gluon tagging can be effective, but has a substantial sensitivity to physics at rather small momentum scales. This is illustrated by the finding in figure 14 that if we seek to tag quark jets, then the background rejection factors obtained with events generated by standard Monte Carlo event generators differ according to which generator, Pythia or Sherpa, we use. The ROC curves obtained are qualitatively similar but have significant quantitative differences. Another finding, illustrated in figures 4 and 5, that points to the same conclusion is that different results are obtained by examining the jet substructure beginning with hadrons or beginning with simulated massive topoclusters. Starting with hadrons gives the most detailed view, while starting with topoclusters removes some of the information that comes from the final, infrared dominated, stages of hadronization. What we see is that including or not this infrared dominated information affects the results.

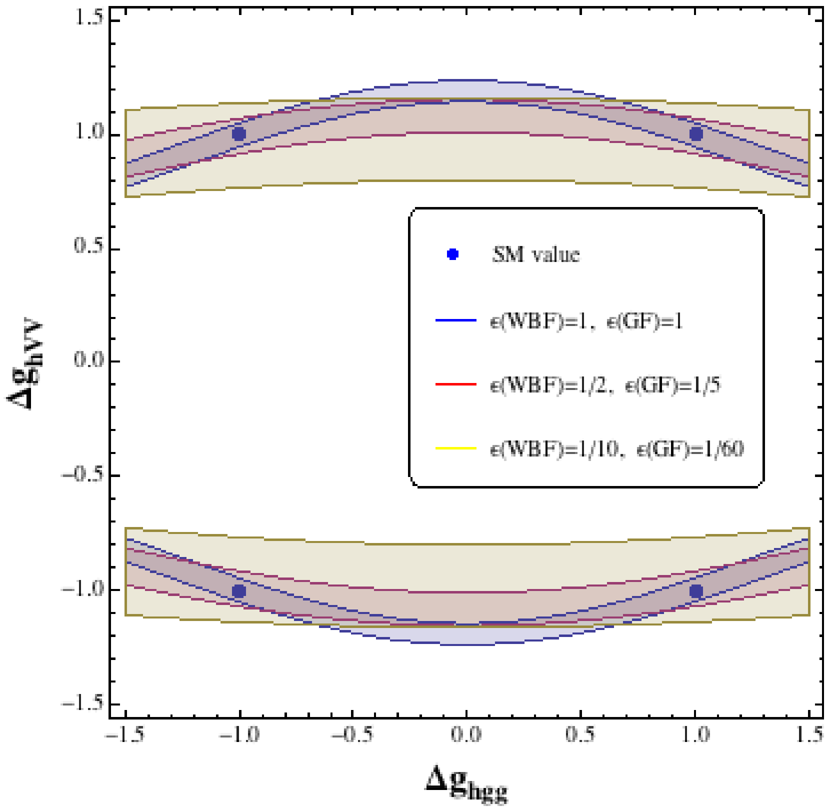

Figure 15. Sensitivity bands for the process $pp \to (h \to ZZ^* \to 4l)jj$ after applying quark-gluon tagging with three different working points, assuming a integrated luminosity of $\mathcal{L} = 1000\ \text{fb}^{-1}$. There is a four-fold ambiguity for the couplings g_{hVV} and g_{hgg}, for which the same number of events as in the Standard Model (corresponding to the point $g_{hVV} = 1$ and $g_{hgg} = 1$) are observed. Coupling modifications are defined as $\Delta g_i \equiv g_{i,\text{mod}}/g_{i,\text{SM}}$.

This tentative conclusion suggests that there is a tradeoff in using quark-gluon tagging between sensitivity to the signals that we are looking for and the reliability of the method. That is, we can improve background rejection and thus increase our chances of finding, say, a signal for new physics. However, we may induce a substantial systemic error in the calculation of the amount of background rejection. Of course, if we can measure the background rejection factor experimentally, this problem is ameliorated. To this end, it is encouraging that, when we try to tag quark jets, the background rejection factor seems to be quite independent of the hard scattering process that creates the jets, as illustrated in figure 13.

We examined several measures of jet substructure that bear on quark-gluon separation. The most realistic case is to apply these measures to simulated topoclusters rather than hadrons, both because topocluster results are likely to be less infrared sensitive and because they are more experimentally practical. In our studies, we retained the mass of each simulated topocluster rather than scaling the momentum so as to set the topocluster mass to zero. This goes beyond the method used by ATLAS, but it improves the quark-gluon separation for the shower deconstruction variable χ. Most of our studies concerned tagging fat jets with radius parameter $R_{\rm fj} = 0.4$. There we found that the variables r_1, r_2 and C_1 exhibited similar performances, as illustrated in figure 8. For other graphs, we chose r_2 as representative of these three. We compared r_2 to the shower deconstruction variable χ. Normally, shower deconstruction divides the fat jet into several smaller jets, called microjets. That is essential when seeking to find heavy particles that decay to several jets. However, in distinguishing quark from gluon QCD jets with a rather small cone size $R_{\rm fj} = 0.4$ for the fat jet, we found that it was better to simply apply the shower deconstruction calculation of χ to a single microjet, identical to the fat jet. The result, from figure 8, is that the ROC curve for χ shows better background rejection than that for r_2.

We examined quark-gluon discrimination also for fatter fat jets, with $R_{\rm fj} = 0.8$, as illustrated in figure 12. There we found that the shower deconstruction method with more than one microjets worked best. However, the improvement over the use of $R_{\rm fj} = 0.4$ fat jets was small.

We conclude, in general agreement with refs. [7–13], that using jet substructure measures to discriminate between quark initiated jets and gluon initiated jets can be helpful for distinguishing signals from backgrounds at the LHC. We have presented results that bear on the use of these methods, but a final judgement can only be reached by using these observables by ATLAS and CMS.

Acknowledgments. We thank Silvan Kuttimalai for support with Sherpa event generation. MS is supported in part by the European Commission through the "HiggsTools" Initial Training Network PITN-GA-2012-316704. DS is supported by the U.S. Department of Energy. DFL is supported by the Alexander von Humboldt Foundation.

References

[1] **CMS** Collaboration, V. Khachatryan *et. al.*, *Search for dark matter, extra dimensions, and unparticles in monojet events in protonproton collisions at $\sqrt{s} = 8$ TeV*, *Eur. Phys. J.* **C75** (2015), no. 5 235, [arXiv:1408.3583].

[2] Y. L. Dokshitzer, V. A. Khoze, and T. Sjostrand, *Rapidity gaps in Higgs production*, *Phys. Lett.* **B274** (1992) 116–121.

[3] D. L. Rainwater, D. Zeppenfeld, and K. Hagiwara, *Searching for $H \to \tau^+\tau^-$ in weak boson fusion at the CERN LHC*, *Phys. Rev.* **D59** (1998) 014037, [hep-ph/9808468].

[4] D. Zeppenfeld, R. Kinnunen, A. Nikitenko, and E. Richter-Was, *Measuring Higgs boson couplings at the CERN LHC*, *Phys. Rev.* **D62** (2000) 013009, [hep-ph/0002036].

[5] C. Englert, R. Kogler, H. Schulz, and M. Spannowsky, *Higgs coupling measurements at the LHC*, arXiv:1511.0517.

[6] V. Del Duca, W. Kilgore, C. Oleari, C. Schmidt, and D. Zeppenfeld, *Gluon fusion contributions to H + 2 jet production*, *Nucl. Phys.* **B616** (2001) 367–399, [hep-ph/0108030].

[7] S. Catani, G. Turnock, and B. R. Webber, *Jet broadening measures in* e^+e^- *annihilation*, *Phys. Lett.* **B295** (1992) 269–276.

[8] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **03** (2011) 015, [arXiv:1011.2268].

[9] J. Gallicchio and M. D. Schwartz, *Quark and Gluon Tagging at the LHC*, *Phys. Rev. Lett.* **107** (2011) 172001, [arXiv:1106.3076].

[10] A. J. Larkoski, G. P. Salam, and J. Thaler, *Energy Correlation Functions for Jet Substructure*, *JHEP* **06** (2013) 108, [arXiv:1305.0007].

[11] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, *Gaining (Mutual) Information about Quark/Gluon Discrimination*, *JHEP* **11** (2014) 129, [arXiv:1408.3122].

[12] B. Bhattacherjee, S. Mukhopadhyay, M. M. Nojiri, Y. Sakaki, and B. R. Webber, *Associated jet and subjet rates in light-quark and gluon jet discrimination*, *JHEP* **04** (2015) 131, [arXiv:1501.0479].

[13] J. R. Andersen *et. al.*, *Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report*, in *9th Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2015) Les Houches, France, June 1-19, 2015*, 2016. arXiv:1605.0469.

[14] **ATLAS** Collaboration, G. Aad *et. al.*, *Light-quark and gluon jet discrimination in pp collisions at* $\sqrt{s} = 7$ TeV *with the ATLAS detector*, *Eur. Phys. J.* **C74** (2014), no. 8 3023, [arXiv:1405.6583].

[15] **CMS Collaboration** Collaboration, *Performance of quark/gluon discrimination in 8 TeV pp data*, Tech. Rep. CMS-PAS-JME-13-002, CERN, Geneva, 2013.

[16] D. E. Soper and M. Spannowsky, *Finding physics signals with shower deconstruction*, *Phys. Rev.* **D84** (2011) 074002, [arXiv:1102.3480].

[17] D. E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, *Phys. Rev.* **D87** (2013) 054012, [arXiv:1211.3140].

[18] D. E. Soper and M. Spannowsky, *Finding physics signals with event deconstruction*, *Phys. Rev.* **D89** (2014), no. 9 094005, [arXiv:1402.1189].

[19] T. Sjostrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852–867, [arXiv:0710.3820].

[20] T. A. Collaboration, *The atlas experiment at the cern large hadron collider*, *Journal of Instrumentation* **3** (2008), no. 08 S08003.

[21] T. C. Collaboration, *The cms experiment at the cern lhc*, *Journal of Instrumentation* **3** (2008), no. 08 S08004.

[22] **ATLAS** Collaboration, G. Aad *et. al.*, *Jet energy measurement with the ATLAS detector in proton-proton collisions at* $\sqrt{s} = 7$ *TeV*, *Eur. Phys. J.* **C73** (2013), no. 3 2304, [arXiv:1112.6426].

[23] W. Lampl, S. Laplace, D. Lelas, P. Loch, H. Ma, S. Menke, S. Rajagopalan, D. Rousseau, S. Snyder, and G. Unal, *Calorimeter clustering algorithms: Description and performance*, .

[24] **ATLAS** Collaboration, G. Aad *et. al.*, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, arXiv:1603.0293.

[25] **CMS Collaboration** Collaboration, *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET*, Tech. Rep. CMS-PAS-PFT-09-001, CERN, 2009. Geneva, Apr, 2009.

[26] A. Katz, M. Son, and B. Tweedie, *Jet Substructure and the Search for Neutral Spin-One Resonances in Electroweak Boson Channels*, *JHEP* **03** (2011) 011, [arXiv:1010.5253].

[27] S. Schaetzel and M. Spannowsky, *Tagging highly boosted top quarks*, *Phys. Rev.* **D89** (2014), no. 1 014007, [arXiv:1308.0540].

[28] A. J. Larkoski, F. Maltoni, and M. Selvaggi, *Tracking down hyper-boosted top quarks*, *JHEP* **06** (2015) 032, [arXiv:1503.0334].

[29] M. Spannowsky and M. Stoll, *Tracking New Physics at the LHC and beyond*, *Phys. Rev.* **D92** (2015), no. 5 054033, [arXiv:1505.0192].

[30] **CMS Collaboration** Collaboration, *Jet Energy Corrections for Multiple Cone Sizes*, tech. rep., CERN, 2009. Geneva.

[31] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, *Better jet clustering algorithms*, *JHEP* **08** (1997) 001, [hep-ph/9707323].

[32] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross-sections in deep inelastic scattering*, in *Monte Carlo generators for HERA physics. Proceedings, Workshop, Hamburg, Germany, 1998-1999*, 1998. hep-ph/9907280.

[33] **ATLAS Collaboration** Collaboration, *Jet mass reconstruction with the ATLAS Detector in early Run 2 data*, Tech. Rep. ATLAS-CONF-2016-035, CERN, Geneva, Jul, 2016.

[34] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, 1992.

[35] A. J. Larkoski, I. Moult, and D. Neill, *Power Counting to Better Jet Observables*, *JHEP* **12** (2014) 009, [arXiv:1409.6298].

[36] S. Ask, J. H. Collins, J. R. Forshaw, K. Joshi, and A. D. Pilkington, *Identifying the colour of TeV-scale resonances*, *JHEP* **01** (2012) 018, [arXiv:1108.2396].

[37] K. Joshi, A. D. Pilkington, and M. Spannowsky, *The dependency of boosted tagging algorithms on the event colour structure*, *Phys. Rev.* **D86** (2012) 114016, [arXiv:1207.6066].

[38] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter, *Event generation with SHERPA 1.1*, *JHEP* **02** (2009) 007, [arXiv:0811.4622].

[39] M. Beltran, D. Hooper, E. W. Kolb, Z. A. C. Krusberg, and T. M. P. Tait, *Maverick dark matter at colliders*, *JHEP* **09** (2010) 037, [arXiv:1002.4137].

[40] J. Goodman, M. Ibe, A. Rajaraman, W. Shepherd, T. M. P. Tait, and H.-B. Yu, *Constraints on Light Majorana dark Matter from Colliders*, *Phys. Lett.* **B695** (2011) 185–188, [arXiv:1005.1286].

[41] P. J. Fox, R. Harnik, J. Kopp, and Y. Tsai, *Missing Energy Signatures of Dark Matter at the LHC*, *Phys. Rev.* **D85** (2012) 056011, [arXiv:1109.4398].

[42] J. Abdallah *et. al.*, *Simplified Models for Dark Matter and Missing Energy Searches at the LHC*, arXiv:1409.2893.

[43] M. R. Buckley, D. Feld, and D. Goncalves, *Scalar Simplified Models for Dark Matter*, *Phys. Rev.* **D91** (2015) 015017, [arXiv:1410.6497].

[44] P. Harris, V. V. Khoze, M. Spannowsky, and C. Williams, *Constraining Dark Sectors at Colliders: Beyond the Effective Theory Approach*, *Phys. Rev.* **D91** (2015) 055009, [arXiv:1411.0535].

[45] **ATLAS** Collaboration, M. Aaboud *et. al.*, *Search for new phenomena in final states with an energetic jet and large missing transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector*, arXiv:1604.0777.

[46] D. L. Rainwater and D. Zeppenfeld, *Observing $H \to W^*W^* \to e^\pm\mu\mp \not{p}_T$ in weak boson fusion with dual forward jet tagging at the CERN LHC*, *Phys. Rev.* **D60** (1999) 113004, [hep-ph/9906218]. [Erratum: Phys. Rev.D61,099901(2000)].

[47] B. E. Cox, J. R. Forshaw, and A. D. Pilkington, *Extracting Higgs boson couplings using a jet veto*, *Phys. Lett.* **B696** (2011) 87–91, [arXiv:1006.0986].

[48] J. R. Andersen, C. Englert, and M. Spannowsky, *Extracting precise Higgs couplings by using the matrix element method*, *Phys. Rev.* **D87** (2013), no. 1 015019, [arXiv:1211.3011].

[49] C. Englert, M. Spannowsky, and M. Takeuchi, *Measuring Higgs CP and couplings with hadronic event shapes*, *JHEP* **06** (2012) 108, [arXiv:1203.5788].

[50] T. Figy, C. Oleari, and D. Zeppenfeld, *Next-to-leading order jet distributions for Higgs boson production via weak boson fusion*, *Phys. Rev.* **D68** (2003) 073005, [hep-ph/0306109].

[51] M. Ciccolini, A. Denner, and S. Dittmaier, *Electroweak and QCD corrections to Higgs production via vector-boson fusion at the LHC*, *Phys. Rev.* **D77** (2008) 013002, [arXiv:0710.4749].

[52] M. Cacciari, F. A. Dreyer, A. Karlberg, G. P. Salam, and G. Zanderighi, *Fully Differential Vector-Boson-Fusion Higgs Production at Next-to-Next-to-Leading Order*, *Phys. Rev. Lett.* **115** (2015), no. 8 082002, [arXiv:1506.0266].

[53] **CMS Collaboration** Collaboration, *Measurements of properties of the Higgs boson and search for an additional resonance in the four-lepton final state at sqrt(s) = 13 TeV*, Tech. Rep. CMS-PAS-HIG-16-033, CERN, Geneva, 2016.

[54] J. R. Andersen and J. M. Smillie, *QCD and electroweak interference in Higgs production by gauge boson fusion*, *Phys. Rev.* **D75** (2007) 037301, [hep-ph/0611281].