Published for SISSA by Springer

Received: July 31, 2018
Revised: September 19, 2018
Accepted: September 23, 2018
Published: September 28, 2018

Top tagging: an analytical perspective

Mrinal Dasgupta,[a] **Marco Guzzi,**[b] **Jacob Rawling**[a] **and Gregory Soyez**[c]

[a] *Lancaster-Manchester-Sheffield Consortium for Fundamental Physics, School of Physics & Astronomy, University of Manchester, Manchester M13 9PL, U.K.*

[b] *Department of Physics, Kennesaw State University, Kennesaw, GA 30144, U.S.A.*

[c] *IPhT, CEA Saclay, CNRS UMR 3681, Université Paris-Saclay, F-91191 Gif-Sur-Yvette, France*

E-mail: mrinal.dasgupta@manchester.ac.uk, mguzzi@kennesaw.edu, jrawling@cern.ch, gregory.soyez@ipht.fr

Abstract: In this paper we study aspects of top tagging from first principles of QCD. We find that the method known as the CMS top tagger becomes collinear unsafe at high p_t and propose variants thereof which are IRC safe, and hence suitable for analytical studies, while giving a comparable performance to the CMS tagger. We also develop new techniques to identify three-pronged jet substructure, based on adaptations of the Y-splitter method and its combination with grooming. A novel feature of our studies, relative to previous calculations of two-pronged substructure, is our use of triple-collinear splitting functions combined with all-order resummation, which owes to the presence of two mass scales of the same order, m_t and m_W, in the signal jet. We carry out leading logarithmic resummed calculations for the various top-taggers, for both background and signal jets, and compare the results to those from parton showers. We also identify and comment on the main features driving tagger performance at high p_t and discuss the role of non-perturbative effects.

Keywords: Jets

ArXiv ePrint: 1807.04767

https://doi.org/10.1007/JHEP09(2018)170

JHEP09(2018)170

Contents

1 Introduction **2**

2 Tagger definitions **4**
- 2.1 The CMS top tagger and new methods 4
- 2.2 The Y_m-splitter method for top tagging 7

3 Analytical calculations at fixed-order **8**
- 3.1 Leading-order calculations in the soft-collinear limit 8
- 3.2 The triple-collinear limit of a QCD jet 11

4 Resummed calculation to all orders **13**
- 4.1 Y_m-splitter 15
  - 4.1.1 Calculation in pure soft and strongly-ordered limit 15
  - 4.1.2 Matching to the triple-collinear limit 17
  - 4.1.3 Y_m-splitter with grooming 18
- 4.2 `TopSplitter` and $\text{CMS}^{3p,mass}$ 21

5 Results **24**
- 5.1 Numerical impact of triple-collinear and resummation effects 25
- 5.2 Comparison to parton showers 27

6 Signal efficiency and performance **31**
- 6.1 Signal efficiency 31
- 6.2 Performance and non-perturbative effects 34

7 Conclusions **38**

A Collinear unsafety of the CMS tagger with no ΔR cut **40**

B Variants of the CMS and Y-splitter taggers **41**
- B.1 Definition of the variants 41
- B.2 Declustering with a ζ_{cut} or z_{cut} condition 41
- B.3 Minimum pairwise condition v. secondary declustering condition 42

C Analytic expressions for the radiators **42**

D Performance at lower energy **44**

JHEP09(2018)170

1 Introduction

Recent years have seen the field of jet substructure mature and develop into one of the key areas of current LHC phenomenology [1–11]. Amongst the numerous applications of substructure methods there are direct searches for new physics beyond the standard model [12–17], crucial studies of the Higgs sector of the standard model [18], precise determination of the top quark mass [19], testing high precision calculations for jets in QCD [20–24], and studies involving jets in heavy-ion collisions [25, 26]. Following the commencement of the LHC run 2 at 13 TeV, electroweak scale particles including the top quark can be extremely boosted, which means their hadronic decays will often result in a single jet. In such situations jet substructure studies have been proven to provide important input which is the key to effectively distinguishing signal from background as well as to improve resolution of signal mass peaks.

One of the most active areas within the field of jet substructure has been the study of boosted top quarks and several techniques are available to study boosted tops including various "top-taggers" [27–34], template tagging [35], shower deconstruction [36], jet shape variables such as N-subjettiness [37, 38], energy correlation functions [39–41] and multivariate methods exploiting machine learning [42–46]. The performance of these tools has been investigated in detail using studies based on Monte Carlo event generators. Many of the above mentioned methods are also increasingly used in experimental analyses at the LHC [47].

An alternative approach to traditional Monte Carlo studies of jet substructure has emerged and gained substantial ground in recent years [39, 48–54]. This new approach is based on directly using perturbative QCD calculations for jet substructure observables. Since the boosted regime with jet masses $m \ll p_t$ is a classic multi-scale problem, and one encounters the feature of large logarithms in p_t/m, perturbative calculations at fixed-order in α_s are not directly useful on their own, and one needs the techniques of analytic resummation to give a satisfactory description of substructure observables in the boosted limit.

Analytic resummed perturbative calculations have been shown to be powerful methods in learning about jet substructure techniques often yielding vital information about features that did not emerge in shower studies prior to the advent of the analytics. Amongst some of the benefits arising from analytical studies, one can list the discovery of flaws such as kinks and bumps in the jet mass spectrum with various taggers [48] which led to the emergence of improved tools [48, 50], the discovery of occasional issues with parton shower descriptions of jet substructure [48], the development of observables which can be computed to high precision in QCD and which display reduced sensitivity to non-perturbative effects [48, 50], giving rise to phenomenological studies with LHC data [23, 24]. The analytical calculations give powerful insight into the physics of jet substructure and into the factors influencing and driving tagger performance in a way that is virtually impossible to extract from limited shower studies unguided by any analytics. There are several spin-offs arising from this insight but most crucially it opens the way to creating optimal tools which are not just performant but also reliable and robust. It is, of course, a relatively simple exercise to use Monte Carlo tools to get an estimate of tagger performance, which

JHEP09(2018)170

is typically done via generating the so-called ROC (Receiver Operating Characteristic) curves which plot the background mistag rate against the signal efficiency achieved with different taggers. However any result that derives from QCD theory should also come with an uncertainty estimate which reflects the theoretical approximations made and ROC curves are no exception to this. However theoretical uncertainties on results produced purely from Monte Carlo methods are not simple to estimate and given the quite basic leading-logarithmic accuracy of parton showers[1] one may worry that such uncertainties, if estimated properly, will be very large. Thus unambiguous statements about comparative tagger performance based solely on Monte Carlo studies are always potentially dangerous and support from analytic calculations gains further importance.

In this article we shall carry out an investigation of aspects of top tagging using analytic resummation as a main tool. While previous analytic studies such as those in refs. [39, 48–54] have focused on the case of W/Z/H tagging, here our aim is to embark on a similar level of understanding for top tagging. We shall mainly explore methods for identifying the top quark based on its three-pronged decays i.e. shall focus on the prong finding aspect of top taggers.

A study that covers all of the existing top-taggers goes beyond the scope of our current article. Instead, to illuminate some of the main features we shall consider a standard method, the CMS top-tagger [29, 30] which is closely related to the Johns Hopkins tagger [27], as well as introduce another method based on the Y_m-splitter tagger [53], itself a variant of the Y-splitter method already used in top-tagging [2, 32], and also investigate the combination of Y_m-splitter with jet grooming. We start in section 2 by defining in detail the CMS tagger and pointing out that it suffers from the issue that it becomes collinear unsafe at high p_t, with potential consequences for precision in QCD calculations.

We therefore propose two new variants of the CMS tagger, $\text{CMS}^{3\text{p,mass}}$ and `TopSplitter`, which are both infrared and collinear (IRC) safe and, especially in the case of `TopSplitter`, more suited to analytical calculations based on resummation. Next, in the same section, we also define the Y_m-splitter method and extend it for the purpose of identifying three-pronged jet substructure in the context of top tagging. We further discuss the combination of Y_m-splitter with grooming which is needed in order to achieve a good performance with Y_m-splitter.

In section 3 we carry out an $\mathcal{O}\left(\alpha_s^2\right)$ leading-order calculation for the CMS tagger and for Y_m-splitter. At this order the CMS tagger is IRC safe and the collinear unsafety arises at next-to-leading order level and beyond. We first carry out a calculation using a simplified picture based on soft emissions which are strongly ordered in emission angles. Next we discuss reasons for why such a picture, which we might expect to be correctly described by most parton shower methods, may be insufficient for the case of top tagging. We explain that a more natural picture to describe top-taggers is instead based on the use of triple-collinear splitting functions which describe the collinear $1 \to 3$ splitting of an energetic parton, with no strong ordering between the final emissions and no

[1] Recent work has shown that some widely used parton showers often fail to achieve even full leading-logarithmic accuracy for well-known simple observables like the thrust distribution [55].

JHEP09(2018)170

soft approximation [56–58]. We then carry out calculations for the various taggers using the triple-collinear splitting functions and phase-space. We note that fixed-order studies for three-pronged jet substructure using triple-collinear splitting functions have previously been carried out in ref. [59].

Section 4 contains a description of the resummation we perform for the different taggers starting with Y_m-splitter and its combination with grooming and then moving on to `TopSplitter`. Here we present the arguments leading to the resummed results in each case as well as leading-logarithmic results for the Sudakov form factors using a fixed-coupling approximation, although our final results also include both the effect of hard-collinear next-to-leading logarithmic corrections as well as running coupling effects. We discuss how to match the Sudakov form factors computed in the soft and strongly ordered approximation, with the leading-order pre-factor computed in the triple-collinear limit.

In section 5 we first discuss the numerical impact of including the triple-collinear splitting function and of various resummation effects, then compare the results of our analytical calculations for QCD background jets with parton level results from using the Pythia shower [60]. We study different analytical approximations to the Sudakov exponent for each tagger compared to the Pythia result and also directly compare the taggers to one another both using our analytical results and using Pythia.

Section 6 contains our studies for signal jets as well as studies of tagger performances with ROC curves generated both analytically and with Monte Carlo. We also investigate in this section the role of non-perturbative effects including both hadronisation and the underlying event. Our conclusions are presented in section 7. An explicit demonstration of the collinear unsafety of the CMS tagger using fixed-order perturbative QCD, a discussion of further tagger variants, and analytical results including running-coupling effects can be found in the appendices.

2 Tagger definitions

In this section we shall describe the default version of the CMS tagger and discuss its potential collinear unsafety issue. We shall define a variation of the CMS method, $\text{CMS}^{\text{3p,mass}}$, that is IRC safe and we shall introduce a new method we call `TopSplitter` that apart from being IRC safe is more amenable to a detailed analytical understanding. We also discuss our implementation of the Y_m-splitter method for top tagging and discuss the combination of Y_m-splitter with grooming, extending the ideas we first introduced in ref. [53].

2.1 The CMS top tagger and new methods

The steps involved in the CMS top tagger are detailed below. The first version of the CMS tagger reported in [29, 30], proceeds as follows:[2]

1. The initial anti-k_t jet [61] is re-clustered using the Cambridge-Aachen (C-A) algorithm [62, 63].

[2]The explicit code can be found as part of CMSSW, see [31] which is what we have used in this paper.

JHEP09(2018)170

2. *Primary decomposition*: the last step of the clustering is undone (starting from the initial jet for the first iteration, or from the result of the last iteration when recursing), giving 2 prongs. These two prongs are examined for the condition

$$p_t^{\text{prong}} > \zeta_{\text{cut}}\, p_t^{\text{jet}}, \tag{2.1}$$

where p_t^{jet} refers to the hard jet transverse momentum. ζ_{cut}, referred to as δ_P in the CMS papers, is a parameter which is usually taken as 0.05. If both prongs pass the cut then the "primary" decomposition succeeds. If both prongs fail the cut then the jet is rejected i.e. is not tagged as a top jet.

If a single prong passes the cut the primary decomposition recurses into the passed prong, until the decomposition succeeds or the whole jet is rejected. Note that during the recurrence, p_t^{jet} (used in (2.1)) is kept as the transverse momentum of the original jet.

3. *Secondary decomposition*: with the two prongs found by the primary decomposition, repeat the declustering procedure as for the primary decomposition, still defining the ζ_{cut} condition (2.1) w.r.t. the original jet p_t. This can result in either both prongs from the primary decomposition being declustered into two sub-prongs, only one prong being declustered, or none. When no further substructure is found in a primary prong, the primary prong is kept intact in the final list of prongs. When two sub-prongs are found both are kept in the final list of prongs. Ultimately, this leads to two, three or four prongs emerging from the original jet. Only jets with three or four sub-prongs are then considered as top candidates.

4. Taking the three highest p_t subjets (i.e. prongs) obtained by the declustering, the algorithm finds the minimum pairwise mass and requires this to be related to the W mass, m_W, by imposing the condition $\min(m_{12}, m_{13}, m_{23}) > m_{\text{min}}$ with $m_{\text{min}} \lesssim m_W$. For practical applications, m_{min} is usually taken as 50 GeV.

5. Note that in the second version of the tagger [30], the decomposition procedure also imposes an angular cut: when examining the decomposition of a subjet S into two prongs i and j, the CMS tagger also requires $\Delta R_{ij} > 0.4 - A p_t^S$ where $\Delta R_{ij} = \sqrt{\Delta y_{ij}^2 + \Delta\phi_{ij}^2}$ and p_t^S refers to the transverse momentum of the subjet.[3] The default value for A is $0.0004\,\text{GeV}^{-1}$.

We also note here that the first version of the tagger [29] does not make a reference to the ΔR condition in the decomposition of a cluster. In fact without a ΔR cut the tagger is *collinear unsafe*. This in turn implies that fixed-order perturbative QCD results for observables can produce divergent results, thereby compromising the reliability of the tagger.

The collinear unsafety arises due to the process of selecting the three hardest prongs out of four prongs (to define the m_{min} cut) which is sensitive to arbitrarily collinear hard

[3] For the p_t scale entering the ΔR condition, ref. [30] mentions using the original jet (resp. the primary prongs) during the primary (resp. secondary) decomposition. However, the code in CMSSW is explicitly using the "local" subjet p_t.

JHEP09(2018)170

radiation (see appendix A for an explicit demonstration of the collinear unsafety aspect, using fixed-order perturbative QCD). With a ΔR cut formal collinear safety is restored but for small values $\Delta R \ll 1$, one will encounter large logarithms in ΔR making a perturbative description of the tagger potentially complicated. Also, given the recommended optimal value for the parameter A, as one progresses towards high p_t values the ΔR cut becomes smaller and eventually vanishes which means that the default CMS tagger will again be collinear unsafe at asymptotically large p_t.

To evade the issue of collinear unsafety one could argue that precision perturbative calculations are not the main aim of jet substructure studies, at least in the context of LHC searches for new physics. However as we stated in the introduction, assessing the uncertainty on results for tagger signal and background efficiencies is far from simple, and with an IRC unsafe tool this becomes even less straightforward. Hence any statements about tagger performance based on ROC curves cannot be formally taken at face value. Moreover not all jet substructure studies are aimed at direct searches for new physics, and substructure tools are widely used in an increasing variety of contexts including for precision studies and comparison between perturbative QCD calculations and experimental data [20–22], possible extractions of the strong coupling [64], and in the case of top quark physics, determinations of the top mass [19]. For such studies, where high precision and small uncertainties are essential, any IRC unsafety issues can severely compromise the validity of the results obtained and conclusions reached. It is therefore desirable to ensure a set of substructure tools that are free from IRC unsafety issues while still yielding the required performance.

Ultimately, this collinear unsafety issue motivated us to investigate alternatives to the ΔR cut imposed by the CMS top tagger and to introduce the following new methods:[4]

- CMS$^{\text{3p,mass}}$: say that the primary decomposition led to the two prongs A and B and that prong A has a secondary decomposition into subprongs A$'$ and A$''$ while B is decomposed into B$'$ and B$''$. Rather than selecting the hardest 3 objects from the set A$'$, A$''$, B$'$, B$''$ as in the standard CMS tagger, one instead examines the invariant masses $m^2_{\mathrm{A}'\mathrm{A}''} = (p_{\mathrm{A}'} + p_{\mathrm{A}''})^2$ and $m^2_{\mathrm{B}'\mathrm{B}''} = (p_{\mathrm{B}'} + p_{\mathrm{B}''})^2$. If $m^2_{\mathrm{A}'\mathrm{A}''} > m^2_{\mathrm{B}'\mathrm{B}''}$ then one simply considers the 3 prongs to be A$'$, A$''$ and B, and vice-versa. In this variant of the CMS method we obtain 3 prongs which can be used in the m_{min} condition without any collinear unsafety issues and without a ΔR cut. We shall refer to this variant as CMS$^{\text{3p,mass}}$ since it produces three prongs based on a selection using invariant masses.

- `TopSplitter`: as we shall clarify in more detail in subsequent sections, it proves to be advantageous in some respects to nominate the emission that would dominate the mass of a prong in the limit where all emissions are soft and strongly ordered in mass, as a product of the declustering, instead of the largest-angle emission passing

[4] We are aware that in the meantime the CMS collaboration, for reasons unrelated to collinear unsafety pointed out for the first time in our current article, have moved away from using the CMS tagger in experimental studies of top-tagging at the LHC. Nevertheless, as we shall demonstrate, the IRC safe variants we propose in this article are effective and performant methods to identify three-pronged substructure and demonstrate our analytical control over top tagging.

JHEP09(2018)170

the $\zeta_{\rm cut}$ as given by the C-A declustering. In order to do so we first keep the same procedure as above for identifying the two prongs A and B that emerge from the primary decomposition. Now consider the decomposition of each of these prongs starting say with prong A. We decluster this precisely as before until we find an emission i that passes the $\zeta_{\rm cut}$ condition. At this stage however we also consider *all* subsequent emissions further down the C-A tree following the hardest branch, together with emission i, and identify the emission j in this set that has the largest value of $p_{tj}\theta_j^2$, i.e contributes the most to the prong mass in the limit that all emissions are soft.[5] We take this emission to be A$''$ i.e. one of the products of the declustering of A. The other product of the declustering is labelled A$'$ as before. It consists of the remaining object to which A$''$ is clustered in the C-A clustering sequence, along with all emissions preceding A$''$ in the C-A tree which passed the $\zeta_{\rm cut}$ condition, such as emission i. We call this new method `TopSplitter`.

Other variants are possible and they will be discussed in appendix B.

2.2 The $\mathrm{Y_m}$-splitter method for top tagging

The use of the Y-splitter method for top tagging was already considered by Brooijmans and made use of in ATLAS studies of top tagging [32, 65].

In refs. [53, 66] it was found that the Y-splitter technique, when supplemented with grooming, was a high-performance method for the tagging of electroweak scale particles that exhibit two pronged decays, especially for p_t values in the TeV range. To be more precise, it was observed in refs. [53, 66] that Y-splitter gives an excellent suppression of QCD background jets due to a large Sudakov suppression factor. However the performance of Y-splitter on signal jets was poor as the lack of an explicit grooming step resulted in loss of signal. Once grooming is performed after Y-splitter (either via mMDT [48] or trimming [6]), while the feature of the background suppression stays largely intact, there is considerable improvement in the signal efficiency. This results in striking gains for the signal significance.

Therefore it also becomes of interest to adapt Y-splitter with grooming to the case of top decays. In ref. [53] we introduced and discussed several variants of the Y-splitter technique for the case of two pronged decays. The variant that emerged as both most robust and performant was a variant we called $\mathrm{Y_m}$-splitter, which makes use of the gen-k_t ($p = 1/2$) algorithm [67] to define distances between objects, in place of the k_t distance [68–70] used in the standard Y-splitter. The use of the gen-k_t ($p = 1/2$) distance (hereafter referred to just as the gen-k_t distance for brevity) guarantees an ordering equivalent to an ordering in mass in the soft limit which facilitates the direct analytical understanding of the tagger behaviour, with the fringe benefit of giving a slightly better performance compared to the standard Y-splitter.

We will also consider pre-grooming with SoftDrop ($\beta = 2$ and $\beta = 0$ (i.e. mMDT)) prior to the application of $\mathrm{Y_m}$-splitter. The $\beta = 2$ pre-grooming option was already explored

[5] Given that this emission is either emission i itself or a smaller angle emission, it is clear that it must also pass the $\zeta_{\rm cut}$ condition.

JHEP09(2018)170

for the tagging of W/Z/H and found to give good performance while highly reducing the sensitivity to non-perturbative effects [53]. The $\beta = 0$ pre-grooming option was not considered in ref. [53] since for the case of W/Z/H tagging studied there, this option was found to reduce the important Sudakov suppression of the background. In the present case however, where we have a coloured object being tagged, the situation will be different as we shall explain in more detail in section 6, and pre-grooming with mMDT becomes a useful option to consider.

To adapt Y_m-splitter for use in top tagging one considers applying it twice in succession, as follows:

1. Perform a primary decomposition of the initial fat jet by doing a first declustering but here based on the gen-k_t ($p = 1/2$) distance measure. On each of the two prongs obtained by undoing the clustering apply the ζ_{cut} condition, eq. (2.1). If the ζ_{cut} condition fails for either of the two prongs, discard the jet as a top candidate, otherwise move to the next step.

2. Decluster both prongs obtained from the primary decomposition (still using the gen-k_t algorithm). The prong that produces the smaller gen-k_t distance in the declustering is kept unaltered. The prong that yielded the larger gen-k_t distance is tested for the ζ_{cut} condition as for the primary decomposition. If the ζ_{cut} condition passes proceed to the next step otherwise the jet is rejected.

3. Take the three prongs that emerge after the secondary decomposition (i.e. the unaltered primary prong and the two secondary prongs which passed the ζ_{cut} condition) and impose the m_{min} condition on the minimum pairwise mass.

Additionally, as mentioned above, we shall consider pre-grooming with mMDT and SoftDrop on the full jet, prior to the application of the above steps. Lastly, we also introduced additional variants for Y_m-splitter similar to the case of the CMS tagger and these are also discussed in appendix B.

3 Analytical calculations at fixed-order

In this section we shall carry out some basic leading-order analytic calculations to help us better understand the action of top taggers on QCD jets. We shall start by using a soft and collinear approximation for emissions within the jet and then discuss improving this approximation in light of the specific requirements for three-pronged jet substructure and top taggers.

3.1 Leading-order calculations in the soft-collinear limit

The standard idea that is exploited in two-body tagging to distinguish signal from background is to exploit the differences in splitting functions between QCD decays and those involving W/Z/H. While the former contain soft enhancements, the latter are regular in

JHEP09(2018)170

the soft limit and hence cutting the soft region via a δ_P or $z_{\rm cut}$ type of condition[6] reduces the background significantly compared to the modest impact on the signal (see e.g. ref. [48] for explicit examples and more details). For the case of three-body hadronic top decays we have instead *two* branchings that are not soft-enhanced namely the branching $t \to b\,W$ and then the two-body W decay to quarks. We should therefore expect that the double application of the $\zeta_{\rm cut}$ condition exploits this feature.

In order to see this most clearly, in this sub-section we perform a leading-order QCD calculation for the jet-mass distribution for QCD jets after the application of top-tagging methods. In the boosted limit the jet mass m is small compared to the jet p_t and we shall work in terms of the standard variable, $\rho = \frac{m^2}{R^2 p_t^2}$, with m the jet mass and where the jet radius R reflects the jet opening angle. In the small-angle limit ρ is invariant under boosts along the jet direction since they scale the jet p_t up by a factor γ and its opening angle by a factor $1/\gamma$, such that the jet mass m is unchanged.

For the application of top taggers, aside from the jet mass we also have the $m_{\rm min}$ condition and hence also define $\rho_{\rm min} = \frac{m_{\rm min}^2}{R^2 p_t^2} \ll 1$. The other parameter which enters our calculations is $\zeta_{\rm cut}$. This is chosen not too small in order to reduce the QCD background i.e. $\zeta_{\rm cut} \gg \rho, \rho_{\rm min}$ but nevertheless $\zeta_{\rm cut} \ll 1$, with the value $\zeta_{\rm cut} = 0.05$ generally favoured in practical applications. We therefore expect that in a perturbative calculation we will encounter large logarithms in the jet masses $\rho, \rho_{\rm min}$ as well as large logarithms in $\zeta_{\rm cut}$ but with the former being numerically dominant over the latter.

A further issue that arises is the potential presence of logarithms of $\rho/\rho_{\rm min}$ at each order in perturbation theory. In practice however, given that we are interested in top tagging, the jet mass $m \sim m_t$ and $m_{\rm min} \sim m_W$ are not strongly ordered, hence logarithms of $\rho/\rho_{\rm min}$ are not necessarily large. Furthermore, $\frac{\rho_{\rm min}}{\rho} \gtrsim \zeta_{\rm cut}$ for $\zeta_{\rm cut} = 0.05$ and $m_{\rm min} = 50\,\text{GeV}$. We will therefore consider that $\rho_{\rm min}/\rho$ is small enough to retain only logarithms of $\rho_{\rm min}/\rho$, but not too small so logarithms of the jet mass dominate over logarithms of $\rho_{\rm min}/\rho$. We shall return to address these assumptions in the next subsection and subsequent sections.

With the above mentioned large logarithms in mind we shall initially specialise to the soft and collinear limit for all emissions i.e. $z_i, \theta_i \ll 1$, where z_i is the fraction of the jet's p_t carried by emission i and θ_i the angle of emission i w.r.t. the jet axis. Moreover to calculate the leading logarithms in jet mass we can further assume that successive emissions are strongly ordered in angles. In order to pass the top-tagger conditions one requires at least two emissions in addition to the hard parton that initiates the jet. Thus the leading order in perturbative QCD for the jet mass distribution, with application of top tagging, is order α_s^2. Assuming that the jet is initiated by a hard quark we start by considering two soft and collinear gluon emissions strongly ordered in emission angles and emitted independently by the hard quark, corresponding to a C_F^2 colour factor.

We start by applying the CMS top tagger and variants thereof. At the leading order, i.e. order α_s^2, the CMS tagger, CMS$^{\rm 3p,mass}$ and `TopSplitter` are all equivalent. The IRC

[6] This would involve a cut of the form $\frac{\min(p_{T,i}, p_{T,j})}{p_{T,i}+p_{T,j}} > z_{\rm cut}$ which uses the local p_T of the cluster being decomposed, i.e. $p_{T,i} + p_{T,j}$ instead of the global p_T of the hard jet in the denominator as is the case for the original CMS δ_P condition.

JHEP09(2018)170

unsafety issue of the CMS tagger occurs at order α_s^3 i.e. at the NLO level in the context of the present calculations. Hence for the purpose of this section we shall refer explicitly to the CMS tagger with the understanding that the results apply equally for our new methods. After the primary C-A declustering of the jet, the larger angle gluon k_1 emerges first and is subjected to the ζ_{cut} condition which leads to the constraint $z_1 > \zeta_{\text{cut}}$. We obtain two subjets: a massless subjet j_1 consisting of parton k_1 and a massive subjet j_2 composed of a hard quark with four-momentum p and the emission k_2. One then declusters j_2 into its massless partonic constituents and retains the jet only if $z_2 > \zeta_{\text{cut}}$.

The tagger places a constraint on the minimum pairwise mass of the three partons p, k_1 and k_2 which can be written as $\min\left(z_1\theta_1^2, z_2\theta_2^2, z_1 z_2 \theta_{12}^2\right) > \rho_{\text{min}}$, where all angles are taken to be measured in units of the jet radius R, meaning in particular that they should be less than 1. In the strongly ordered limit we have that $\theta_1 \gg \theta_2$ and $\theta_{12} \approx \theta_1$. Therefore, the minimum pairwise mass is the minimum of $z_2\theta_2^2$ and $z_1 z_2 \theta_1^2$. At leading order accuracy, we then have to consider two cases: either the first emission dominates the mass, meaning we have $\rho \sim z_1\theta_1^2$, or the second emission dominates the mass, i.e. $\rho \sim z_2\theta_2^2$. In other words, we can take $\rho \sim \max(z_1\theta_1^2, z_2\theta_2^2)$ and approximate the jet mass distribution as follows

$$\begin{aligned}\frac{1}{\sigma}\left(\frac{d\sigma}{d\rho}\right)^{\text{LO,soft}-\text{collinear}} &= \bar{\alpha}^2 \int \frac{dz_1}{z_1}\frac{dz_2}{z_2}\frac{d\theta_1^2}{\theta_1^2}\frac{d\theta_2^2}{\theta_2^2} \\ &\quad \times \Theta(\theta_2^2 < \theta_1^2 < 1)\,\delta(\rho - \max(z_1\theta_1^2, z_2\theta_2^2)) \\ &\quad \times \Theta(z_1 > \zeta_{\text{cut}})\,\Theta(z_2 > \zeta_{\text{cut}})\,\Theta(\min(z_2\theta_2^2, z_1 z_2\theta_1^2) > \rho_{\text{min}}), \end{aligned} \tag{3.1}$$

where we defined $\bar{\alpha} = \frac{C_F \alpha_s}{\pi}$, taking for definiteness the case of a quark initiated jet.

The above integrations can be straightforwardly done to obtain the following simple result:

$$\begin{aligned}\frac{\rho}{\sigma}\left(\frac{d\sigma}{d\rho}\right)^{\text{LO,soft}-\text{collinear}} &\overset{\frac{\rho_{\text{min}}}{\rho} < \zeta_{\text{cut}}}{=} \bar{\alpha}^2 \ln^2 \frac{1}{\zeta_{\text{cut}}} \ln \frac{\rho}{\rho_{\text{min}}}, \\ &\overset{\frac{\rho_{\text{min}}}{\rho} > \zeta_{\text{cut}}}{=} \bar{\alpha}^2 \ln^2 \frac{\rho}{\rho_{\text{min}}} \left(\frac{3}{2}\ln\frac{1}{\zeta_{\text{cut}}} - \frac{1}{2}\ln\frac{\rho}{\rho_{\text{min}}}\right). \end{aligned} \tag{3.2}$$

The essential functioning of the tagger at leading-order is encoded in the above equation. For comparison, remember that the leading logarithmic behaviour for the QCD background jet mass distribution is double logarithmic i.e. $\frac{\rho}{\sigma}\frac{d\sigma}{d\rho} \sim \bar{\alpha}^2 \ln^3 \frac{1}{\rho}$. After applying the CMS tagger these large logarithms in jet mass have been replaced by logarithms of ζ_{cut} or ρ_{min}/ρ which are not essentially large. This is similar to the action of taggers in the two-body case.

We can also perform a similar calculation for the Y_m-splitter technique defined in section 2.2. The essential difference with the CMS tagger is the use of the gen-k_t distance with its parameter p taken to be 1/2 (instead of the C-A declustering used for the CMS tagger). With only two emissions in the jet, both emissions need to satisfy $z_i > \zeta_{\text{cut}}$. At leading logarithmic accuracy, one can still assume that $\theta_1 \gg \theta_2$, hence $\theta_{12} \approx \theta_1$, and the expression for the Y_m-splitter cross-section also takes the form of (3.1) (see as well the discussion about Θ^{tagger} in the next section). This means that the result (3.2) is also valid for Y_m-splitter.

JHEP09(2018)170

Similar calculations can be carried out for the terms involving secondary emissions i.e those involving say an initial quark emitting a gluon which splits to a gg or $q\bar{q}$ pair i.e. the $C_F C_A$ and $C_F T_R n_f$ channels. In the former case one obtains again a three powers of a logarithm in either ρ/ρ_{min} or ζ_{cut} as in (3.2). The $C_F n_f$ term has at most two logarithms in ρ/ρ_{min} or ζ_{cut} due to the absence of a soft singularity in the $g \to q\bar{q}$ splitting.

While eq. (3.2) captures the basic physics of the tagger in the limit $1 \gg (\rho_{\text{min}}/\rho, \zeta_{\text{cut}}) \gg \rho$, a number of comments are in order. First of all we have used the approximation of strong angular-ordering which is intended to capture logarithms in ρ/ρ_{min} and ζ_{cut}. Additionally we also used the soft approximation in performing the calculation which does not reproduce the constant terms stemming from hard collinear emissions, or power corrections in ζ_{cut} or ρ_{min}/ρ. The former constant contributions, in particular, are known to be numerically significant in practice [48]. The standard method to include hard collinear splitting is to correct the soft approximation, used above, with the full splitting function i.e. make the replacement $\frac{dz}{z} \to \frac{1+(1-z)^2}{2z} dz$ for the integral over energy fractions in eq. (3.1). We should then expect a product of leading-order splitting functions to appear, which account for both hard branchings i.e. the region where z_1, z_2 are both finite. Moreover, beyond the soft limit, the gen-k_t distances, involved in the Y_{m}-splitter calculation, would no longer be identical to the mass. All these changes are straightforward to implement and do not require a fundamental change of the basic angular-ordered picture above.

More crucially perhaps, as we already observed, the approximation $\rho_{\text{min}} \ll \rho$ and $\zeta_{\text{cut}} \ll 1$, while convenient analytically, is in practice not a good approximation for the case of top tagging. Without these strong ordering assumptions, we are led to a situation where the only genuinely large logarithms in the boosted limit are those in ρ or equivalently ρ_{min} but not those of ρ/ρ_{min}. In other words we should regard eq. (3.2) as an approximation to a result of the form

$$\frac{1}{\sigma}\left(\frac{d\sigma}{d\rho}\right)^{\text{LO,triple-collinear}} = \frac{\alpha_s^2}{\rho} f_q\left(\rho, \rho_{\text{min}}, \zeta_{\text{cut}}\right). \tag{3.3}$$

In the above equation f_q is a function that needs to be computed in full i.e. without any soft or collinear approximation and where the suffix q indicates a quark initiated jet. It contains the contributions from $C_F^2, C_F C_A$ and $C_F T_R n_f$ colour factors on an equal footing. The only approximation inherent in writing eq. (3.3) is the approximation of small $\rho \ll 1$, corresponding to appearance of the $1/\rho$ factor, which is justified by working in the boosted limit $m^2 \ll p_t^2$. Thus we need to examine the collinear decay of an initial parton to three partons, producing a small jet mass ρ, but with *no ordering between the three partons themselves, in either energy or angle*. The appropriate extension of eq. (3.2) requires the use of *triple-collinear* $(1 \to 3)$ *splitting functions*. Calculations based on these shall be the subject of the next section.

3.2 The triple-collinear limit of a QCD jet

Here we shall use the $1 \to 3$ splitting functions [56–58] to compute the differential distribution in the jet mass ρ, for the CMS and Y_{m}-splitter methods.

JHEP09(2018)170

Consider for example the collinear decay of an initial quark to a quark and two gluons, taking the Abelian C_F^2 term of the triple-collinear splitting functions as an example. The explicit functional form for the spin-averaged splitting function is

$$\langle \hat{P}^{(\mathrm{ab})}_{g_1 g_2 q_3} \rangle = C_F^2 \left[\frac{s_{123}^2}{2 s_{13} s_{23}} z_3 \left(\frac{1+z_3^2}{z_1 z_2} \right) + \frac{s_{123}}{s_{13}} \left(\frac{z_3(1-z_1)+(1-z_2)^3}{z_1 z_2} \right) - \frac{s_{23}}{s_{13}} \right] + (1 \leftrightarrow 2)\,. \tag{3.4}$$

For the other colour configurations (involving C_A and n_f), we refer the reader to the original references [56–58]. Here s_{ij} and s_{ijk} are the usual kinematic invariants $(p_i+p_j)^2$ and $(p_i+p_j+p_k)^2$ respectively. The z_i are energy fractions defined w.r.t. the original parton's energy so that we have $\sum_i z_i = 1$. Also, in what follows below we shall need only the splitting functions in four space-time dimensions and hence have set the dimensional regularisation parameter ϵ to zero above and in all subsequent applications.

The phase-space in the triple-collinear limit can be written as

$$d\Phi_3 = \frac{(p_t R)^4}{\pi} (z_1 z_2 z_3)\, dz_2\, dz_3\, d\theta_{12}^2\, d\theta_{23}^2\, d\theta_{13}^2 \Delta^{-1/2} \Theta(\Delta)\,, \tag{3.5}$$

with the Gram determinant Δ given by [71, 72]

$$\Delta = 4\theta_{13}^2\theta_{23}^2 - (\theta_{12}^2 - \theta_{13}^2 - \theta_{23}^2)^2. \tag{3.6}$$

We then carry out an integral over the triple-collinear phase-space which includes the action of the taggers encoded as a sequence of kinematical cuts. We compute the jet mass distribution as an integral of the schematic form

$$\left(\frac{\rho}{\sigma}\frac{d\sigma}{d\rho}\right)^{\mathrm{LO,triple-collinear}} = \left(\frac{\alpha_s}{2\pi}\right)^2 \int d\Phi_3 \frac{\langle \hat{P} \rangle}{s_{123}^2}\, \Theta^{\mathrm{jet}}\, \Theta^{\mathrm{tagger}}(\zeta_{\mathrm{cut}}, \rho_{\mathrm{min}})\, \rho\, \delta\left(\rho - \frac{s_{123}}{R^2 p_t^2}\right), \tag{3.7}$$

where $\langle \hat{P} \rangle$ denotes the spin-averaged triple-collinear splitting function, including the proper symmetry factor for identical particles, for the splitting of an initial quark (or gluon if considering a gluon-initiated jet), the Θ^{jet} condition denotes the constraint for all three partons to be in the same anti-k_t jet of a given radius R

$$\Theta^{\mathrm{jet}} = \sum_{i>j\neq k} \Theta\Big(d_{ij}^{(\text{anti-}k_t)} < \min\big(d_{ik}^{(\text{anti-}k_t)}, d_{jk}^{(\text{anti-}k_t)}\big)\Big)\, \Theta(\theta_{ij} < R) \Theta(\theta_{(i+j)k} < R), \tag{3.8}$$

and the condition Θ^{tagger} represents the action of the substructure taggers. In particular, Θ^{tagger} contains constraints from the ζ_{cut} and ρ_{min} conditions which will regulate the soft and collinear divergences of the $1 \to 3$ splitting functions. Accordingly we can carry out the computation of the jet mass distribution entirely in 4 dimensions and only real-emission terms contribute at the leading order α_s^2.

For any of the taggers we have introduced, we have, at order α_s^2,

$$\begin{aligned}\Theta^{\mathrm{tagger}}(\zeta_{\mathrm{cut}}, \rho_{\mathrm{min}}) = \sum_{i>j\neq k} \Theta\left(d_{ij}^{(\mathrm{tagger})} < \min(d_{ik}^{(\mathrm{tagger})}, d_{jk}^{(\mathrm{tagger})})\right) \\ \times \Theta\left(\min(z_k, 1-z_k) > \zeta_{\mathrm{cut}}\right) \Theta\left(\min(z_i, z_j) > \zeta_{\mathrm{cut}}\right) \\ \times \Theta\left(\min(\rho_{ij}, \rho_{jk}, \rho_{ki}) > \rho_{\mathrm{min}}\right),\end{aligned} \tag{3.9}$$

JHEP09(2018)170

where the only difference between the CMS (recall that there is no difference at order α_s^2 between the default CMS, CMS$^{\text{3p,mass}}$ and `TopSplitter`) and the Y_m-splitter taggers is in the distance measure they use:

$$d_{ij}^{(\text{CMS})} = \theta_{ij}^2, \tag{3.10}$$

$$d_{ij}^{(\text{Y}_\text{m}\text{-splitter})} = \min(z_i, z_j)\theta_{ij}^2. \tag{3.11}$$

However, at this order of the perturbation theory, eq. (3.9) is equivalent to the simpler form

$$\Theta^{\text{tagger}}\left(\zeta_{\text{cut}}, \rho_{\text{min}}\right) = \Theta\left(\min(z_1, z_2, z_3) > \zeta_{\text{cut}}\right)\Theta\left(\min(\rho_{12}, \rho_{13}, \rho_{23}) > \rho_{\text{min}}\right), \tag{3.12}$$

which is the same for the CMS and Y_m-splitter taggers.

It is worth noting that the triple-collinear splitting functions and phase-space are *not* presently included in current parton shower models implemented in any of the main general purpose Monte Carlo event generator codes.[7] Parton showers instead include the strongly-ordered limit of the triple-collinear functions where the triple-collinear functions factorise into a product of two leading-order splitting kernels. It is simple to make this link explicit by expanding the triple-collinear functions about the strongly ordered limit. For instance for the C_F^2 term reported in eq. (3.4) we can explicitly take the limit $\theta_{23}^2 \ll \theta_{13}^2$ and perform an expansion in the smallest angle θ_{23}^2. Writing $\theta_{12}^2 = \theta_{13}^2 + \theta_{23}^2 - 2\theta_{13}\theta_{23}\cos\phi$ and introducing the splitting variables z and z_p such that $z_1 = 1-z, z_2 = z(1-z_p), z_3 = zz_p$, one obtains upon series expansion in θ_{23}^2:

$$\frac{\langle \hat{P}^{(\text{ab})}_{g_1 g_2 q_3}\rangle}{s_{123}^2} d\Phi_3 = C_F^2 \frac{d\theta_{13}^2}{\theta_{13}^2}\frac{d\theta_{23}^2}{\theta_{23}^2}\Theta(\theta_{23} < \theta_{13})\, dz\, dz_p \frac{d\phi}{2\pi}\left(\frac{1+z^2}{1-z} \times \frac{1+z_p^2}{1-z_p} + \mathcal{O}\left(\theta_{23}\right)\right). \tag{3.13}$$

The above form exhibits the factorisation of the leading order splitting kernels which is expected in the strongly ordered limit.[8] Additionally taking the soft limit i.e. $z, z_p \to 1$ brings us back to the approximations used to derive (3.2).

4 Resummed calculation to all orders

Eq. (3.3), making use of the $1 \to 3$ splitting function to obtain $f_q(\rho, \rho_{\text{min}}, \zeta_{\text{cut}})$, is sufficient to obtain the small ρ (and ρ_{min}) behaviour at order α_s^2. However, the large logarithms of ρ or ρ_{min} need to be resummed to all orders in α_s. For this, in addition to the two hard collinear emissions described by the $1 \to 3$ splitting, we need to add an arbitrary number of real or virtual soft and collinear emissions and consider the constraints on them. As is standard in resummation, this is expected to yield a Sudakov form factor that multiplies the leading-order result. In this section, we derive the explicit form of that Sudakov for the `TopSplitter`, CMS$^{\text{3p,mass}}$ and Y_m-splitter top taggers. The IRC unsafety of the default CMS tagger prevents a similar analysis being directly carried out for that case.

[7] For recent attempts at partially including these effects in parton showers we refer the reader to [73].

[8] In general i.e. beyond the Abelian C_F^2 term a fully factorised structure is obtained *after* an azimuthal integration in the strongly ordered limit.

JHEP09(2018)170

Before digging into the details of each tagger, let us clarify the accuracy of our resummation. First of all, previous work has shown [48, 52, 53, 66] that a leading-logarithmic (LL) calculation is usually sufficient to grasp the main features of substructure tools. In that context, our resummation should definitely include large logarithms of the jet mass (ρ or ρ_{min}) to LL accuracy. These logarithms are the most relevant for describing boosted jets and we shall see that including them holds the key to understanding the basic details of the top-taggers.

However, since both ρ_{min}/ρ and ζ_{cut} are somewhat smaller than unity for practical applications, it might also be of interest to include logarithms in ρ_{min}/ρ and ζ_{cut} in the resummation. Indeed, without including these terms one may worry about their impact on our analytical picture for top tagging. With this in mind our resummation accuracy goal will ideally be double-logarithmic, but where the scale of the logarithms can be either ρ (or ρ_{min}), ρ_{min}/ρ or ζ_{cut}. We will also confirm that in the final result the logarithms of ρ or ρ_{min} dominate over logarithms of ρ_{min}/ρ and ζ_{cut} in the region relevant for phenomenology, as one might have naively expected purely on grounds of their numerical size.

While our final resummation accuracy is double-logarithmic, or more precisely leading-logarithmic after inclusion of running coupling effects, we shall also retain some sources of single-logarithmic corrections, notably via the inclusion of hard-collinear contributions which arise from considering the full splitting functions rather than just their soft-enhanced terms. This is again standard in the existing resummed calculations for jet substructure (see e.g. ref. [48]) and is sometimes referred to as modified leading logarithmic accuracy [50].

We would like to stress that strictly from the point of view of our logarithmic accuracy, we do not need the full structure of the triple-collinear splitting functions that we have used above to compute the leading-order pre-factor, that will eventually multiply the Sudakov exponent. Instead one could just treat the pre-factor in the soft limit with strong angular-ordering, which is sufficient to retain all double-logarithmic terms in the pre-factor. Using the triple-collinear splitting functions means that we have instead chosen to be more careful in our treatment of the pre-factor by retaining terms that are formally subleading from the viewpoint of our logarithmic resummation accuracy. In effect we thus perform a form of matching so that at order α_s^2 our result coincides with using the full triple-collinear splitting function, while beyond order α_s^2 our result should contain all potentially large double-logarithmic terms, counting logs of $\rho, \rho_{\text{min}}, \rho/\rho_{\text{min}}$, and ζ_{cut} on the same footing. In practice we are able to achieve this goal for the `TopSplitter` and Y_{m}-splitter taggers including also Y_{m}-splitter with general SoftDrop gre-grooming. Instead for the case of the $\text{CMS}^{\text{3p,mass}}$ tagger it does not prove to be simple to include logarithms of ζ_{cut} and ρ/ρ_{min} on the same footing as double logarithms in ρ or ρ_{min}. Accordingly for $\text{CMS}^{\text{3p,mass}}$ we do not attempt to retain all possible double logarithms, focusing mainly on the numerically dominant logarithms in ρ or ρ_{min}. This level of accuracy is still sufficient to provide us insight into the behaviour of the tagger in the region relevant for phenomenology.

The structure and details of the resummed result depend on the tagger being considered. We first discuss the Y_{m}-splitter case due to its simpler structure.

JHEP09(2018)170

4.1 $\mathbf{Y_m}$-splitter

At double-logarithmic accuracy, we can consider emissions to be soft and strongly ordered, here in gen-k_t distance, or, equivalently, in mass. If one wishes to simultaneously retain the information that is present in the triple-collinear limit however, we have to lift the requirement of strong ordering and the soft approximation, for the two emissions that are declustered by the taggers, while still retaining these approximations for all remaining emissions. In the first instance however it is instructive to impose the soft and strong-ordered requirement on *all* emissions including the declustered ones, which gives us the leading-logarithmic accuracy we seek. Subsequently we shall match our result to the triple-collinear limit.

4.1.1 Calculation in pure soft and strongly-ordered limit

We first consider two soft emissions k_1 and k_2 both emitted by a hard parton leg with colour factor C_R, where $C_R = C_F$ for a quark initiated jet and $C_R = C_A$ for a gluon initiated jet. We denote by x_i and θ_i the momentum fraction and angle of emission k_i defined w.r.t. the hard emitting parton. At leading logarithmic level we can assume strong ordering in the gen-k_t distance or equivalently in masses. Hence we assume $\rho \approx \rho_1 \equiv x_1\theta_1^2 \gg \rho_2 \equiv x_2\theta_2^2$, which implies that emission k_1 is the first to be declustered and k_2 is the second, while k_1 also sets the jet mass ρ.

Next, we consider multiple soft emissions ordered in gen-k_t distance. Emissions k_1 and k_2 are, by construction, the ones obtained by the declustering procedure and subject to the ζ_{cut} requirement. Due to the gen-k_t ordering they are also the emissions that dominate the pairwise masses entering the ρ_{min} condition.[9] Thus all the tagger constraints are fully determined by the declustered partons k_1 and k_2 which produce the leading-order pre-factor, cf. eq. (3.2).

One then has a veto on any additional emission with gen-k_t distance (or, equivalently, mass) larger than $\rho_2 = x_2\theta_2^2$. For (primary) emissions from the leading parton p, this corresponds to the shaded (red) region in the plot of figure 1. In this region, virtual emissions are still allowed, yielding a Sudakov form factor $\exp(-R^{(\text{primary})}_{\text{Y}_\text{m}\text{-splitter}})$, with $R^{(\text{primary})}_{\text{Y}_\text{m}\text{-splitter}}$ corresponding to the shaded area:

$$R^{(\text{primary})}_{\text{Y}_\text{m}\text{-splitter}}(\rho_2) = \int \frac{d\theta^2}{\theta^2}\frac{dz}{z}\,\frac{\alpha_s(z\theta p_t R)C_R}{\pi}\,\Theta(z\theta^2 > \rho_2), \tag{4.1}$$

where we also took into account running coupling effects.

Furthermore, since $\rho_2 \ll \rho_1$, we should also veto (secondary) emissions from k_1 in between these two scales. Consider such an emission from k_1 to be soft and to carry a momentum fraction $z \ll 1$ of the momentum of its parent k_1 implying it has momentum fraction $x_1 z$ w.r.t. the jet p_t. It is emitted at angles smaller than θ_1 due to angular ordering. Such emissions should also be vetoed if they give a gen-k_t distance larger than ρ_2.

[9] Technically, the gen-k_t distance between prong i and prong j differs from the pairwise mass in two ways: first by an overall factor proportional to $z_i + z_j$, and second by factors of the form $1 - z$ which are irrelevant in the resummation limit $z \ll 1$. Overall, this means that the relative ordering of the emissions in mass is the same as their relative gen-k_t ordering.

JHEP09(2018)170

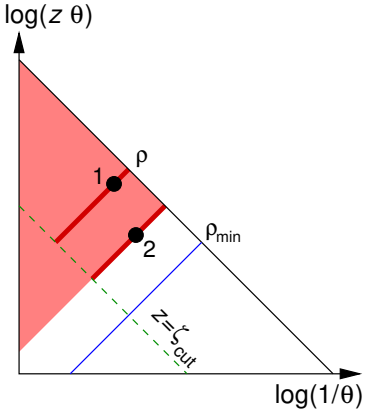


Figure 1. Lund plane corresponding to the Y_m-splitter tagger for top tagging. The emission that has a larger gen-k_t distance or mass is labelled 1 and that with the next largest mass is labelled 2. Emissions 1 and 2 both pass the ζ_{cut} condition shown using the dashed line. The mass ρ_2 lies between ρ and ρ_{min} as shown. The red shaded region represents the region over which emissions are vetoed and leads to the appearance of the primary emission contribution to the Sudakov form factor. A similar configuration where the second-largest mass emission, 2, is emitted as a secondary emission from emission 1, is not shown in this plot.

This is not shown in figure 1 and gives an additional contribution (with $x_1 = \rho_1/\theta_1^2$)

$$R^{(\text{secondary})}_{\text{Y}_\text{m}\text{-splitter}}(\rho_2;\rho_1,\theta_1) = \int \frac{d\theta^2}{\theta^2}\frac{dz}{z}\,\frac{\alpha_s(zx_1\theta p_t R)C_A}{\pi}\,\Theta(zx_1\theta^2>\rho_2)\,\Theta(\theta<\theta_1). \tag{4.2}$$

Note that in the expression for secondary emissions, the ordering in gen-k_t distance (imposed by the declustering procedure of Y_m-splitter), differs from the mass by an x_1 factor.[10]

Ultimately, the Sudakov form factor is

$$\mathcal{S}_{\text{Y}_\text{m}\text{-splitter}}(\rho_2;\rho_1,\theta_1) = \exp\Big[-R_{\text{Y}_\text{m}\text{-splitter}}(\rho_2;\rho_1,\theta_1)\Big], \tag{4.3}$$

$$R_{\text{Y}_\text{m}\text{-splitter}}(\rho_2;\rho_1,\theta_1) = R^{(\text{primary})}_{\text{Y}_\text{m}\text{-splitter}}(\rho_2) + R^{(\text{secondary})}_{\text{Y}_\text{m}\text{-splitter}}(\rho_2;\rho_1,\theta_1). \tag{4.4}$$

In the strongly-ordered (and soft) limit, the resummed result, including the pre-factor is therefore

$$\left(\frac{\rho}{\sigma}\frac{d\sigma}{d\rho}\right)^{\text{resum}} = \int_0^1 \frac{d\theta_1^2}{\theta_1^2}\frac{d\theta_2^2}{\theta_2^2}\frac{dx_1}{x_1}\frac{dx_2}{x_2}\,\frac{\alpha_s(x_1\theta_1 p_t R)C_R}{\pi}\,\frac{\alpha_s(x_2\theta_2 p_t R)C_R}{\pi}\Theta(x_1>\zeta_{\text{cut}})$$
$$\times\,\Theta(x_2>\zeta_{\text{cut}})\,\Theta(\rho_{\text{min}}<x_2\theta_2^2<x_1\theta_1^2)\,\rho\,\delta(\rho-x_1\theta_1^2)\,\mathcal{S}_{\text{Y}_\text{m}\text{-splitter}}. \tag{4.5}$$

The above result coincides with the LO result in eq. (3.2) at order α_s^2, i.e. when switching off the running of the strong coupling and the Sudakov form factor and replacing C_R by C_F. As part of our accuracy goal which aims at correctly retaining *all* double logarithms, our pre-factor should also contain the $\mathcal{O}\left(\ln^3\zeta_{\text{cut}}\right)$ terms neglected in eq. (3.2). Moreover

[10] Note that it is still true that inside the k_1 prong, i.e. amongst all the secondary emissions off k_1, the emission that has the largest gen-k_t distance also dominates the mass of the prong.

JHEP09(2018)170

we should account for all possible branchings that contribute to the pre-factor including the case where emission k_2 is emitted as a secondary emission off k_1, which for a quark initiated jet yields a $C_F C_A$ contribution to the pre-factor. However since, in the next subsection, we eventually use the triple-collinear splitting functions and kinematics to compute our pre-factor, it is guaranteed that all such terms (along with subleading terms relevant beyond the soft and strongly ordered approximation) are correctly retained.

Below we give results in the fixed-coupling case to highlight the different logarithms that are present in the above expressions:

$$R^{(\text{primary})}_{\text{Y}_\text{m}\text{-splitter}}(\rho_2) \overset{\text{f.c.}}{=} \frac{\alpha_s C_R}{2\pi} \ln^2 \rho_2, \tag{4.6}$$

$$R^{(\text{secondary})}_{\text{Y}_\text{m}\text{-splitter}}(\rho_2; \rho_1, \theta_1) \overset{\text{f.c.}}{=} \frac{\alpha_s C_A}{2\pi} \ln^2(\rho_2/\rho). \tag{4.7}$$

After integration over ρ_2, the numerically dominant logarithms will be of the form of a series in $\alpha_s \ln^2 \rho$ where we treat ρ and ρ_{min} on the same footing, which multiplies the leading order pre-factor and originates purely from the veto on primary emissions. Given the range of the ρ_2 integration between ρ_{min} and ρ, secondary emissions can only contribute terms which are at most as singular as $\alpha_s \ln^2(\rho/\rho_{\text{min}})$. We may therefore anticipate that secondary emissions turn out to be relatively significant only when $\rho_{\text{min}} \ll \rho$, which is an observation we shall return to later.

While the fixed-coupling results we have reported here (and the corresponding results derived for other taggers later in this section) are computed purely in the soft limit, our final results include not just the running coupling effects but also the impact of hard collinear corrections via inclusion of the "B_1" resummation coefficients associated to the splitting kernels, which are beyond our formal double logarithmic accuracy. The full expressions, including running-coupling effects and hard-collinear splittings are given in appendix C.

4.1.2 Matching to the triple-collinear limit

As we argued at the end of section 3.1 (cf. eq. (3.3)), a more accurate calculation of the pre-factor multiplying the Sudakov form factor involves lifting the assumption of strong ordering between the two leading emissions k_1 and k_2 (and that of their softness). One should then use the $1 \to 3$ splitting function for calculating the pre-factor. This is best described using the kinematics of section 3.2, i.e. a system of 3 partons p_1, p_2 and p_3, carrying jet momentum fractions z_i, with $\sum_i z_i = 1$, and with pairwise angles θ_{ij}.

Matching this to the resummed results obtained in section 4.1.1 comes with two conditions. First, we need to make sure that the emissions on which the Sudakov depends (i.e. p, k_1 and k_2 in the previous section), which are by construction the emissions on which the two ζ_{cut} constraints are imposed, are also the relevant emissions that are constrained by the ρ_{min} condition imposed on the pre-factor, cf. the Θ^{tagger} factor in eq. (3.7). Once this condition is satisfied, we will need to map the emissions p, k_1 and k_2 onto the triple-collinear parton system.

For the first condition, as already mentioned, the fact that we are using a gen-k_t declustering guarantees that the emissions picked by the declustering procedure are also the emissions that dominate the pairwise prong masses.

JHEP09(2018)170

We are therefore left with mapping the emissions p, k_1 and k_2 onto the triple-collinear system. Within our double-logarithmic accuracy, this is equivalent to redefine the arguments of the Sudakov (θ_1, ρ_1 and ρ_2) in terms of the kinematics of partons p_1, p_2 and p_3 to account for the lifting of the strong-ordering and softness assumptions. The Sudakov form factor can then still be formally written as eqs. (4.1) and (4.2). However there is a freedom in defining θ_1, ρ_1 and ρ_2 since our only constraint is to recover the proper soft and strongly-ordered limit.

Ultimately, the all-order version of eq. (3.7) becomes

$$\frac{\rho}{\sigma}\frac{d\sigma}{d\rho} = \int d\Phi_3 \, \frac{\langle \hat{P} \rangle}{s_{123}^2} \, \frac{\alpha_s(k_{t1})}{2\pi} \frac{\alpha_s(k_{t2})}{2\pi} \Theta^{\text{jet}} \, \Theta^{\text{tagger}} \, \delta\left(\rho - \frac{s_{123}}{R^2 p_t^2}\right) \mathcal{S}_{\text{Y}_\text{m}\text{-splitter}}(\rho_2; \rho_1, \theta_1), \quad (4.8)$$

where we still have to specify θ_1, ρ_i and k_{ti}. For definiteness, let us consider the situation where the partons p_2 and p_3 are clustered first, followed by a clustering of the (p_2, p_3) pair with p_1 for which we adopted the following prescription:

$$\theta_1 = \theta_{1(2+3)}, \qquad \theta_2 = \theta_{23}, \quad (4.9)$$

$$\rho_1 = \min(z_1, 1-z_1)\theta_1^2, \qquad \rho_2 = \min(z_2, z_3)\theta_2^2, \quad (4.10)$$

$$k_{t1} = \min(z_1, 1-z_1)\theta_1 p_t R, \qquad k_{t2} = \min(z_2, z_3)\theta_2 p_t R, \quad (4.11)$$

$$x_1 \equiv \rho_1/\theta_1^2 = \min(z_1, 1-z_1). \quad (4.12)$$

which can be easily verified to agree with the resummed expressions above, in the soft and strongly-ordered limit. It is perhaps worth re-emphasising that using the form of a double-logarithmic Sudakov form factor multiplying the triple-collinear splitting functions produces uncontrolled terms beyond our leading-logarithmic accuracy. The main purpose of introducing the triple-collinear splitting is to calculate the order α_s^2 pre-factor as precisely as possible while beyond $\mathcal{O}\left(\alpha_s^2\right)$ only double logarithmic terms are controlled.

4.1.3 $\mathbf{Y_m}$-splitter with grooming

In ref. [66] it was shown that Y-splitter was a high performance tool for tagging two-pronged decays only when supplemented with grooming e.g. via mMDT [48] or trimming [6]. It was also shown that the order in which Y-splitter and grooming were used on the jet was crucial to the performance. Grooming jets *after* using Y-splitter resulted in a subleading impact on the crucial large Sudakov suppression of the QCD background seen with Y-splitter, hence maintaining this desirable feature. On the other hand grooming significantly increased the signal efficiency over that seen with plain Y-splitter. The improved signal efficiency and largely unmodified background suppression resulted in striking gains to the signal significance ($S/\sqrt{B}$). On the other hand using grooming tools such as trimming and mMDT *before* Y-splitter was seen in comparison to not give a good performance , since the background Sudakov suppression factor changed from the Y-splitter Sudakov to the less effective trimming and mMDT Sudakov factors respectively. An exception to this, noted in ref. [53], was SoftDrop pre-grooming with positive β (typically, $\beta = 2$), where grooming had a smaller impact on the Y-splitter Sudakov, while the signal efficiency and sensitivity to non-perturbative effects were still considerably improved. In this respect of achieving

JHEP09(2018)170

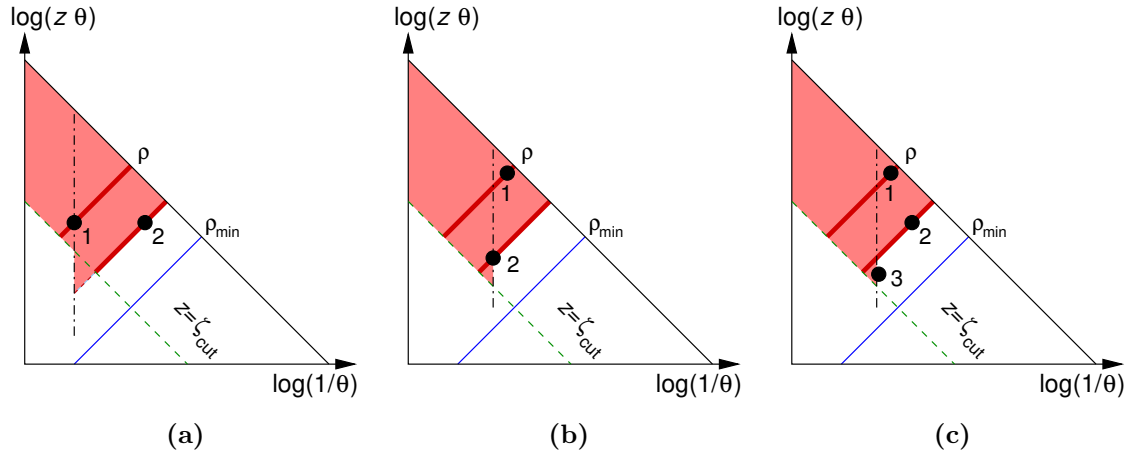


Figure 2. Lund plane corresponding to Y_m-splitter with grooming. Emissions 1 and 2 have respectively the largest and second-largest gen-k_t distance (or mass). The 3 plots (a), (b) and (c) correspond to three different possibilities where the emissions 1, 2 and 3 respectively are the largest-angle emissions that pass the mMDT condition.

a high performance while minimising non-perturbative effects, the SoftDrop ($\beta = 2$) pre-grooming option followed by Y_m-splitter emerged as one of the most effective and reliable methods in the analysis of ref. [53].

In the present case, i.e. for top tagging, it shall turn out to be interesting to explore both $\beta = 0$ (i.e. mMDT) and $\beta > 0$ pre-grooming options. Indeed based on our previous work [53] we may anticipate that pre-grooming with mMDT produces a Sudakov that resembles the mMDT Sudakov suppression factor. As we shall show in the next section, this is also the essential behaviour shown by the CMS tagger (when one considers our IRC safe extensions thereof).

We first consider SoftDrop pre-grooming for $\beta = 0$ i.e. with mMDT. After applying mMDT to the jet we apply Y_m-splitter as adapted by us for top-tagging (see section. 2.2 for details). For simplicity we use the ζ_{cut} condition with the same value for both grooming and Y_m-splitter. One needs to consider three separate cases represented in figure 2:

1. The largest angle emission that passes mMDT is also the largest gen-k_t (or equivalently largest mass) emission from those that remain after grooming (figure 2a). This emission is also the first to be declustered by Y_m-splitter and sets the final jet mass ρ.

2. The largest angle emission that passes mMDT is the second largest gen-k_t (mass) emission and is hence the second emission to be declustered by Y_m-splitter (figure 2b).

3. The largest angle emission that passes mMDT is lower in mass than either of the two emissions declustered by Y_m-splitter (figure 2c). This situation first occurs at the level of three real emissions i.e. at order α_s^3.

To obtain the result corresponding to the first case, consider the emissions k_1 with largest mass $z_1\theta_1^2 = \rho_1 = \rho$ and k_2 with second largest mass $z_2\theta_2^2 = \rho_2$ as before, with $\theta_1 > \theta_2$ and with $z_1, z_2 > \zeta_{\text{cut}}$. The first emission is by assumption the emission that

JHEP09(2018)170

passes mMDT which means that all emissions larger in angle that fail the $\zeta_{\rm cut}$ condition have been groomed away. Here the mMDT stops and one applies $\mathrm{Y_m}$-splitter to the jet, declustering k_1 and k_2 and imposing the $\rho_{\rm min}$ conditions as before. The condition that there are no emissions with mass larger than ρ_2 after mMDT (except the real emission k_1) would produce the standard mMDT Sudakov in ρ_2. However in the present case a key difference with mMDT is the fact that mMDT stops at emission k_1. This implies that emissions k_i at angles smaller than θ_1 that have $z_i < \zeta_{\rm cut}$ are no longer removed by mMDT. If such emissions set a mass larger than ρ_2 they must be vetoed as well which gives an extra contribution to that arising from the mMDT Sudakov down to ρ_2 (as explicitly visible in figure 2a). The result for the overall Sudakov exponent for the primary emission contribution may be written as

$$R^{(1),\text{primary}}(\rho_2;\rho_1,\theta_1) = R_{\text{mMDT}}(\rho_2) + R^{\text{angle}}_{\text{mMDT}}(\theta_1,\rho_2), \tag{4.13}$$

where $R^{(1)}$ represents the Sudakov that applies in the situation that the emission that passes the mMDT condition is also the largest mass emission, $R_{\text{mMDT}}(\rho_2)$ is the standard mMDT Sudakov down to mass ρ_2 and $R^{\text{angle}}_{\text{mMDT}}(\theta_1,\rho_2)$ is the contribution from vetoing emissions that are at smaller angles than θ_1, have $z < \zeta_{\rm cut}$ and set a mass larger than ρ_2. A straightforward calculation in the fixed-coupling approximation gives

$$R^{\text{angle}}_{\text{mMDT}}(\theta_1,\rho_2) = \frac{C_R\alpha_s}{2\pi}\left[\ln^2\frac{\zeta_{\rm cut}\theta_1^2}{\rho_2}\Theta\left(\zeta_{\rm cut}\theta_1^2 > \rho_2\right)\right], \tag{4.14}$$

while the mMDT Sudakov is the usual known result [48]

$$R_{\text{mMDT}}(\rho_2) = \frac{C_R\alpha_s}{2\pi}\left[\Theta\left(\zeta_{\rm cut} > \rho_2\right)\left(2\ln\frac{1}{\zeta_{\rm cut}}\ln\frac{1}{\rho_2} - \ln^2\frac{1}{\zeta_{\rm cut}}\right) + \Theta\left(\rho_2 > \zeta_{\rm cut}\right)\ln^2\frac{1}{\rho_2}\right]. \tag{4.15}$$

For the second case where the emission that passes mMDT is k_2 i.e the second largest in mass the mMDT evolution down to ρ_2 is unmodified and hence we obtain

$$R^{(2),\text{primary}}(\rho_2;\rho_1,\theta_1) = R_{\text{mMDT}}(\rho_2). \tag{4.16}$$

Finally in the third case where an emission k_3 triggers mMDT before either k_1 or k_2, for such an emission to be allowed its mass should be smaller than the mass set by k_2. This contribution cancels against corresponding virtual corrections as for the case of the standard mMDT calculation (corresponding to the small triangular areas with mass smaller than ρ_2 in figures 2b and 2c). Such configurations can thus be ignored.

In addition to the primary emission contributions considered above, there is also a secondary emission contribution to the Sudakov. Secondary emission contributions are not modified by mMDT pre-grooming and hence in either of the two cases considered above, the secondary emission result coincides with that already obtained for $\mathrm{Y_m}$-splitter in eq. (4.7).

Therefore ultimately, the background distribution can still be written in the form of eq. (4.8) now with a primary Sudakov given by eqs. (4.13) and (4.16). Note that, on top of showing different Sudakov suppressions, the different kinematic cases from figure 2 will also be weighted differently when inserted in eq. (4.8).

JHEP09(2018)170

The main feature of the mMDT pre-grooming and Y_m-splitter is that the result closely resembles the known mMDT result i.e. one inherits the Sudakov structure of the pre-grooming method. At small jet masses ρ_2, the Sudakov has an $\alpha_s \ln \zeta_{\rm cut} \ln \rho_2$ behaviour which gives a smaller suppression than the $\alpha_s \ln^2 \rho_2$ behaviour of Y_m-splitter. Differences from the pure mMDT result arise due to the extra $R^{\rm angle}_{\rm mMDT}$ term and due to secondary emissions. In both cases the argument of the double logarithm produced has a ratio involving either θ_1^2/ρ_2 or ρ/ρ_2 which again can be expected to be modest contributions, except possibly at small values of $\rho_{\rm min}$.

For the case of SoftDrop pre-grooming using a general $\beta > 0$ the same general arguments hold as for the $\beta = 0$ mMDT results derived above. The only change one needs to make is that the condition for an emission to pass the SoftDrop constraint now becomes $z > \zeta_{\rm cut}\theta^\beta$. Hence we get

$$R^{\rm angle}_{\rm SD}(\theta_1, \rho_2, \beta) = \frac{C_R \alpha_s}{2\pi} \frac{2}{2+\beta} \ln^2 \frac{\zeta_{\rm cut}\theta_1^{2+\beta}}{\rho_2} \Theta\left(\zeta_{\rm cut}\theta_1^{2+\beta} > \rho_2\right) \tag{4.17}$$

while $R_{\rm mMDT}(\rho_2)$ is replaced by the SoftDrop Sudakov down to ρ_2:

$$R_{\rm SD}(\rho_2, \beta) = \frac{C_R \alpha_s}{2\pi} \left[\Theta(\zeta_{\rm cut} > \rho_2) \left(\ln^2 \frac{1}{\rho_2} - \frac{2}{2+\beta} \ln^2 \frac{\zeta_{\rm cut}}{\rho_2}\right) + \Theta\left(\rho_2 > \zeta_{\rm cut}\right) \ln^2 \frac{1}{\rho_2}\right]. \tag{4.18}$$

The primary-emission Sudakov therefore becomes

$$R^{(1),\rm primary}_{\rm SD+Y_m\text{-splitter}}(\rho_2; \rho_1, \theta_1) = R_{\rm SD}(\rho_2) + R^{\rm angle}_{\rm SD}(\theta_1, \rho_2), \tag{4.19}$$

while we also have as for the mMDT pre-grooming case the contribution $R^{(2),\rm primary}_{\rm SD+Y_m\text{-splitter}}(\rho_2) = R_{\rm SD}(\rho_2)$.

As for the case of Y_m-splitter, our final results also include running-coupling effects and hard-collinear splittings, and are given in appendix C.

4.2 `TopSplitter` and CMS$^{\rm 3p,mass}$

As discussed before and explicitly demonstrated in appendix A, the CMS tagger is unsuitable for precise theoretical computations involving top jets, due to its IRC unsafety at high p_t. Instead we shall consider our extensions of the tagger i.e. the CMS$^{\rm 3p,mass}$ variant and the method we call `TopSplitter` which also originates from the CMS tagger. In fact the `TopSplitter` method turns out to be the most amenable to a resummed calculation to the accuracy we were able to achieve for Y_m-splitter and Y_m-splitter with grooming, namely the resummation of logarithms of $\rho/\rho_{\rm min}$ and $\zeta_{\rm cut}$ on the same footing as logarithms of ρ or $\rho_{\rm min}$. Hence we shall consider this method first.

For the `TopSplitter` tagger we shall need to consider Cambridge-Aachen (C-A) declustering of the jet rather than the gen-k_t declustering relevant for Y_m-splitter. Therefore now we need to consider emissions that are soft and strongly ordered in angle. In particular if we assume that emission k_1 is declustered first then there is a veto on any emission at an angle larger than θ_1 with $z > \zeta_{\rm cut}$ (emissions with $z < \zeta_{\rm cut}$ are groomed away by the primary declustering procedure). On declustering the emission k_1 we are left with k_1

JHEP09(2018)170

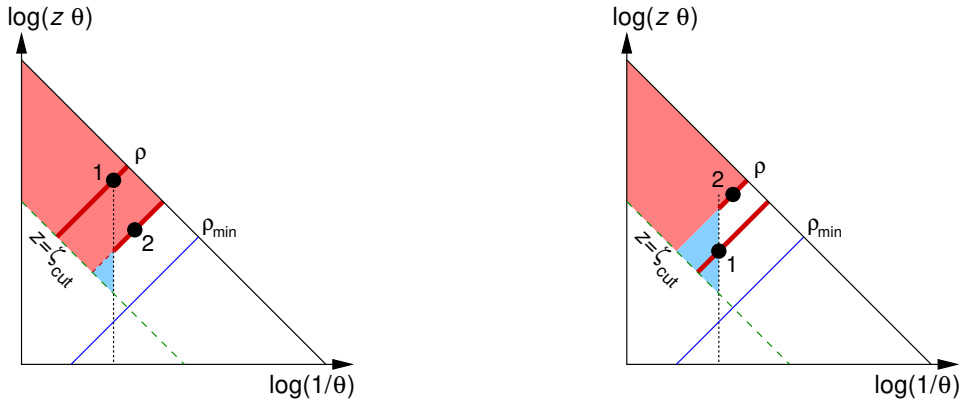


Figure 3. Lund plane corresponding to the `TopSplitter` tagger. Emissions are labelled such that emission 1 is the first one selected by the tagger, i.e. $\theta_1 \gg \theta_2$, and emission 2 dominates the prong mass. The two plots correspond to (a) $\rho_1 \gg \rho_2$ and (b) $\rho_1 \ll \rho_2$. The red shaded region shows the veto region for emissions with $z > \zeta_{\text{cut}}$ and $\rho > \rho_2$ while the blue shaded triangle region corresponds to the extra contribution that arises from requiring that there are also no emissions with $\theta > \theta_1$ and $z > \zeta_{\text{cut}}$.

and a massive prong p. The tagger then proceeds to decluster prong p. The declustering produces a second emission k_2, also with $z > \zeta_{\text{cut}}$, which, by definition of the `TopSplitter` method, dominates the mass of the prong and hence we impose a veto on all emissions that set a larger mass than $\rho_2 = z_2\theta_2^2$.[11] The veto on emissions in the prong is only active for emissions with $z > \zeta_{\text{cut}}$. To see this, note that emissions in the prong at angles larger than that of emission k_2 and with $z < \zeta_{\text{cut}}$ are groomed away by the secondary declustering step of the tagger, while emissions with angles smaller than that of k_2 and with $z < \zeta_{\text{cut}}$ cannot dominate the mass in any case. Furthermore, with the above `TopSplitter` procedure of selecting k_2 so that it dominates the prong mass, only emissions k_1 and k_2 enter into the construction of the minimum pairwise mass and contribute to the pre-factor that multiplies the Sudakov exponent.

The situation is depicted in the Lund plane in figure 3. Emissions k_1 and k_2 are shown corresponding to $\theta_1 \gg \theta_2$ with either $\rho_1 \gg \rho_2$ (figure 3a) or $\rho_1 \ll \rho_2$ (figure 3b). The red shaded region shows a veto on all emissions with mass larger than ρ_2 and $z > \zeta_{\text{cut}}$ as argued above. A further blue shaded region shows additional emissions that are vetoed since they have $\theta > \theta_1$ and $z > \zeta_{\text{cut}}$. Analogous to the case of Y_m-splitter we can write a resummed result of the form given in eq. (4.8) but with a different Sudakov form factor $\mathcal{S}_{\texttt{TopSplitter}}$ which can be written as

$$\mathcal{S}_{\texttt{TopSplitter}} = \exp\left[-R_{\texttt{TopSplitter}}\left(\rho_2; \rho_1, \theta_1\right)\right]. \tag{4.20}$$

The Sudakov exponent $R_{\texttt{TopSplitter}}$ receives contributions from both vetoes on primary and secondary emissions i.e. $R_{\texttt{TopSplitter}} = R^{(\text{primary})}_{\texttt{TopSplitter}} + R^{(\text{secondary})}_{\texttt{TopSplitter}}$. The veto on primary

[11] Note that the fact that the declustered emission is the one that dominates the prong mass owes precisely to our construction of `TopSplitter` which picks the emission with largest $p_{ti}\theta_i^2$ as a product of the declustering.

JHEP09(2018)170

emissions was discussed above and its explicit form is:

$$R^{(\text{primary})}_{\texttt{TopSplitter}}(\rho_2;\rho_1,\theta_1) = \int \frac{d\theta^2}{\theta^2}\frac{dz}{z}\,\frac{\alpha_s(z\theta p_t R)C_R}{\pi}\,\Theta(z>\zeta_{\text{cut}})\,\Theta(z\theta^2>\rho_2 \text{ or } \theta>\theta_1). \tag{4.21}$$

We should also consider the case of secondary emissions from k_1 which would prevent emission k_2 from being declustered if they have mass larger than ρ_2 and energy fraction w.r.t. the jet p_t greater than ζ_{cut}, hence we must also veto such emissions. In this case we obtain

$$R^{(\text{secondary})}_{\texttt{TopSplitter}}(\rho_2;\rho_1,\theta_1) = \int^{\theta_1^2} \frac{d\theta^2}{\theta^2}\frac{dz}{z}\,\frac{\alpha_s(zx_1\theta p_t R)C_A}{\pi}\,\Theta(x_1 z>\zeta_{\text{cut}})\,\Theta(zx_1^2\theta^2>\rho_2). \tag{4.22}$$

In the above equation z and θ are the energy fraction and angle of the secondary emission with respect to the emitting parent k_1, which itself has energy fraction x_1 and angle θ_1 with respect to the hard jet p_t and direction respectively. Also, as for the Y_m-splitter case discussed in section 4.1, one could also have a situation where emission k_2 is emitted as a secondary emission from k_1. Again, this situation is automatically accounted for by matching to the triple-collinear splitting function. The main features of the results, for the above defined contributions, are again best illustrated by using a fixed-coupling approximation and we refer to appendix C for our full expressions including running-coupling results and hard-collinear splittings. It is instructive to further separate the contributions to $R^{(\text{primary})}_{\texttt{TopSplitter}}$ and write it as the sum of the contributions due to the red shaded region ($z\theta^2 > \rho_2$) and the extra blue shaded region ($z\theta^2 < \rho_2$ and $\theta > \theta_1$) in figure 3, $R^{(\text{primary})}_{\texttt{TopSplitter}} = R^{(\text{red})}_{\texttt{TopSplitter}} + R^{(\text{blue})}_{\texttt{TopSplitter}}$. For the red shaded region we obtain just the usual result for the mMDT already mentioned in eq. (4.15),

$$R^{(\text{red})}_{\texttt{TopSplitter}}(\rho_2) = R_{\text{mMDT}}(\rho_2), \tag{4.23}$$

while the blue triangle contributes as below:

$$R^{(\text{blue})}_{\texttt{TopSplitter}}(\rho_2;\theta_1) = \frac{\alpha_s C_R}{2\pi}\left[\ln^2\frac{\rho_2}{\zeta_{\text{cut}}\theta_1^2}\,\Theta(\rho_2>\zeta_{\text{cut}}\theta_1^2) - \ln^2\frac{\rho_2}{\zeta_{\text{cut}}}\,\Theta(\rho_2>\zeta_{\text{cut}})\right]. \tag{4.24}$$

The corresponding expression for $R^{(\text{secondary})}_{\texttt{TopSplitter}}$ in a fixed-coupling approximation is also simple to obtain:

$$R^{(\text{secondary})}_{\texttt{TopSplitter}}(\rho_2;\rho_1,\theta_1) = \frac{C_A\alpha_s}{2\pi}\left[\ln^2\frac{x_1\rho}{\rho_2}\Theta\left(\rho_2<x_1\rho\right) - \ln^2\frac{\zeta_{\text{cut}}\rho}{\rho_2}\Theta\left(\rho_2<\zeta_{\text{cut}}\rho\right)\right]. \tag{4.25}$$

Some comments about the results obtained here are in order. Although we have identified various different contributions, the most relevant contribution to the tagger behaviour comes from the $R^{(\text{red})}_{\texttt{TopSplitter}}$ term which is essentially the same Sudakov as was originally obtained for the modified mass-drop tagger (mMDT/SD($\beta=0$)) [48]. This is because as we mentioned before the largest logarithms are those in jet mass, and in the limit of small jet mass $\rho_2 \ll \zeta_{\text{cut}}$, we see a single-logarithmic Sudakov suppression due to $R^{(\text{red})}_{\texttt{TopSplitter}}$. The remaining terms i.e those due to secondary emissions and $R^{(\text{blue})}_{\texttt{TopSplitter}}$ produce, in the

JHEP09(2018)170

limit of small jet mass, only leading logarithms of $\zeta_{\rm cut}$ and ρ_1/ρ_2 or equivalently $\rho/\rho_{\rm min}$. We retain these terms here for the reasons mentioned before, namely to assess their impact on the tagger behaviour.

Lastly we are left with mapping the variables entering the Sudakov onto the triple-collinear set of emissions p_1, p_2, p_3 as for the $\mathrm{Y_m}$-splitter case. We again exploit our freedom to choose the precise definitions, with the only constraint being to recover the leading-logarithmic results after taking the soft and strongly-ordered limits. We then define (again for the case where p_2 and p_3 are clustered first followed by p_1 with the (p_2, p_3) pair):

$$\rho_1 = z_1(1-z_1)\theta_1^2, \qquad \rho_2 = z_2 z_3 \theta_2^2, \tag{4.26}$$

$$k_{t1} = z_1(1-z_1)\theta_1 p_t R, \qquad k_{t2} = z_2 z_3 \theta_2 p_t R, \tag{4.27}$$

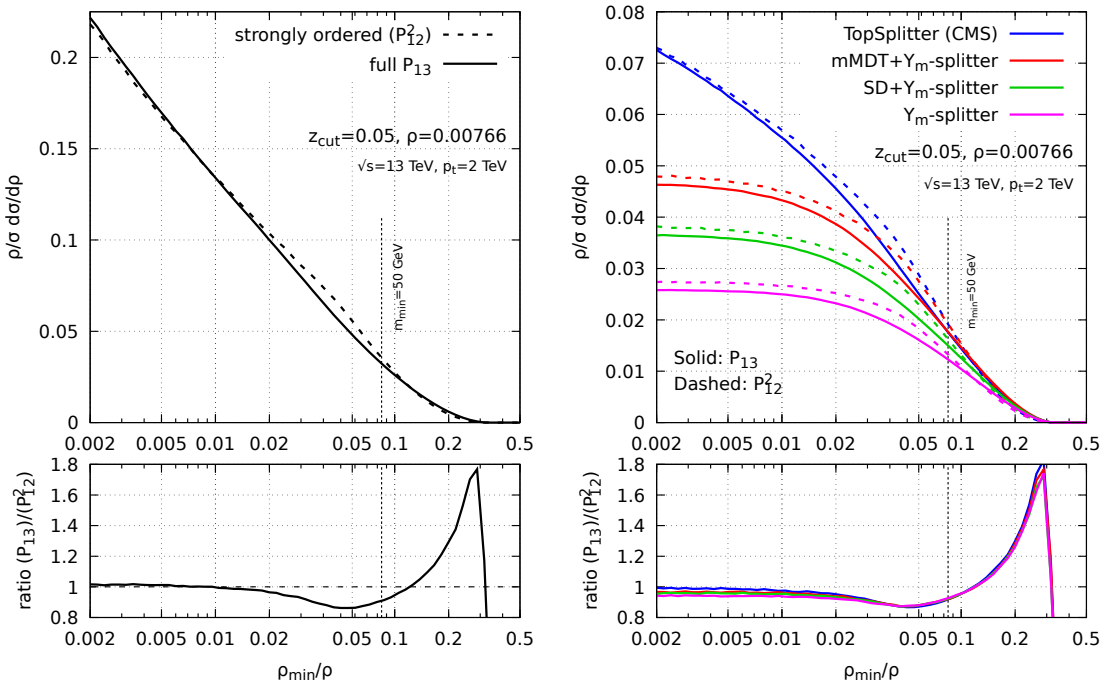
$$\theta_1 = \theta_{1(2+3)}, \qquad \theta_2 = \theta_{23}. \tag{4.28}$$

Having discussed the case of `TopSplitter`, we now turn to the $\mathrm{CMS}^{\mathrm{3p,mass}}$ tagger. The main difference with the `TopSplitter` case is simply the fact that when declustering a prong, one takes the largest angle emission within the prong, i.e the one declustered first by the tagger, as a product of the declustering. This emission is not guaranteed to dominate the mass of the prong however, and hence there is a possible mismatch between the emissions that are declustered by the tagger and those that enter the minimum pairwise mass condition, in particular for the pairwise mass that involves one of the branches from the secondary declustering and the branch left intact from the primary declustering. Configurations for which there is such a mismatch produce additional corrections with leading-log terms involving logs of $\rho_{\rm min}/\rho$ and $\zeta_{\rm cut}$. In this case, the resummation of such terms is possible but substantially more complicated than for the `TopSplitter` case. Since the behaviour of the $\mathrm{CMS}^{\mathrm{3p,mass}}$ tagger now depends on up to three emissions (two emissions dominating the $\zeta_{\rm cut}$ condition and one additional emission dominating the $\rho_{\rm min}$ condition), the matching with the triple-collinear splitting is no longer systematically achievable. Given the small impact of the additional terms, both numerically and for our understanding of the tagger behaviour, we will neglect them. Hence, in our analytical treatment, the result for the $\mathrm{CMS}^{\mathrm{3p,mass}}$ variant of the CMS tagger coincides with that we presented for `TopSplitter`. We shall later verify that the performance of $\mathrm{CMS}^{\mathrm{3p,mass}}$ and `TopSplitter`, as given by parton shower models, is in fact consistent with our observations.

5 Results

We have previously noted that given the fact that ρ and $\rho_{\rm min}$ are not widely disparate in practice, the approximation of strong angular-ordering may not be sufficient to satisfactorily capture the impact of top taggers on QCD jets. It is thus clearly interesting to attempt to compare the results obtained using the triple-collinear splitting functions to those produced by the strong angular-ordering approximation, especially since it is the latter picture that is effectively included in most parton shower descriptions of QCD jets. At the same time given that we have devoted most of this article to discussing resummation effects in detail, it is also worthwhile to consider the numerical importance of resummation

JHEP09(2018)170



(a) No resummation **(b)** With resummation

Figure 4. Comparison of the results obtained for quark jets in the strongly-ordered limit (dashed curves) with the results using the full triple-collinear splitting function (solid curves), (a) without including resummation effects, (b) including resummation effects. In both cases, the top panel shows the distributions $\rho/\sigma\, d\sigma/d\rho$ and the lower panel shows the ratio between the results including the full triple-collinear splitting and the strongly-ordered limit.

and especially to understand the relative contributions of various different contributions to the Sudakov exponents. We shall devote the current section to these studies.

5.1 Numerical impact of triple-collinear and resummation effects

We first discuss the effect of including the full triple-collinear splitting function instead of working in the strongly-ordered limit. This is shown in figure 4 both with and without inclusion of resummation effects. As expected, in the limit $\rho_{\min} \ll \rho$, both results agree, although the ratio does not exactly converge to 1 in the case where resummation effects are included simply because the Sudakov form factor weights differently different regions of phase-space. For situations closer to what is used for phenomenology, i.e. $m_{\min} \approx 50\,\text{GeV}$ (highlighted by the vertical dotted line on the plots), the inclusion of the full triple-collinear splitting function only introduces a correction of about 10% once all effects are considered. This means that unless one uses larger values of $\rho_{\min}$, closer to the endpoint of the distribution, the effect of the triple-collinear splitting functions is modest and should not substantially modify pure parton shower descriptions of top tagging mistag rates.

Next, we move on to the discussion of resummation effects. We have plotted in figure 5 our results for $\rho/\sigma d\sigma/d\rho$, obtained from (4.8) adapted to each tagger, varying $\rho_{\min}$. The

JHEP09(2018)170

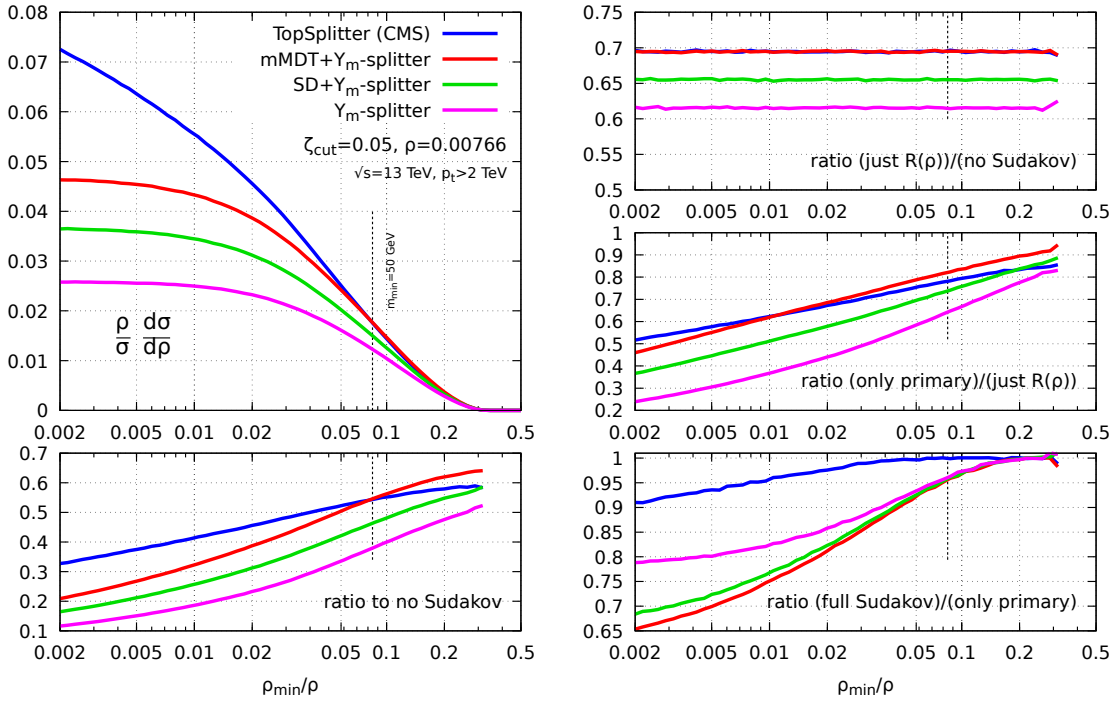


Figure 5. Study of the resummation effects for the various taggers as a function of ρ_{min}. The top-left plot shows the distribution $\rho/\sigma d\sigma/d\rho$, and the bottom-left plot shows the overall Sudakov effect (the ratio of $\rho/\sigma d\sigma/d\rho$ with and without the Sudakov form factor). On the right, different levels of approximation to the Sudakov were made: (top) simply using the jet-mass Sudakov (plain, SD or mMDT depending on the tagger) down to the scale ρ, (middle) considering instead the full Sudakov from primary emissions, and (bottom) adding secondary emissions. Each plot shows the ratio to the previous approximation.

plot shows the overall effect of the resummation on the left and the effects split in a series of contributions on the right. Focusing on the left plot first, we see that the resummation has a sizeable impact, suppressing the QCD background by a factor $\sim 2-3$, in the phenomenological region. As expected, the effects increase when further reducing ρ_{min}.

The series of plots on the right of figure 5 aim at gauging the relative importance of various types of contribution to the Sudakov. Here we studied 4 different levels of approximations for the Sudakov form factor. First, we generated results without a Sudakov form factor (i.e. with just the leading-order α_s^2 calculation with the $1 \to 3$ splitting function), then those with just a simplified Sudakov exponent involving resummation of only logarithms of ρ via the radiator $R(\rho)$. $R(\rho)$ is taken as the plain jet-mass radiator for the case of Y_m-splitter, the appropriate groomed jet-mass radiator for Y_m-splitter with grooming and the mMDT radiator for `TopSplitter`. Next, we studied results involving only primary emissions and finally the full result including all double logarithms on the same footing and including secondary emissions. The three plots show the ratio of the results obtained with one approximation relative to what was obtained with the previous (more crude) approximation.

JHEP09(2018)170

The top plot shows the effect of the jet-mass like Sudakov $\exp(-R(\rho))$, compared to not including any Sudakov. This is expected to capture the dominant logarithms, i.e. the most enhanced by logarithms of ρ, in the phenomenological region. We see indeed that they come with a large suppression factor. Furthermore, we see that the suppression is larger for $\mathrm{Y_m}$-splitter than for SD+$\mathrm{Y_m}$-splitter, itself more suppressed than mMDT+$\mathrm{Y_m}$-splitter and `TopSplitter`, following the size of the region vetoed by the Sudakov factor.

In the middle plot, we now include the full primary Sudakov (recall that the plot shows the relative impact of the full primary Sudakov compared to just including "$\exp(-R(\rho))$"). Although this is expected to have a smaller effect than the resummation of the dominant logarithms of ρ, typically trading a logarithm of ρ for a logarithm of ρ_{min}/ρ, we see that the effect remains sizeable, in fact, almost as large as the $\exp(-R(\rho))$ Sudakov. Again, decreasing ρ_{min} the effect of the full primary Sudakov becomes more pronounced, dominating the trend seen at small ρ_{min} in the overall Sudakov effect (bottom-left plot). As before, the Sudakov suppression is reduced when the level of pre-grooming is increased. Note that, although this is not explicitly shown in the plot, we have also tested the relative importance of the "blue" primary Sudakov compared to the (dominant) "red" contribution in the case of `TopSplitter` and found that it had a very small effect of order of a few percent at most.

Finally, the bottom-right plot studies the effect of adding the secondary emission suppression, shown as the ratio of the results with the full Sudakov compared to only including primary emissions. This is expected to involve only additional logarithms of ρ_{min}/ρ and ζ_{cut} and it indeed turns out to have a small impact in the region relevant for phenomenology, again increasing when moving to smaller values of ρ_{min}.

Before moving to a comparison to parton-shower Monte-Carlo simulations, we note that in the strongly-ordered limit and using a fixed-order approximation for the Sudakov, it is possible to simplify (at least some of) the integrations over emissions 1 and 2 in (4.8). For the case of $\mathrm{Y_m}$-splitter, all the integrations can be performed analytically. The full analytic result explicitly highlights the expected logarithmic dependences and nicely reproduces the various trends observed in figure 5. For other taggers, we could only obtain interesting expressions in the limit $\rho_{\mathrm{min}}/\rho \ll \zeta_{\mathrm{cut}} \ll \rho$ or $\zeta_{\mathrm{cut}} \ll \rho_{\mathrm{min}}/\rho \ll \rho$ which again involved the expected dominant logarithms and corroborated the behaviours seen numerically in figure 5.

5.2 Comparison to parton showers

Having obtained analytic results for the different taggers, in this section we shall compare the analytics to Monte Carlo (MC) simulations. Here we will be interested in parton level MC results, since we are comparing to a purely perturbative calculation, i.e. we shall use Pythia 8.230 [60] parton-level events to compare to our all-order resummed analytic predictions. The impact of non-perturbative effects (hadronisation and MPI) will be considered when we discuss tagger performances in the next section.

We consider jet production in dijet processes at the LHC with $\sqrt{s} = 13\,\mathrm{TeV}$. We first focus on subprocesses involving only quark jets in the final state (by selecting $qq \to qq$ hard matrix elements) and will discuss gluon jets (obtained through the $gg \to gg$ matrix element) later. We define jets using the anti-k_t algorithm [61], as implemented in FastJet [67, 74],

JHEP09(2018)170

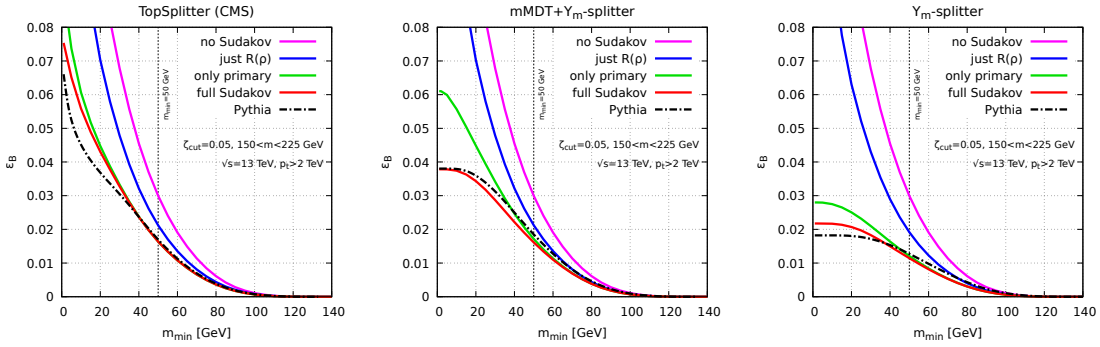


Figure 6. Comparison between analytic results and Pythia simulations for the QCD background efficiency. Results are plotted as a function of the m_{min} cut. For the analytic results, we have included the same levels of approximation to the Sudakov as in figure 5.

and use a jet radius $R = 1$ and a transverse momentum selection cut such that $p_t > 2\,\text{TeV}$. For all the taggers we use $\zeta_{\text{cut}} = 0.05$. To study the background efficiency (mistag rate) of the taggers we work in a mass window around the signal mass $150 < m < 225\,\text{GeV}$.

First we compare in figure 6 our analytic predictions to parton shower results for the background efficiency, obtained by integrating over the signal mass window, as a function of m_{min}. We consider the case of `TopSplitter` for which the LL resummation structure resembles that for the CMS$^{\text{3p,mass}}$ variant of the CMS tagger but where we control all double logarithms and not just those in ρ, as would be the case for CMS$^{\text{3p,mass}}$. We also consider both Y_{m}-splitter alone as well as its combination with pre-grooming via mMDT and SoftDrop ($\beta = 2$). For each tagger we show the analytic results using the same four levels of approximation to the Sudakov as for figure 5. We also show the result from Pythia for comparison.

Let us first consider the `TopSplitter` results. We note that the best agreement across all m_{min} values is provided by the use of the full Sudakov. In the phenomenologically relevant region with $m_{\text{min}} \sim 50\,\text{GeV}$ one obtains perfect agreement with Pythia by using the full Sudakov while using $R(\rho)$ alone in the Sudakov exponent gives a noticeable difference with Pythia which increases at small m_{min}. At smaller m_{min} beyond the phenomenologically relevant region, one sees that Pythia starts to depart somewhat from the analytic resummation. The feature in the Pythia results at small m_{min} is not evident in the analytic calculations but occurs in a region where the pure parton shower predictions are potentially subject to significant non-perturbative corrections. To see this one notes that jet masses $\sim 40\,\text{GeV}$ can be produced by emissions with energy $\sim 1\,\text{GeV}$ in conjunction with a hard parton with energy $\sim 2\,\text{TeV}$. Hence the difference between Pythia's parton shower (without hadronisation) and analytics at such low masses is largely of academic interest. As already observed in figure 5, secondary emissions have only a modest role over most of the m_{min} range though at smaller m_{min} there is evidence that they have the effect of shifting the resummed result closer towards those from Pythia.

Next we discuss the plain Y_{m}-splitter case. We again observe a good general agreement of the full resummed result with Pythia across a broad range of m_{min} with some difference

JHEP09(2018)170

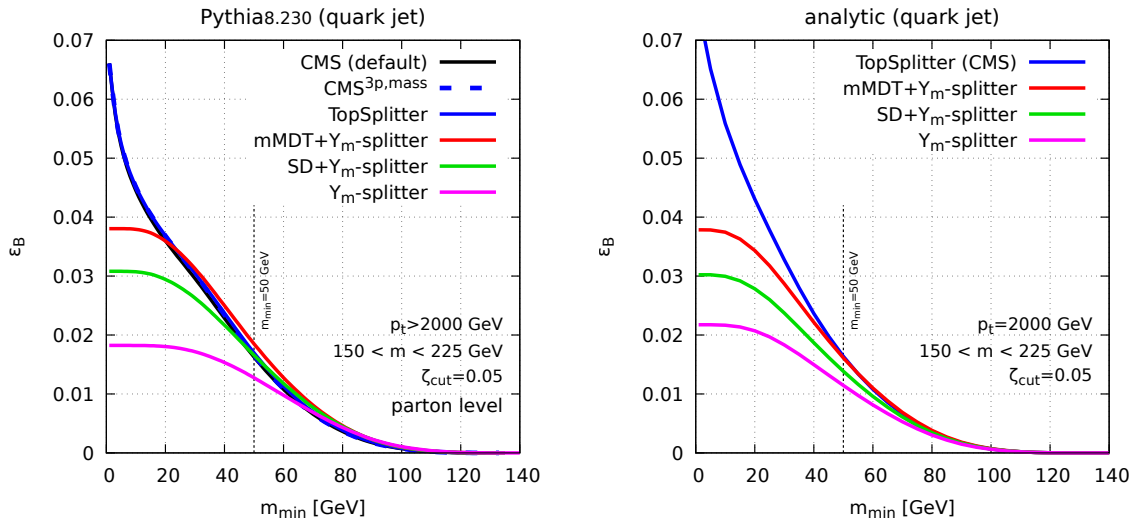


Figure 7. Background efficiency as a function of the m_{min} cut for various taggers. The left plot shows results obtained from Monte Carlo simulations with the Pythia8 generator and the right plot shows the results of our analytic calculations discussed in the main text.

visible at smaller m_{min} values somewhat beyond the phenomenologically relevant values. Secondary emissions play a more visible role here at smaller m_{min} than for `TopSplitter` and noticeably move the result closer to that from Pythia. However we again note here that, as for the case of `TopSplitter` mentioned above, a comparison between analytics and Monte Carlo at low masses is subject to modification by non-perturbative effects which have been neglected in our analytical estimate and turned off in Pythia.

Similar comments apply to the groomed variants of Y_m-splitter with again a good general agreement for the full resummed result with Pythia and a demonstrable improvement from including resummation effects beyond those in the naive $R(\rho)$ function.

For ease of comparison between the different taggers, we also show in figure 7 results for the background efficiency or mistag rate ϵ_B of the different taggers as a function of m_{min} on the same plot, with MC results shown on the left and analytic results on the right. Taggers with a lower ϵ_B suppress the background more, which is desirable, although the final performance depends also on the impact of the taggers on signal jets, which we discuss in the next section. As far as the main purpose of this section is concerned — comparing expectations from analytics with results from MC parton showers — one can say that the general features of the MC results are well reproduced by the analytics. In particular one notes the ordering in the performance of taggers that is predicted by the analytics also emerges in the parton shower results. We would naturally expect, as has also been observed before [53] for the case of two-pronged signal jet substructure, that Y_m-splitter suppresses the background most effectively due to the large double-logarithmic Sudakov form factor obtained there. This expectation is clearly borne out by both the analytical and MC results. We would also expect that Y_m-splitter with pre-grooming using SoftDrop ($\beta = 2$) would give the SoftDrop Sudakov which reduces the background less than Y_m-splitter but

JHEP09(2018)170

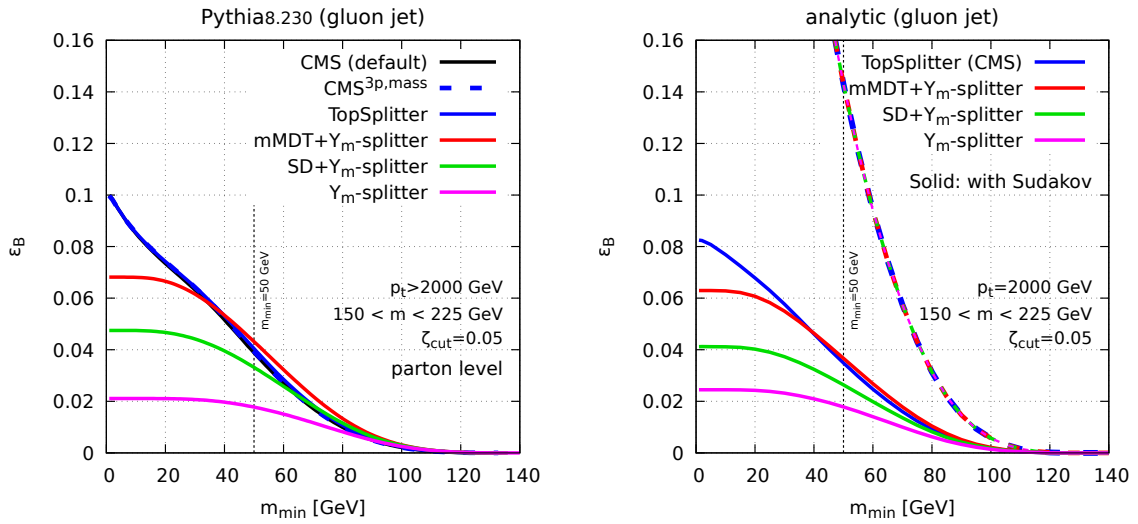


Figure 8. Same as figure 7 but this time for gluon-initiated jets. Also shown in the analytic calculations on the right is the result without the application of the Sudakov form factors, using just the pre-factor from the triple-collinear splitting function for an initial gluon.

still more than other methods with a smaller Sudakov suppression. This also emerges in the MC studies albeit at somewhat smaller m_{min} than predicted by the analytics. Next one would expect the mMDT pre-groomed Y_m-splitter which has an essentially mMDT style Sudakov suppression (a smaller suppression than that expected from SoftDrop ($\beta = 2$)) however still retaining the Y_m-splitter result at the level of secondary emissions. Again, especially at slightly smaller m_{min} relative to the analytics, the MC results follow this expected trend. Lastly we have the case of `TopSplitter` where relative to the Y_m-splitter based methods one would expect the smallest Sudakov suppression since both primary and secondary emissions are impacted by the ζ_{cut} condition. Once again MC results confirm this expectation.

Perhaps most crucially, at the phenomenological working point of 50 Gev there is no significant difference visible in the analytics between the results for `TopSplitter` and those for mMDT+Y_m-splitter and this is also what emerges in the parton shower results. A small difference is visible between the above two methods and SoftDrop ($\beta = 2$) in the analytics while the spread in MC results is not visible yet. Finally there is a clear difference between the above three methods and Y_m-splitter visible in both analytics and MC.

A further comment is due on the MC results for the original CMS tagger, labelled as CMS (default) in figure 7 compared to those for the $\text{CMS}^{\text{3p,mass}}$ variant and `TopSplitter`. The MC predictions for these methods are in remarkably good agreement with one another, being virtually coincident over the entire m_{min} range. This suggests that both $\text{CMS}^{\text{3p,mass}}$ and `TopSplitter` are good IRC safe alternatives to the CMS tagger, with `TopSplitter` having the advantage of being more amenable to an accurate resummation of all double-logarithmic enhanced terms.

For completeness, we show the results obtained for gluon-initiated jets in figure 8. Again, the overall behaviour and ordering between the taggers are correctly reproduced

JHEP09(2018)170

by the analytic calculations. However, we see larger quantitative differences, in particular at small $m_{\min}$ than what was seen for quark-initiated jets. This is likely due to the fact that the Sudakov factors are larger in the case of gluon jets, e.g. we see a suppression by a factor $\sim 4-6$ for $m_{\min} = 50\,\text{GeV}$, relative to the result without a Sudakov shown using the dashed lines in figure 8. Therefore, subleading corrections, not included in our calculation, also have a larger impact. Since the QCD background in the boosted limit is dominated by quark-initiated jets ($> 80\%$ at 2 TeV), this has little impact on practical applications (and we will focus on quark-initiated jets in what follows).

Lastly, from the results of this section alone it may be tempting to conclude that the Y_m-splitter method should be the preferred option for top tagging. Indeed for studies of jet substructure with signal jets initiated by a colourless electroweak boson decay, the impact on the QCD background was most often the decisive factor in dictating tagger performance [48]. In the present case with a coloured parton also initiating the signal, one also needs to consider the impact of QCD radiation for the signal jets too. The final word on tagger performance will therefore involve also an analysis of the signal, which is the subject of the next section.

6 Signal efficiency and performance

Having studied the action of top taggers on QCD background jets we shall here consider the case of signal jets. As a consequence we shall then produce ROC curves purely from analytics and compare those to curves obtained from Monte Carlo event generators. Finally we shall examine here the role of non-perturbative effects.

6.1 Signal efficiency

To study signal jets we consider the case of boosted top production in a given hard process, with the top exhibiting a three-pronged hadronic decay to a b quark and a hadronically decaying W boson, which form the constituents of the top jet at leading order. One then has to take account of the action of the top taggers on the three-pronged system.

The basic leading-order result can be obtained as for the QCD case

$$\frac{d\sigma}{d\rho} = \int d\Phi_3 \; |\mathcal{M}_{t\to bqq}|^2 \, \delta\left(\rho - \frac{s_{123}}{R^2 p_T^2}\right) \Theta^{\text{tagger}}\left(\rho_{\min}, \zeta_{\text{cut}}\right) \Theta^{\text{jet}}, \tag{6.1}$$

where $\mathcal{M}_{t\to bqq}$ is the matrix element for the top decay process, $d\Phi_3$ the three-body phase space in the collinear limit as before, and with the tagger and jet finding conditions now applied to the top decay products. For an on-shell top quark we have that $\rho = \rho_t = \frac{m_t^2}{R^2 p_T^2}$. The above result is simple to compute numerically and in implementing the result for the squared matrix-element we have made the usual substitution of the W boson and top quark propagators by a Breit-Wigner form.

In contrast to the case of colour singlet (e.g. W/Z/H) decays however, the above tree-level result is not sufficient to give a good description of substructure and tagging efficiency for top jets. The obvious reason for this is that the top quark is a coloured object and hence one must consider the role of accompanying QCD radiation. Soft gluons are emitted

JHEP09(2018)170

by the virtual top quark in the course of producing an on-shell final-state top and further emissions occur during the top-decay process.

While multiple soft emissions are generally thought to be less significant in heavy quark production than is the case for light quarks, in the highly boosted region where $m_t^2 \ll p_t^2$ the top quark can be considered as being essentially light and all-order resummation effects start to become important. In particular, in the boosted regime, we may ignore the effect of the dead cone [75] of order $m_t^2/p_t^2 \sim \rho$, which does not play a role at the logarithmic accuracy in ρ that we are concerned with here. At the same time while soft gluon emission in top production and decay is known to have a complicated emission pattern [76] especially for gluon energies near or below the top width, again at our leading double-logarithmic accuracy where we are concerned with only soft and relatively collinear radiation, these complications can be neglected. Hence one can treat the radiation as for the massless case to be essentially stemming from a single fast moving colour line along the jet direction.

The soft emissions from the top quark, which are recombined into the top jet, will contribute to the mass of the jet. We consider the jet mass distribution after the further application of the various top-taggers which, as for the case of the QCD background, places constraints on the accompanying soft gluon emission within the top jet, and leads to Sudakov form factors multiplying the top production and decay probability.

Given our calculations in the previous sections for QCD jets it should simple to understand the basic features of the resulting Sudakov factor multiplying the leading-order electroweak factor eq. (6.1), for the different taggers. Two important differences from the QCD case are firstly that the electroweak top decay treated via the pre-factor already dominates the jet mass condition since it produces a jet mass equal to the top mass for an on-shell top decay, and secondly the fact that the W boson radiated off the top is a colour singlet and hence does not radiate gluons unlike a primary gluon emission from say a quark jet which, as we have accounted for in the QCD jet case, acts as a source of relevant secondary emissions.

A treatment of signal jets at the same level as we have performed for background jets, i.e. one where $\ln \zeta_{\text{cut}}$ and $\ln \rho/\rho_{\text{min}}$ are also resummed, proves to be substantially more complicated, e.g. because the ordering of the three prongs found by the taggers' double declustering procedure will in general involve different combinations of the b quark and the W decay products. Additionally, gluon emissions from the top could contribute to shifting its mass. Their effect would depend on both their interplay with the tagger (including the dynamics of the three top decay products) and on the mass window cut imposed on the tagged jet. The latter introduces yet another non-trivial scale in the calculation.

As a consequence of these extra complications, for top jets we shall not try to achieve a full double-logarithmic accuracy including logarithms of ρ_{min}/ρ and ζ_{cut} on equal footing with logarithms of ρ. Instead, we shall primarily focus on getting the dominant behaviour in ρ. Since we are also interested in investigating the strongly-ordered limit, we will use a mass-like Sudakov down to the scale $\min(\rho_1, \rho_2)$. This means that `TopSplitter` and mMDT+Y_m-splitter would use a mMDT Sudakov R_{mMDT}, eq. (4.15), SD+Y_m-splitter would use a Soft-Drop Sudakov R_{SD}, eq. (4.18), and ungroomed Y_m-splitter would use a plain jet-mass Sudakov $R^{(\text{primary})}_{\text{Y}_\text{m}\text{-splitter}}$, all taken at the scale $\min(\rho_1, \rho_2)$. In the case of the

JHEP09(2018)170

Y_m-splitter taggers, $\min(\rho_1, \rho_2)$ is by definition equal to ρ_2 in the strongly-ordered limit and is the natural scale for the Sudakov. In the case of `TopSplitter`, one could instead expect a mixture of the θ_1^2 angular scale and the ρ_2 mass-like scale (cf. figure 3). Since one can trade θ_1^2 for ρ_1 up to subleading logarithms of ζ_{cut}, the scale $\min(\rho_1, \rho_2)$ is also appropriate.

We have also investigated the impact of other choices which have the same formal accuracy as the choice mentioned above. Specifically, we have checked that the corrections, compared to using the same form of the Sudakov, but taken at the scale ρ, were within 20% in the phenomenologically relevant region, which should really be seen as the ballpark uncertainty on our calculations for signal jets. Additionally, we have also considered using the full primary Sudakov form factor, derived for quark jets in section 4, which should also achieve the job of capturing the bulk of the radiation from the top and bottom quarks in the strongly-ordered limit. We found results very similar to the ones obtained with the simpler mass-like Sudakov taken at the scale $\min(\rho_1, \rho_2)$ and hence we continue to use the latter as our default form.

Our results for the top distribution can therefore be written as

$$\frac{\rho}{\sigma}\frac{d\sigma}{d\rho} = \int d\Phi_3 \, |\mathcal{M}_{t\to bqq}|^2 \, \rho\delta\left(\rho - \frac{s_{123}}{R^2 p_T^2}\right) \Theta^{\text{tagger}}\left(\rho_{\min}, \zeta_{\text{cut}}\right) \Theta^{\text{jet}} \exp\left[-R^{(\text{mass})}_{\text{tagger}}\right], \quad (6.2)$$

with

$$R^{(\text{mass})}_{\texttt{TopSplitter}} = R^{(\text{mass})}_{\text{mMDT+Y}_\text{m}\text{-splitter}} = R_{\text{mMDT}}(\min(\rho_1, \rho_2)), \quad (6.3)$$

$$R^{(\text{mass})}_{\text{SD+Y}_\text{m}\text{-splitter}} = R_{\text{SD}}(\min(\rho_1, \rho_2)), \quad (6.4)$$

$$R^{(\text{mass})}_{\text{Y}_\text{m}\text{-splitter}} = R^{(\text{primary})}_{\text{Y}_\text{m}\text{-splitter}}(\min(\rho_1, \rho_2)), \quad (6.5)$$

where ρ_1 and ρ_2 are defined according to eq. (4.26) for `TopSplitter` and eq. (4.10) for the Y_m-splitter variants.

With the top-decay result supplemented by Sudakov form factors we ought to be able to capture the main features seen in the performance of taggers using Monte Carlo event generators. To that aim, we show the signal efficiency as a function of $m_{\min}$ in figure 9. We see that the effects of the Sudakov seem over-estimated in the analytics relative to the MC results, but that ordering between the taggers is reasonably reproduced. We also show in figure 9, via the dashed curves, the impact of not including any Sudakov form factor in the analytic calculations for signal jets, in which case the analytic results are the same for all taggers and differ substantially from Pythia results.

The analytical result shows also a minor difference between `TopSplitter` and mMDT+Y_m-splitter, where our analytic calculation predicts a larger suppression for the latter which is not observed in the Pythia simulations. Since our treatment of top jets does not reach the same accuracy as what was obtained for QCD jets in section 4, such differences between analytics and MC results should be expected. In this precise case of comparing `TopSplitter` with mMDT+Y_m-splitter, the observed difference has to be driven by the different definitions for ρ_1 and ρ_2 (as a function of the parton kinematics from the

JHEP09(2018)170

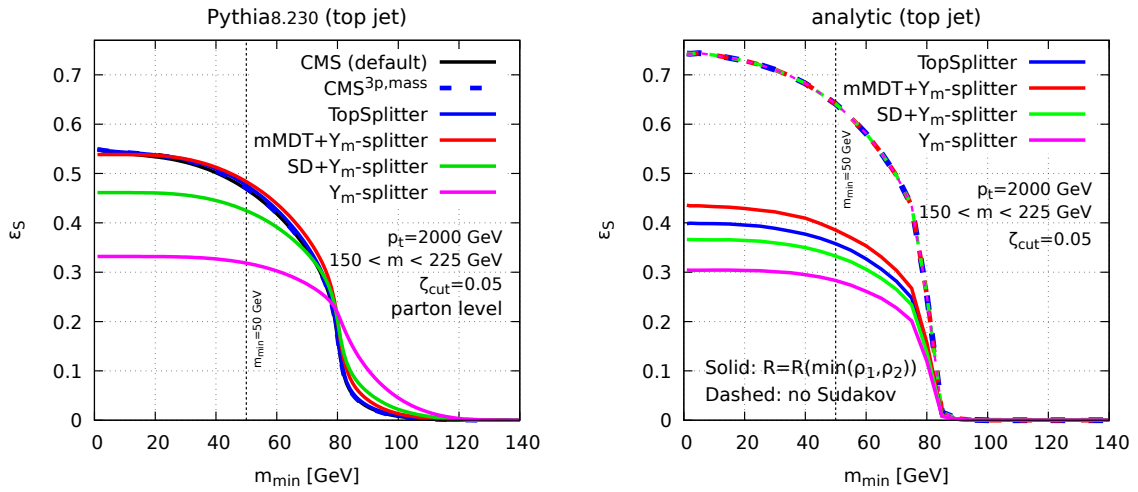


Figure 9. Same as figure 7 this time for signal (top) jets. Also shown is the analytic result without the inclusion of a Sudakov form factor.

triple-collinear splitting), eqs. (4.19) and (4.26). Indeed, while for QCD jets we expect emissions with momentum fractions close to $\zeta_{\rm cut}$, the situation will be more symmetric for top jets, meaning in practice a smaller value for ρ_1 and ρ_2 in the case of `TopSplitter` compared to mMDT+$\mathrm{Y_m}$-splitter. This smaller value translates in a larger $R_{\rm mMDT}(\min(\rho_1,\rho_2))$ and hence a smaller signal efficiency for `TopSplitter` (again, compared to mMDT+$\mathrm{Y_m}$-splitter).[12] These differences are clearly beyond our targeted accuracy.

The main message that emerges from our studies in the current section is that a Sudakov form factor is essential to describe the behaviour of the taggers on signal jets. The basic form of the Sudakov that we have used in the signal case is sufficient to understand the main features of top taggers but a more precise statement on tagging efficiency, as we have for instance for QCD background jets, would require a more detailed analytic calculation for signal jets which is beyond the scope of our present work. Finally we remark that on the Monte-Carlo side, we also note that no observable differences are seen in figure 9 between the various CMS-related taggers. In the following section we will look at tagger performance using both parton shower and analytic methods.

6.2 Performance and non-perturbative effects

We now discuss the performance of the various taggers using the standard ROC curves which show the background efficiency or mistag rate plotted against the signal efficiency. For a given signal efficiency, the tagger with the lowest mistag rate is considered the most performant.

A point that is worth noting is that due to a very similar Sudakov suppression seen for the signal and the background, any gains that are produced by Sudakov suppression of

[12] If we were instead using a simple mass Sudakov taken at the scale ρ for top jets — achieving the same formal accuracy as what have used so far — we would obtain the exact same signal efficiency for `TopSplitter` and mMDT+$\mathrm{Y_m}$-splitter.

JHEP09(2018)170

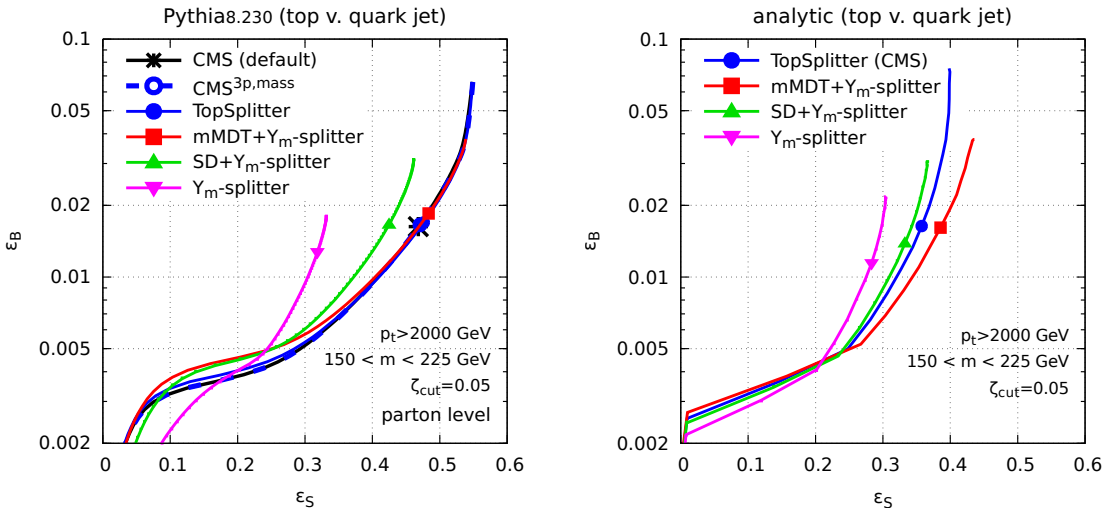


Figure 10. ROC curves corresponding to the $m_{\min}$ scan shown in figure 7. This time, the Pythia results also include the default CMS tagger for comparison (in black). The thicker points correspond to the default value $m_{\min} = 50\,\text{GeV}$.

emissions from a QCD jet are largely offset by a corresponding suppression of the signal. Therefore a large Sudakov suppression is not necessarily beneficial for the case of top tagging in contrast to the tagging say of colour singlet electroweak and Higgs bosons. An exception to the above may in principle be expected to occur for the case of gluon jets where the Sudakov suppression of the background is indeed more than that for the signal, owing to the larger colour factor for emissions from gluon jets. In general however the background will be a mix of quark and gluon jets, with the quark jet component being dominant at higher p_t where Sudakov effects are stronger for a fixed jet mass. For this reason we start by looking at the highest phenomenologically relevant p_t values, i.e. in the TeV region, with quark jets alone.

Figure 10 shows the ROC curves one obtains for $p_t = 2\,\text{TeV}$ with a pure quark jet background. The curves correspond to tagging in a mass window $150 < m < 225\,\text{GeV}$, use $\zeta_{\text{cut}} = 0.05$, as done throughout our studies, and both parton level results from Pythia (left) as well as analytical results (right) are shown. A first observation is that except at fairly low signal efficiencies, a larger Sudakov results in a larger mistag rate for a given signal efficiency i.e. a worse performance. Based on this observation we would expect to see a definite ordering in the results for tagger performance. From an analytical viewpoint the smallest Sudakov suppression belongs to `TopSplitter` (and $\text{CMS}^{\text{3p,mass}}$) and mMDT+$\text{Y}_{\text{m}}$-splitter taggers. A somewhat larger Sudakov suppression is seen in our analytic calculations for SoftDrop ($\beta = 2$) and the largest suppression is for Y_{m}-splitter with a double-logarithmic plain-mass type Sudakov. This globally corresponds to the ordering seen in the analytic ROC curves above a signal efficiency $\epsilon_s \sim 0.2$. Instead for lower signal efficiencies the ordering is inverted so that taggers with a large Sudakov perform better. A larger signal efficiency is however what we clearly would desire from a phenomenological viewpoint and so taggers with a smaller Sudakov would be favoured. The results from

JHEP09(2018)170

the Pythia parton shower are in general agreement with our analytical expectations and a similar ordering is seen for those results. However, while the analytic results indicate some difference in performance between `TopSplitter` and mMDT+Y_m-splitter, these are seen to perform essentially identically in Monte Carlo studies at higher signal efficiencies. Such differences can be easily ascribed to the less precise treatment of the signal Sudakov in the analytics, discussed in section 6.2.

It is also noteworthy that no differences are seen in parton shower results between the default CMS tagger, `TopSplitter`, CMS^{3p,mass} and the mMDT+Y_m-splitter methods, at least for efficiencies $\epsilon \gtrsim 0.35$. Apart from the IRC unsafe CMS tagger all these other methods have the common feature of an essentially mMDT style Sudakov at low masses, albeit with differences of detail. The main message appears to be that it is possible to create a family of taggers which are IRC safe but give a similar performance to the default CMS tagger with the family being defined by its key feature of an mMDT style Sudakov.

Finally, we show in appendix D that our observations are still valid at lower jet p_t (1 TeV or even lower down to about 500 GeV) albeit with a reduced difference between the taggers, attributed to a reduction of the phase-space available for radiation and the decreased importance of Sudakov effects. In particular, it means that the `TopSplitter` can be considered as an effective and more robust replacement of the CMS top tagger over a wide range of p_t values relevant to phenomenology.

A discussion of tagger performance and reliability is not complete without a discussion of non-perturbative effects. As we mentioned before, ROC curves produced using event generators are subject to a theoretical uncertainty. However estimating the uncertainty on such ROC curves is a far from simple exercise even conceptually, largely owing to the sole reliance on Monte Carlo event generators. It is however safe to say that results for methods which are either IRC unsafe like the CMS tagger, or those that receive large non-perturbative corrections, must be considered to suffer from a larger theoretical uncertainty than IRC safe methods which additionally show only small non-perturbative corrections, even if that uncertainty cannot be easily quantified. Therefore examining the impact of non-perturbative corrections is important in order to more reliably assess the performance of a tagger.

Figure 11 shows a plot of the signal efficiency divided by the square-root of the background efficiency, also known as the signal significance, which quantifies the tagger performance on the y axis, while at the same time showing the sensitivity to non-perturbative effects on the x axis. To estimate the latter, the measure we have chosen is the ratio of the background efficiency at hadron level to that at parton level to assess the impact of hadronisation (for a fixed $m_{\min}$ cut) in the left plot, and the ratio of the background efficiency at hadron level including UE to that without the UE on the right plot of figure 11. Similar studies have also been carried out in the past for the case of W/Z/H tagging, for instance in ref. [53].

A number of points follow from consideration of figure 11. Firstly the inclusion of non-perturbative effects as measured by the deviation of the results from unity along the x axis does not have a very substantial effect for a wide range of signal significances, with the notable exception of Y_m-splitter which due to its inherent lack of grooming suffers

JHEP09(2018)170

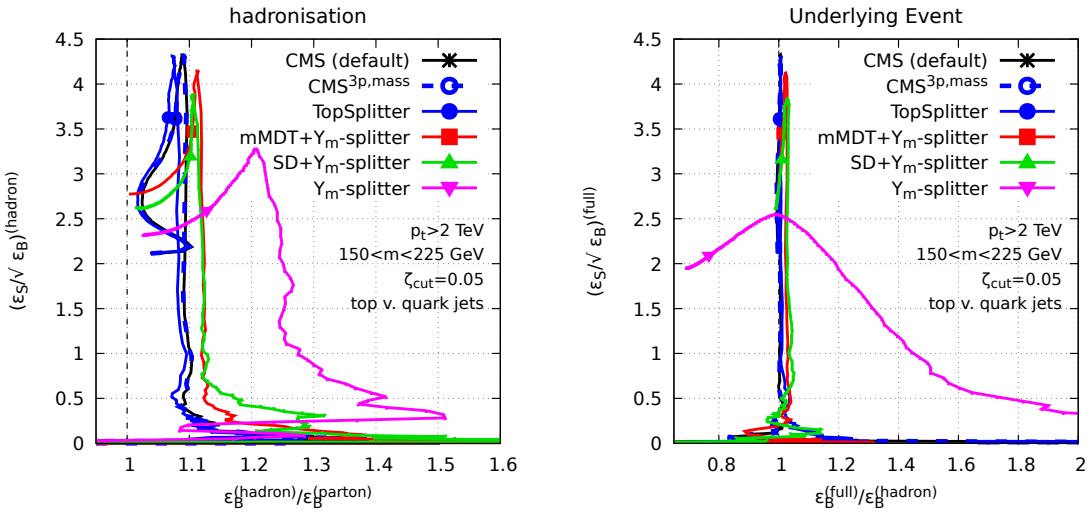


Figure 11. Both plots show how the sensitivity to non-perturbative effect (x axis) and the discriminating power (y axis) evolve when varying the cut on m_{min} for different taggers. Left: effects observed when switching on hadronisation, i.e. going from parton level to hadron level. Right: effects observed when including the Underlying Event. The symbols correspond to $m_{\text{min}} = 50\,\text{GeV}$.

significantly from both hadronisation and UE effects. For other methods the hadronisation effects are no larger than around the 15% level with even smaller effects for the default CMS tagger, CMS$^{\text{3p,mass}}$ and `TopSplitter`. A remarkable degree of similarity and effectiveness across methods is seen with regard to removal of contamination from the UE with the only exception again being Y_{m}-splitter which is entirely expected from previous studies of Y-splitter [66] . On this basis the non-perturbative studies do not have any sizeable impact on the main conclusions that we reached from the parton level analysis before. The best taggers in terms of sheer performance are also the ones which are most resilient to non-perturbative effects, which is in contrast to what is seen for W/Z/H tagging where Y_{m}-splitter followed by grooming using mMDT or trimming far outperforms other methods, but at the cost of large non-perturbative effects. When one factors in IRC safety which is a key element in assessing the robustness of a tool, one should replace the CMS tagger with either CMS$^{\text{3p,mass}}$ or `TopSplitter` which leads to no loss of performance. If one further factors in analytic calculability then the `TopSplitter` method emerges as the one over which we have the best theoretical control certainly for phenomenologically relevant m_{min} values, while at the same time maximising the performance,

Finally, we have also tested that these conclusions remain valid down to (at least) jet p_t's of 500 GeV, where the `TopSplitter` non-perturbative corrections remain in the 15-20% range, followed by mMDT+Y_{m}-splitter and SD+Y_{m}-splitter around 30%. In this context, it would also be interesting to investigate a version of the Y_{m}-splitter tagger where one first applies Recursive Soft Drop [77], e.g. with two layers of grooming with $\beta = 0$ ("recursive mMDT") or $\beta = 2$, or infinite recursion with $\beta = 2$.

JHEP09(2018)170

7 Conclusions

In this article we have studied aspects of top-tagging from first principles of QCD using the methods of analytic resummation supported by Monte Carlo studies. The aim has been to try and identify the main physical principles that are at play and hence better understand the effect of using top tagging methods on background and signal jets.

To consider an explicit example of a tool that has been used directly in the context of LHC phenomenology, we started by studying the CMS top tagger. Here we discovered the issue of the CMS tagger's collinear unsafety at high p_t, with potential adverse consequences for precise theoretical predictions. The collinear unsafety of the CMS tagger was seen to originate in the step of selecting three prongs from four on the basis of their energy. Hence we proposed variants of the CMS tagger that are explicitly IRC safe even at high p_t. One variant that we named CMS$^{\text{3p,mass}}$ selects three prongs from four based on the invariant mass while another variant we named `TopSplitter`, selects the emission which dominates the prong mass in the soft limit as a product of the declustering. While both methods are collinear safe, `TopSplitter` is simpler from the viewpoint of the analytical calculations we aimed at in this article.

In addition to the above methods which are all based on C-A declustering of a jet, we introduced new methods based on gen-k_t declustering. Here we adapted our previously suggested Y_m-splitter method [53] for use in top tagging. Our earlier studies based on W/Z/H tagging have shown that Y_m-splitter when additionally supplemented by some form of grooming has the potential to be a high performance tool [53], which led us in this paper to investigate a combination of grooming with Y_m-splitter.

For the QCD background, we carried out leading logarithmic in jet mass analytical calculations for Y_m-splitter, mMDT + Y_m-splitter, SoftDrop ($\beta = 2$) + Y_m-splitter, `TopSplitter` and the CMS$^{\text{3p,mass}}$ taggers. For all but the last case we were able to supplement a resummation of large logarithms in the jet masses ρ or ρ_{min} with additional resummation of leading logarithms in ζ_{cut} and ρ/ρ_{min}, counting them on the same footing as logarithms of ρ or ρ_{min}. Our results were seen to take the form of an order α_s^2 pre-factor which multiplies a Sudakov exponent arising from resummation. We argued that an accurate calculation of the pre-factor should require going beyond the picture of strong ordering in angles or energies of emissions and should involve instead the use of triple-collinear splitting functions. Such splitting functions are not included in the Pythia shower, or indeed in other well-known showers, commonly used to study tagger performance. Ultimately however the triple-collinear splitting functions gave a somewhat modest $\sim 10\%$ effect for $m_{\text{min}} = 50\,\text{GeV}$ relative to the strong angular-ordering approximation which is in principle correctly included in the Pythia shower.

A comparison of our analytical calculations for QCD background jets with the Pythia parton shower revealed general good agreement across a wide range of m_{min} values and excellent agreement at the phenomenological working point of $m_{\text{min}} = 50\,\text{GeV}$ for all methods for which resummed results exist (i.e. all our taggers except the collinear-unsafe default CMS tagger.) Our full resummation including logarithms of ζ_{cut} and ρ/ρ_{min} was seen to be required in order to obtain better agreement with Pythia and becomes crucial

JHEP09(2018)170

to include especially at small $\rho_{\rm min}$. The basic conclusion from our analytic versus Monte Carlo comparisons is that we appear to have very good analytic control over top taggers studied and developed in this paper, when applied to QCD jets

In terms of performance we have found that, as may readily be anticipated, taggers with a larger Sudakov suppression are more effective at removing the QCD background. Our analytics suggest that Y_m-splitter with its plain jet mass type Sudakov suppression should therefore produce the lowest background mistag rate and this expectation is confirmed by the Pythia shower. We also found that the ordering of background mistag rates between taggers which emerges in our analytics is indeed reproduced in the Pythia shower at parton level. It is noteworthy that the default CMS tagger produces identical results in the Pythia parton shower to our newly-proposed alternatives CMS$^{\rm 3p,mass}$ and `TopSplitter`.

We also studied the effect of top taggers on signal jets initiated by a top quark. The resulting jet mass distributions also receive a Sudakov suppression factor similar to that for QCD background, due to the colour charge of the top quark, although here our analytical calculations were less precise than those we carried out for the QCD background and we neglected retaining full control over logarithms of $\zeta_{\rm cut}$ and $\rho/\rho_{\rm min}$. We discovered that the impact on signal together with background is such that, at high p_t, taggers with a larger Sudakov suppression generally perform less well, at least for reasonably large signal efficiencies, than those with a smaller Sudakov, assuming a pure quark background. Therefore the plain Y_m-splitter method is less performant than Y_m-splitter with SoftDrop ($\beta = 2$) pre-grooming, in turn less performant than Y_m-splitter with mMDT pre-grooming which produces a Sudakov which resembles more closely the mMDT Sudakov, rather than the plain mass type of Sudakov seen with Y_m-splitter. The mMDT pre-groomed Y_m-splitter, `TopSplitter`, the default CMS tagger and the CMS$^{\rm 3p,mass}$ tagger gave essentially identical performance at signal efficiencies larger than about 0.35, i.e. the Pythia ROC curves for these methods coincide. For lower signal efficiencies the default CMS tagger, CMS$^{\rm 3p,mass}$ and `TopSplitter` still remain very close to one another in performance. The analytics however suggested some modest differences also between `TopSplitter` and mMDT+Y_m-splitter even at higher signal efficiencies, which was not seen in the parton shower studies.

We evaluated also the role of non-perturbative effects. With the exception of plain Y_m-splitter we noticed that all the taggers are quite resilient to non-perturbative effects due to their inherent grooming aspect. Hadronisation effects were found to be no more than $\sim 15\%$ for phenomenologically relevant values of signal efficiency while the underlying event contribution was generally less than a few percent.

Overall we found that it is possible to develop a range of IRC safe methods, for tagging three-pronged jet substructure, which can be understood from first principles of QCD (i.e. largely independently from MC results). The more performant techniques at high p_t are ones which have a common feature of a smaller mMDT style Sudakov suppression. Our alternatives to the default CMS tagger are virtually identical in performance to the CMS tagger, with `TopSplitter` emerging as our preferred method, due to the higher accuracy of the corresponding analytical calculation. While it is our understanding that the CMS collaboration have in the meantime moved away from using the CMS top tagger in experimental studies involving top tagging, our analytical studies based on its variants have

JHEP09(2018)170

helped reveal some of the key physical principles involved in the problem of top tagging using jet substructure, in the same way that early studies of the mass drop tagger (mDT) paved the way for a concrete understanding of two-pronged jet substructure, giving rise to today's tools such as mMDT and SoftDrop [48, 50].

Most importantly, armed with a range of methods and a detailed understanding of their impact, we have acquired both some flexibility and insight which will be important for also studying the optimal combination of top taggers with jet shape variables such as N-subjettiness or energy correlation functions, and to explore the origin and nature of the further gains due to using jet shapes. In future work we intend to enhance our understanding of top-tagging by considering such combinations, which are widely used in LHC studies, also from an analytical viewpoint.

Acknowledgments

MD thanks the U.K's STFC for financial support via grant ST/P000274/1. MD also thanks the CERN theoretical physics department for a scientific associateship and for hospitality during the course of this work as well as the School of Physics and Astronomy at the University of Manchester for sabbatical leave which facilitated this work. MD thanks the CEA Saclay and the French CNRS for financial support and hospitality during the course of this work. JR thanks the ERC and the U.K.'s STFC for financial support via grant ST/N504178/1. GS is supported in part by the French Agence Nationale de la Recherche, under grant ANR-15-CE31-0016.

A Collinear unsafety of the CMS tagger with no ΔR cut

The collinear unsafety of the CMS top tagger can be explicitly shown using a fixed-order study. As described in section 2.1, the collinear unsafety appears when some substructure can be found in both primary prongs. This requires at least 4 particles in the jet. One method to obtain such jets is to generate e^+e^- collisions with QCD particles in the final state and to boost the whole event along the x axis to obtain a collimated jet. In practice, we have used the Event2 [78, 79] generator with a centre-of-mass energy of 80 GeV, boosted to 1 TeV. We then reconstruct the jets with the Cambridge/Aachen [62] with $R = 1$ and keep jets above 500 GeV. We measure the cross-section for the jets to pass either the CMS top tagger or the CMS$^{3p,\text{mass}}$ with $\zeta_{\text{cut}} = 0.05$ and $m_{\text{min}} = 30$ GeV.

This setup allows us to study the boosted jets to order α_s (with up to 3 particles in the jet) and α_s^2 (with up to 4 particles in the jet). We note that since the tagger requires at least 3 particles in the jet, the order α_s is actually the leading order here and there is no contribution from the 2-loop contribution at order α_s^2, which is not available in Event2.

In figure 12 we plot the cross-section for jets passing the CMS tagger as a function of the internal cut-off used in Event2. For the default CMS top tagger, we see an obvious logarithmic dependence on the cut-off as a result of the collinear unsafety of the tagger. Switching instead to the CMS$^{3p,\text{mass}}$ tagger, the cross-section converges rapidly when the cut-off is decreased, showing that the collinear unsafety has been cured.

JHEP09(2018)170

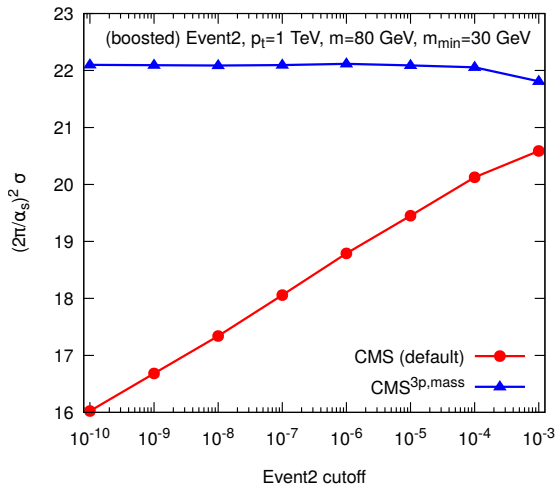


Figure 12. Cross-section for passing the CMS tagger as a function of the Event2 cut-off.

We note that in the context of a resummed calculation, this collinear unsafety will be tamed by the associated Sudakov form factor, i.e. the default CMS tagger although collinear unsafe, remains Sudakov safe [80, 81]. This potentially explains why little differences are seen in practice between the CMS, CMS^{3p,mass} and `TopSplitter` taggers in full Monte-Carlo simulations. The collinear unsafety would however make it delicate to reliably estimate the theoretical uncertainties associated with the CMS top tagger.

B Variants of the CMS and Y-splitter taggers

Here, we consider additional variants of the CMS and Y_m-splitter taggers. We first define them and then briefly compare them to the default versions discussed in the main text.

B.1 Definition of the variants

The variants are as follows:

1. z_{cut} *condition*: one can modify the CMS tagger such that one uses a $z_{\rm cut}$ type condition in performing the decomposition. This would involve a cut of the form $\frac{\min(p_{T,i},p_{T,j})}{p_{T,i}+p_{T,j}} > z$ which uses the local p_T of the cluster being decomposed, i.e. $p_{Ti}+p_{Tj}$ instead of the global p_T of the hard jet in the denominator as is the case for the $\zeta_{\rm cut}$ condition used originally by CMS and in the main body of the paper.

2. ρ_{min} *condition only on secondary declustering*: variants where the taggers proceed exactly like the default CMS^{3p,mass}, `TopSplitter` and Y_m-splitter but we impose the $m_{\rm min}$ condition only on the 2 prongs produced in the secondary declustering instead of all 3 pairwise combinations.

B.2 Declustering with a $\zeta_{\rm cut}$ or $z_{\rm cut}$ condition

We start by comparing the performance of the taggers when imposing a $z_{\rm cut}$ condition (compared to the default $\zeta_{\rm cut}$ condition). Our results are plotted in figure 13 for Pythia simulations (left plot) and for our analytic calculation (right plot).

JHEP09(2018)170

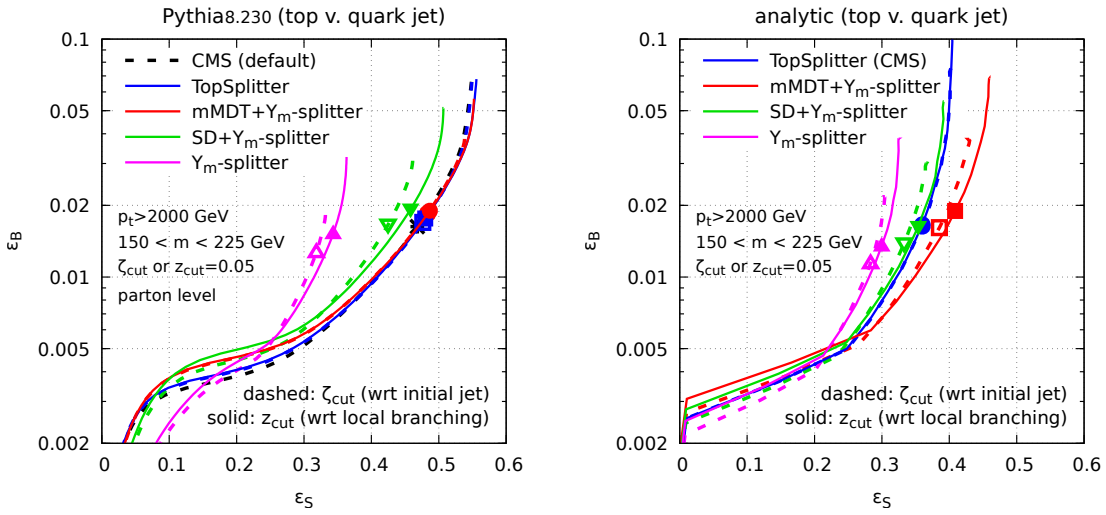


Figure 13. Comparison of the taggers performance when using a z_{cut} condition (solid lines) compared to the default ζ_{cut} condition (dashed lines). Left: Pythia simulations, right: our analytic calculation.

Overall, we see little differences between the two variants, in particular, for the CMS-related taggers. For the Y_m-splitter taggers, we see a small difference in performance, with the versions using a ζ_{cut} condition performing marginally better at small signal efficiencies and the versions using a z_{cut} condition performing slightly better at large signal efficiency. Our analytic calculations reproduce these differences correctly although the predicted difference in the case of mMDT+Y_m-splitter is not seen in the Pythia simulations. This difference seems driven by the signal (top) efficiency which is anyway not as well controlled as the QCD background in our analytic calculations.

B.3 Minimum pairwise condition v. secondary declustering condition

In figure 14, we compare the performance of the variants of the taggers derived by imposing the ρ_{min} condition only on the secondary declustered branch, to the default `TopSplitter` and Y_m-splitter. We see little difference between the default version (dashed lines) and the corresponding variant (solid lines) at large signal efficiency. However, at small signal efficiency, the default version of the taggers clearly outperforms the variants, i.e. favouring the case where the ρ_{min} condition is imposed on the minimum pairwise mass. These behaviours are well captured by our analytic results.

C Analytic expressions for the radiators

For completeness, we list in this appendix our results for the radiators derived in section 4, including running-coupling effects and hard-collinear splittings. Our results are written in terms of the "building blocks" introduced in [52]. For our purpose in this paper, the only

JHEP09(2018)170

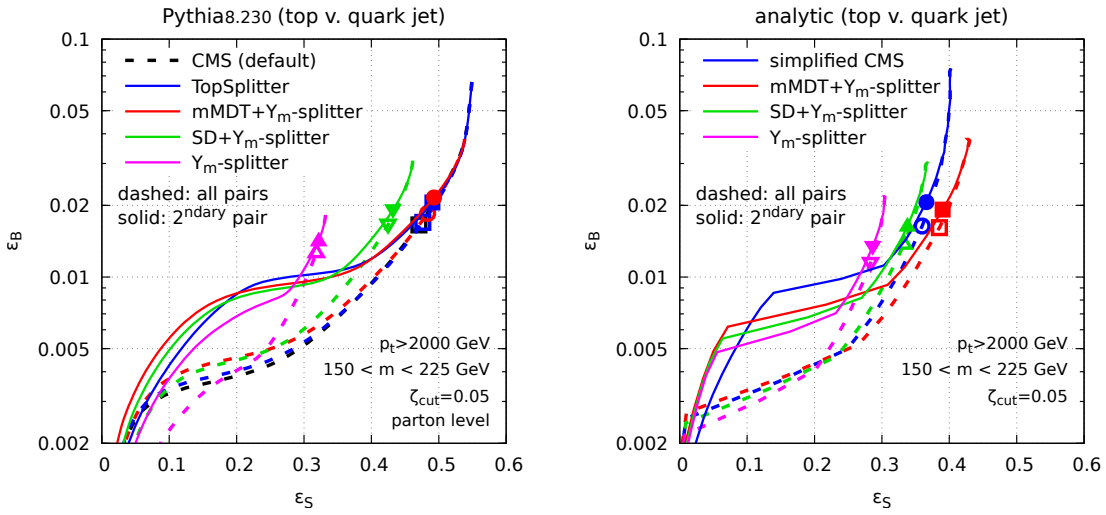


Figure 14. Comparison of the taggers performance when imposing the $\rho_{\min}$ condition only on the secondary declustering (solid lines) compared to imposing the $\rho_{\min}$ condition on all three pairwise masses (dashed lines). Left: Pythia simulations, right: our analytic calculation.

building block we need is[13]

$$T_{-\beta,2}(\kappa_{\min},\kappa_{\max};C_R) = \int_0^1 \frac{d\theta^2}{\theta^2}\frac{dz}{z}\frac{\alpha_s(z\theta p_t R)}{2\pi}\Theta(z<\kappa_{\max}\theta^\beta)\Theta(z\theta^2>\kappa_{\min}) \tag{C.1}$$
$$= \frac{C_R}{2\pi\alpha_s\beta_0^2}\left[\frac{U(\lambda_{\max})}{1+\beta}+U(\lambda_{\min})\right.$$
$$\left.-\frac{2+\beta}{1+\beta}U\left(\frac{\lambda_{\max}+(1+\beta)\lambda_{\min}}{2+\beta}\right)\right]\Theta(\kappa_{\max}>\kappa_{\min})$$

with $\alpha_s \equiv \alpha_s(p_t R)$, $\lambda_i = 2\alpha_s\beta_0 \ln 1/\kappa_i$ and $U(\lambda) = (1-\lambda)\ln(1-\lambda)$. In particular, we have

$$T_{02}(\kappa_{\min},\kappa_{\max};C_R) = \frac{C_R}{2\pi\alpha_s\beta_0^2}\left[U(\lambda_{\max})+U(\lambda_{\min})-2U\left(\frac{\lambda_{\max}+\lambda_{\min}}{2}\right)\right]\Theta(\kappa_{\max}>\kappa_{\min}) \tag{C.2}$$

which corresponds to the standard mass Sudakov. We also note that $T_{-\beta,0}$ vanishes in the $\beta\to\infty$ limit. In practice, we have used a one-loop running coupling with $\alpha_s(M_Z) = 0.1383$ (matching the value used in Pythia).

With this at hand, we can write all the radiators introduced in section 4 as follows:

$$R^{(\text{primary})}_{\text{Y}_\text{m}\text{-splitter}} = T_{02}(\rho_2, b_i; C_R), \tag{C.3}$$

$$R^{(\text{secondary})}_{\text{Y}_\text{m}\text{-splitter}} = T_{02}(\rho_2/\theta_1, \rho_1/\theta_1 b_g; C_A), \tag{C.4}$$

$$R^{(\text{primary})}_{\text{SD+Y}_\text{m}\text{-splitter}} = T_{02}(\rho_2, b_i; C_R) - T_{-\beta,2}(\rho_2, \zeta_{\text{cut}}; C_R) + T_{02}(\rho_2/\theta_1, \zeta_{\text{cut}}\theta_1^{1+\beta}; C_R), \tag{C.5}$$

$$R^{(\text{red})}_{\texttt{TopSplitter}} = T_{02}(\rho_2, b_i; C_R) - T_{02}(\rho_2, \zeta_{\text{cut}}; C_R), \tag{C.6}$$

[13] Compared to [52], we neglected the β_1 and K contributions, and we will introduce the "B" terms — corresponding to hard collinear splittings — via a shift of the $k_{\max}$ argument.

JHEP09(2018)170

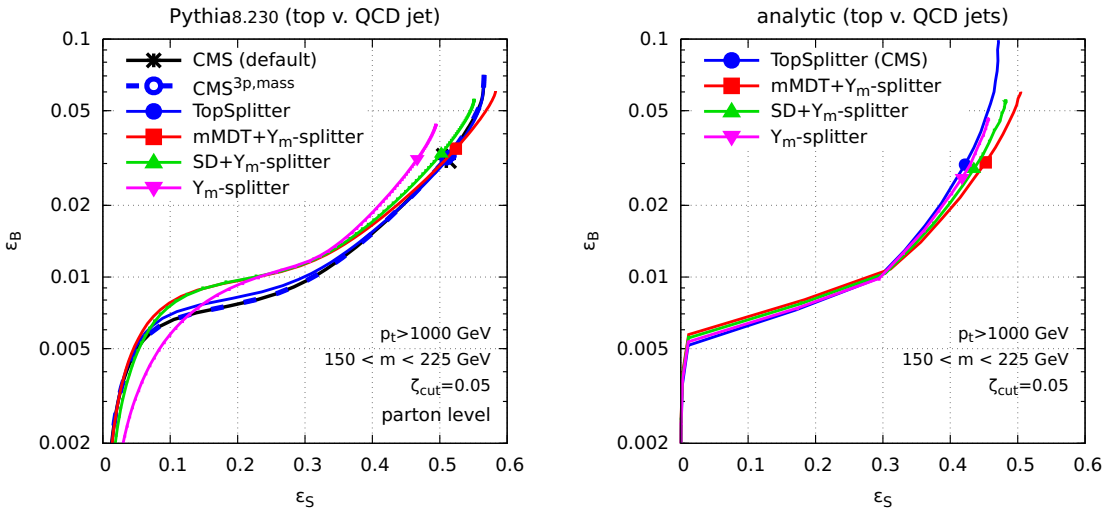


Figure 15. ROC curves obtained for 1 TeV top v. QCD jets when varying $m_{\min}$. For the analytic calculation we have assumed a quark fraction of 2/3, roughly corresponding to the matrix elements used in the Pythia simulation. The rest is as in figure 10.

$$R^{(\text{blue})}_{\texttt{TopSplitter}} = T_{02}(\zeta_{\text{cut}}\theta_1, \rho_2/\theta_1; C_R) - T_{02}(\zeta_{\text{cut}}, \rho_2; C_R), \tag{C.7}$$

$$R^{(\text{secondary})}_{\texttt{TopSplitter}} = T_{02}(\theta_1\rho_2/\rho_1, \rho_1/\theta_1 b_g; C_A) - T_{02}(\theta_1\rho_2/\rho_1, \zeta_{\text{cut}}\theta_1; C_A), \tag{C.8}$$

$$R_{\text{mMDT}}(\rho) = T_{02}(\rho, b_i; C_R) - T_{0,2}(\rho, \zeta_{\text{cut}}; C_R), \tag{C.9}$$

$$R_{\text{SD}}(\rho) = T_{02}(\rho, b_i; C_R) - T_{-\beta,2}(\rho, \zeta_{\text{cut}}; C_R), \tag{C.10}$$

with $b_i = \exp(B_i)$ corresponding to the hard-collinear splittings.

D Performance at lower energy

Throughout this paper, for the purpose of verifying our analytical calculations we have focused on ultra boosted jets with $p_T \sim 2\,\text{TeV}$. It is therefore natural to check whether our main conclusions remain valid at lower jet p_t.

We show our findings for 1 TeV jets in figure 15. A first observation is that due to the somewhat reduced importance of Sudakov effects, the differences between taggers are less visible than for the 2 TeV case both in the analytics and in the parton shower results, which both show a smaller spread between results with different tagging methods.

As before, the ordering between the performance of the Y_m-splitter taggers is well reproduced. Differences due to the (pre-)grooming procedure are also reduced compared to what was seen in figure 10 for 2 TeV jets, which is expected as the phase-space removed by the grooming procedure is reduced. The differences between the CMS-related and Y_m-splitter taggers are not very well reproduced. In the region relevant for phenomenology this is driven by the efficiency for signal (top) jets, which is controlled less well in the analytical calculations than for the QCD background case. Except at small signal efficiencies where

JHEP09(2018)170

the TopSplitter performs marginally worse than the CMS and CMS^{3p,mass} taggers, all three taggers perform equivalently at larger signal efficiency i.e. in the phenomenologically relevant region.

Finally, if we go down to yet smaller p_t, e.g. 500 GeV, the differences between the taggers are even further suppressed, but our main conclusion that TopSplitter is a good overall default choice, remains unchanged.

References

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

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

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

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

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

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

[7] A. Abdesselam et al., *Boosted objects: a probe of beyond the standard model physics*, *Eur. Phys. J.* **C 71** (2011) 1661 [arXiv:1012.5412] [INSPIRE].

[8] A. Altheimer et al., *Jet substructure at the Tevatron and LHC: new results, new tools, new benchmarks*, *J. Phys.* **G 39** (2012) 063001 [arXiv:1201.0008] [INSPIRE].

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

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

[11] A.J. Larkoski, I. Moult and B. Nachman, *Jet substructure at the Large Hadron Collider: a review of recent advances in theory and machine learning*, arXiv:1709.04464 [INSPIRE].

[12] CMS collaboration, *Search for vector-like T and B quark pairs in final states with leptons at* $\sqrt{s} = 13$ *TeV*, *JHEP* **08** (2018) 177 [arXiv:1805.04758] [INSPIRE].

[13] CMS collaboration, *Search for a heavy resonance decaying into a Z boson and a Z or W boson in* $2\ell 2q$ *final states at* $\sqrt{s} = 13$ *TeV*, *JHEP* **09** (2018) 101 [arXiv:1803.10093] [INSPIRE].

JHEP09(2018)170

[14] CMS collaboration, *Search for a heavy resonance decaying into a vector boson and a Higgs boson in semileptonic final states at* $\sqrt{s} = 13$ *TeV*, CMS-PAS-B2G-17-004 (2017).

[15] ATLAS collaboration, *Search for* $W' \to tb$ *decays in the hadronic final state using pp collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, *Phys. Lett.* **B 781** (2018) 327 [arXiv:1801.07893] [INSPIRE].

[16] ATLAS collaboration, *Search for light resonances decaying to boosted quark pairs and produced in association with a photon or a jet in proton-proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, arXiv:1801.08769 [INSPIRE].

[17] ATLAS collaboration, *Search for heavy particles decaying into top-quark pairs using lepton-plus-jets events in proton–proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, *Eur. Phys. J.* **C 78** (2018) 565 [arXiv:1804.10823] [INSPIRE].

[18] CMS collaboration, *Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair*, *Phys. Rev. Lett.* **120** (2018) 071802 [arXiv:1709.05543] [INSPIRE].

[19] A.H. Hoang, S. Mantry, A. Pathak and I.W. Stewart, *Extracting a short distance top mass with light grooming*, arXiv:1708.02586 [INSPIRE].

[20] S. Marzani, L. Schunk and G. Soyez, *A study of jet mass distributions with grooming*, *JHEP* **07** (2017) 132 [arXiv:1704.02210] [INSPIRE].

[21] S. Marzani, L. Schunk and G. Soyez, *The jet mass distribution after soft drop*, *Eur. Phys. J.* **C 78** (2018) 96 [arXiv:1712.05105] [INSPIRE].

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

[23] CMS collaboration, *Measurement of the differential jet production cross section with respect to jet mass and transverse momentum in dijet events from pp collisions at* $\sqrt{s} = 13$ *TeV*, CMS-PAS-SMP-16-010 (2016).

[24] ATLAS collaboration, *Measurement of the soft-drop jet mass in pp collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, *Phys. Rev. Lett.* **121** (2018) 092001 [arXiv:1711.08341] [INSPIRE].

[25] Y. Mehtar-Tani and K. Tywoniuk, *Groomed jets in heavy-ion collisions: sensitivity to medium-induced bremsstrahlung*, *JHEP* **04** (2017) 125 [arXiv:1610.08930] [INSPIRE].

[26] M. Connors, C. Nattrass, R. Reed and S. Salur, *Jet measurements in heavy ion physics*, *Rev. Mod. Phys.* **90** (2018) 025005.

[27] D.E. Kaplan, K. Rehermann, M.D. Schwartz and B. Tweedie, *Top tagging: a method for identifying boosted hadronically decaying top quarks*, *Phys. Rev. Lett.* **101** (2008) 142001 [arXiv:0806.0848] [INSPIRE].

[28] T. Plehn, G.P. Salam and M. Spannowsky, *Fat jets for a light Higgs*, *Phys. Rev. Lett.* **104** (2010) 111801 [arXiv:0910.5472] [INSPIRE].

[29] CMS collaboration, *A Cambridge-Aachen (C-A) based jet algorithm for boosted top-jet tagging*, CMS-PAS-JME-09-001 (2009).

[30] CMS collaboration, *Boosted top jet tagging at CMS*, CMS-PAS-JME-13-007 (2013).

[31] https://github.com/cms-sw/cmssw/blob/master/RecoJets/JetAlgorithms/interface/CMSTopTagger.h

JHEP09(2018)170

[32] G. Brooijmans, *High p_T hadronic top quark identification*, ATL-PHYS-CONF-2008-008 (2008).

[33] J. Thaler and L.-T. Wang, *Strategies to identify boosted tops*, *JHEP* **07** (2008) 092 [arXiv:0806.0023] [INSPIRE].

[34] G. Kasieczka et al., *Resonance searches with an updated top tagger*, *JHEP* **06** (2015) 203 [arXiv:1503.05921] [INSPIRE].

[35] L.G. Almeida et al., *Template overlap method for massive jets*, *Phys. Rev.* **D 82** (2010) 054034 [arXiv:1006.2035] [INSPIRE].

[36] D.E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, *Phys. Rev.* **D 87** (2013) 054012 [arXiv:1211.3140] [INSPIRE].

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

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

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

[40] A.J. Larkoski, I. Moult and D. Neill, *Building a better boosted top tagger*, *Phys. Rev.* **D 91** (2015) 034035 [arXiv:1411.0665] [INSPIRE].

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

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

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

[44] J. Barnard, E.N. Dawe, M.J. Dolan and N. Rajcic, *Parton shower uncertainties in jet substructure analyses with deep neural networks*, *Phys. Rev.* **D 95** (2017) 014018 [arXiv:1609.00607] [INSPIRE].

[45] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy flow polynomials: a complete linear basis for jet substructure*, *JHEP* **04** (2018) 013 [arXiv:1712.07124] [INSPIRE].

[46] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning top taggers or the end of QCD?*, *JHEP* **05** (2017) 006 [arXiv:1701.08784] [INSPIRE].

[47] L. Asquith et al., *Jet substructure at the large hadron collider: experimental review*, arXiv:1803.06991 [INSPIRE].

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

[49] M. Dasgupta, A. Fregoso, S. Marzani and A. Powling, *Jet substructure with analytical methods*, *Eur. Phys. J.* **C 73** (2013) 2623 [arXiv:1307.0013] [INSPIRE].

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

[51] A.J. Larkoski, I. Moult and D. Neill, *Analytic boosted boson discrimination*, *JHEP* **05** (2016) 117 [arXiv:1507.03018] [INSPIRE].

JHEP09(2018)170

[52] M. Dasgupta, L. Schunk and G. Soyez, *Jet shapes for boosted jet two-prong decays from first-principles*, *JHEP* **04** (2016) 166 [arXiv:1512.00516] [INSPIRE].

[53] M. Dasgupta, A. Powling, L. Schunk and G. Soyez, *Improved jet substructure methods: Y-splitter and variants with grooming*, *JHEP* **12** (2016) 079 [arXiv:1609.07149] [INSPIRE].

[54] A.J. Larkoski, I. Moult and D. Neill, *Factorization and resummation for groomed multi-prong jet shapes*, *JHEP* **02** (2018) 144 [arXiv:1710.00014] [INSPIRE].

[55] M. Dasgupta et al., *Logarithmic accuracy of parton showers: a fixed-order study*, *JHEP* **09** (2018) 033 [arXiv:1805.09327] [INSPIRE].

[56] J.M. Campbell and E.W.N. Glover, *Double unresolved approximations to multiparton scattering amplitudes*, *Nucl. Phys.* **B 527** (1998) 264 [hep-ph/9710255] [INSPIRE].

[57] S. Catani and M. Grazzini, *Collinear factorization and splitting functions for next-to-next-to-leading order QCD calculations*, *Phys. Lett.* **B 446** (1999) 143 [hep-ph/9810389] [INSPIRE].

[58] S. Catani and M. Grazzini, *Infrared factorization of tree level QCD amplitudes at the next-to-next-to-leading order and beyond*, *Nucl. Phys.* **B 570** (2000) 287 [hep-ph/9908523] [INSPIRE].

[59] M. Field et al., *Three-prong distribution of massive narrow QCD jets*, *Phys. Rev.* **D 87** (2013) 094013 [arXiv:1212.2106] [INSPIRE].

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

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

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

[63] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross-sections in deep inelastic scattering*, in the proceedings of *Monte Carlo generators for HERA physics*, April 27–30, Hamburg, Germany (1998), hep-ph/9907280 [INSPIRE].

[64] J.R. Andersen et al., *Les Houches 2017: physics at TeV colliders standard model working group report*, in the proceedings of the 10$^{\text{th}}$ *Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2017)*, June 5–23, Les Houches, France (2018), arXiv:1803.07977 [FERMILAB-CONF-18-122].

[65] ATLAS collaboration, *Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, ATLAS-CONF-2015-036 (2015).

[66] M. Dasgupta, A. Powling and A. Siodmok, *On jet substructure methods for signal jets*, *JHEP* **08** (2015) 079 [arXiv:1503.01088] [INSPIRE].

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

[68] S. Catani et al., *New clustering algorithm for multi-jet cross-sections in e^+e^- annihilation*, *Phys. Lett.* **B 269** (1991) 432 [INSPIRE].

[69] S. Catani, Y.L. Dokshitzer, M.H. Seymour and B.R. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, *Nucl. Phys.* **B 406** (1993) 187 [INSPIRE].

JHEP09(2018)170

[70] S.D. Ellis and D.E. Soper, *Successive combination jet algorithm for hadron collisions*, *Phys. Rev.* **D 48** (1993) 3160 [hep-ph/9305266] [INSPIRE].

[71] A. Gehrmann-De Ridder and E.W.N. Glover, *A complete $O(\alpha\alpha_s)$ calculation of the photon + 1 jet rate in e^+e^- annihilation*, *Nucl. Phys.* **B 517** (1998) 269 [hep-ph/9707224] [INSPIRE].

[72] D. Bertolini, J. Thaler and J.R. Walsh, *The first calculation of fractional jets*, *JHEP* **05** (2015) 008 [arXiv:1501.01965] [INSPIRE].

[73] S. Höche and S. Prestel, *Triple collinear emissions in parton showers*, *Phys. Rev.* **D 96** (2017) 074017 [arXiv:1705.00742] [INSPIRE].

[74] M. Cacciari and G.P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, *Phys. Lett.* **B 641** (2006) 57 [hep-ph/0512210] [INSPIRE].

[75] Y.L. Dokshitzer, V.A. Khoze and S.I. Troian, *On specific QCD properties of heavy quark fragmentation ('dead cone')*, *J. Phys.* **G 17** (1991) 1602 [INSPIRE].

[76] Y.L. Dokshitzer, V.A. Khoze, L.H. Orr and W.J. Stirling, *Properties of soft radiation near $t\bar{t}$ and W^-W^- threshold*, *Nucl. Phys.* **B 403** (1993) 65 [hep-ph/9302250] [INSPIRE].

[77] F.A. Dreyer, L. Necib, G. Soyez and J. Thaler, *Recursive soft drop*, *JHEP* **06** (2018) 093 [arXiv:1804.03657] [INSPIRE].

[78] S. Catani and M.H. Seymour, *The dipole formalism for the calculation of QCD jet cross-sections at next-to-leading order*, *Phys. Lett.* **B 378** (1996) 287 [hep-ph/9602277] [INSPIRE].

[79] S. Catani and M.H. Seymour, *A General algorithm for calculating jet cross-sections in NLO QCD*, *Nucl. Phys.* **B 485** (1997) 291 [*Erratum ibid.* **B 510** (1998) 503] [hep-ph/9605323] [INSPIRE].

[80] A.J. Larkoski and J. Thaler, *Unsafe but calculable: ratios of angularities in perturbative QCD*, *JHEP* **09** (2013) 137 [arXiv:1307.1699] [INSPIRE].

[81] A.J. Larkoski, S. Marzani and J. Thaler, *Sudakov safety in perturbative QCD*, *Phys. Rev.* **D 91** (2015) 111501 [arXiv:1502.01719] [INSPIRE].

JHEP09(2018)170