Published for SISSA by Springer

Received: April 2, 2015
Accepted: June 8, 2015
Published: June 30, 2015

Resonance searches with an updated top tagger

Gregor Kasieczka,[a] **Tilman Plehn,**[b] **Torben Schell,**[b] **Thomas Strebler**[a] **and Gavin P. Salam**[c]

[a] *Institute for Particle Physics, ETH Zürich, Otto-Stern-Weg 5, CH-8093 Zürich, Switzerland*
[b] *Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, D-69120 Heidelberg, Germany*
[c] *CERN, PH-TH, CH-1211 Geneva 23, Switzerland*

E-mail: gregorka@phys.ethz.ch, plehn@uni-heidelberg.de, schell@thphys.uni-heidelberg.de, stthomas@student.ethz.ch, gavin.salam@cern.ch

Abstract: The performance of top taggers, for example in resonance searches, can be significantly enhanced through an increased set of variables, with a special focus on final-state radiation. We study the production and the decay of a heavy gauge boson in the upcoming LHC run. For constant signal efficiency, the multivariate analysis achieves an increased background rejection by up to a factor 30 compared to our previous tagger. Based on this study and the documentation in the appendix we release a new HEPTopTagger2 for the upcoming LHC run. It now includes an optimal choice of the size of the fat jet, N-subjettiness, and different modes of Qjets.

Keywords: QCD Phenomenology, Jets

ArXiv ePrint: 1503.05921

doi:10.1007/JHEP06(2015)203

JHEP06(2015)203

Contents

JHEP06(2015)203

1 Introduction

After the discovery of the Higgs boson, a keystone of the Standard Model, one main task for the upcoming LHC runs will be searches for physics beyond the Standard Model. Several open experimental and theoretical questions point to additional particles or structures at energies above the electroweak energy scale [1]. A very generic feature of many extensions of the Standard Model is the presence of additional heavy particles which preferentially decay to a pair of top quarks [2–7]. One example for such a resonance could be a heavy neutral Z'-gauge boson with a TeV-scale mass. Historically, such states were only searched for using semi-leptonically decaying top pairs. There, a kinematic reconstruction is based on an approximate reconstruction of the missing neutrino momentum through a W-mass or top mass condition. In the last LHC run this search channel was supplemented by resonance searches based on boosted, hadronically decaying top pairs. In the corresponding ATLAS analysis [8–12] the HEPTopTagger [13–15] and the template tagger [16–19] each showed a similar reach, comparable with the semileptonic channel. This experimental success is based on rapid progress in the field of jet substructure both experimentally and theoretically, which will gain even more momentum during the 13 TeV LHC run.

The field of top and Higgs tagging [20–23] started essentially as a Gedankenexperiment to illustrate recombination jet algorithms [24, 25]. After some early attempts for example to tag hadronically decaying tops [26–29] it took off with the development of the BDRS Higgs tagger with its mass drop condition [30] and a filtering step targeting underlying event and pile-up [31]. The first top taggers were simple, deterministic algorithms which could identify and reconstruct hadronically decaying top quarks including subjet b-tagging [13, 32–34]. They were based on deliberately simple structures and algorithms, to firmly establish subjet methods in ATLAS and CMS. After the experimental success of these completely new analysis tools in the first run of the LHC, the upcoming run will benefit from more advanced

top tagging methods. Those include multivariate taggers [35], template taggers [16–19], as well as shower deconstruction [36] or event deconstruction [37].[1] For those specialized tools the challenge will be to still provide a universal top tagging approach, which on the one hand allows for optimal experimental results, but on the other hand identifies and reconstructs boosted top quarks independent of the specialized analysis framework.

Over time, the original HEPTopTagger [13] has gone through several rounds of improvements. The first modification included a re-formulation of the algorithm, leading to the trademark A-shaped kinematic cuts [14, 15]. One of the key observations leading to these cuts is that in the absence of a b-tag it is not helpful to uniquely identify the two W-decay jets because in typical top decays there will be two jet-jet combinations which reconstruct to an invariant mass around 80 GeV [40]. The first set of new, additional variables [41] then included a combination of the usual filtered top mass [31] with a pruned top mass [42–44]. In this upgrade we introduced a fat jet radius up to $R = 1.8$ for moderately boosted tops and allow for a choice of Cambridge-Aachen [45, 46] and k_T [47–51] jet algorithms in all internal clustering and filtering steps except for the mass drop condition. This improves the tagging performance for highly boosted tops [41]. Recently, the algorithm was slightly changed to avoid background shaping [35]. In the same study we added a low-p_T mode based on Fox-Wolfram moments [52–54] to incorporate angular correlations, extending the tagging coverage to transverse momenta as low as $p_{T,t} = 150\,\text{GeV}$.

In this paper we present a detailed study of the HEPTopTagger2, collecting all previous modifications, as well as a whole range of new features targeted at multivariate analyses and statistical approaches to single events [55–57]. The main body of the paper will focus on Z' searches, where final-state jet radiation turns out to be the limiting factor of the original tagger. After resolving the issue with final-state radiation we will step by step improve the tagging algorithm by defining and including additional kinematic information. Finally, we will compare the multivariate tagging performance with the leading projections based on event deconstruction [37].

The main background in fully hadronic $Z' \to t\bar{t}$ searches is QCD multi-jets production, which allows us to directly translate all our findings into a performance study based on tagging $t\bar{t}$ pairs in the Standard Model. We will show these results together with a review of the complete HEPTopTagger2 algorithm and the code interface in the appendix.

2 Resonance reconstruction

The key challenge of any top tagger is its broad range of applications and the related optimization of the algorithms and codes. For example, the HEPTopTagger was developed to solve the combinatorial problems in $t\bar{t}H$ searches [13]. The first public tagging code was presented for supersymmetric top partner searches in semi-leptonic top decays [14, 15]. Its proposed applications include single top production to experimentally separate the s-channel and t-channel production processes [58]. However, its experimental application during the first LHC run was the search for heavy resonances decaying to hadronic $t\bar{t}$

[1] Why a kinematic selection as naive as 'top buckets' [38, 39] also seems to work is beyond the comprehension of the authors.

JHEP06(2015)203

searches [8–12]. For such a resonance search the kinematic top tagger in combination with a b-tag showed a similar performance as the usual, approximate reconstruction of semileptonic $t\bar{t}$ pairs. In this paper we will present a set of improvements towards the HEPTopTagger2 for a Z' search at the 13 TeV LHC. Many of these improvements can be applied to other LHC processes, as will be discussed in the appendix.

In using all available information from a pair of boosted top quarks, event deconstruction is currently giving the leading performance estimates for heavy resonance searches [37]. For the analysis in the main body of this paper we will follow the analysis framework of ref. [37], to eventually allow for a comparison in section 4. For the signal we therefore use Pythia8 [59] to generate $Z' \to t\bar{t}$ events with $m_{Z'} = 1500\,\text{GeV}$ at 13 TeV collider energy. Assuming the same couplings as for the Standard Model Z-boson would yield a width of $\Gamma(Z') = 47\,\text{GeV}$; to be consistent with the assumed experimental resolution in ref. [37] we increase the width to 65 GeV and only simulate the vector couplings. However, we will see that this choice of the physical Z' width does not affect our results which are based on the reconstructed fat jet kinematics. For the Z' decay we assume a 100% branching ratio to top pairs. The two backgrounds are continuum $t\bar{t}$ production which we simulate assuming $p_{T,t} > 400\,\text{GeV}$, and QCD di-jet production, also requiring $p_{T,j} > 400\,\text{GeV}$. Again, we rely on Pythia8, keeping in mind that for the pure QCD background our di-jet rate might not be a conservative estimate. All top quarks are forced to decay hadronically. Our simulations for the main body of the paper include underlying event but do not account for pile-up or detector effects, unless explicitly mentioned. For a completely realistic study of the signal and background efficiencies of the new HEPTopTagger2 we will have to rely on upcoming experimental studies. For our multivariate tagging analyses we optimize the background rejection with respect to the pure QCD background, because it is by far dominant.

Decay kinematics. On the analysis level we first select events with at least two fat jets with

$$p_{T,\text{fat}} > 400\,\text{GeV} \qquad \text{and} \qquad |y_{\text{fat}}| < 2.5\,, \tag{2.1}$$

reconstructed using the C/A algorithm [45, 46] with cone size $R = 1.5$, as implemented in FastJet [50, 51]. We limit ourselves to the two hardest fat jets in each event for the Z' search. The corresponding cut flow is given in table 1. Using the old default HEPTopTagger setup [14, 15] we find a double top tagging efficiency of $\varepsilon_{2\text{tags}} = 14\%$ in the signal, as shown in table 1. If we apply a fixed invariant mass window $m_{tt} \in [1200, 1600]\,\text{GeV}$ on the tagged and reconstructed top quarks, the Z' tagging efficiency is $\varepsilon_{Z'} = 10.2\%$. For the $t\bar{t}$ background we find mis-tagging probabilities of $\varepsilon_{2\text{tags}} = 13.7\%$ and $\varepsilon_{Z'} = 3.3\%$. For the QCD background sample the double mistag rates are $\varepsilon_{2\text{tags}} = 6.6 \cdot 10^{-4}$ and $\varepsilon_{Z'} = 1.5 \cdot 10^{-4}$. The QCD jets background exceeds the continuum top pair production by a factor five after all cuts.

A straightforward improvement of the basic analysis shown in table 1 should be to replace the mass window by a boosted decision tree (BDT) analysis, as implemented in Tmva [60, 61], based on the reconstructed invariant mass m_{tt}. Although a multivariate approach for a single variable may not be appropriate, the BDT including solely m_{tt} es-

JHEP06(2015)203

	$Z' \to t\bar{t}$	$t\bar{t}$	QCD
generator level	10^5	10^5 (1.76 pb)	$8 \cdot 10^6$ (1.93 nb)
≥ 2 fat jets eq. (2.1)	69142	85284 (1.50 pb)	$6.7 \cdot 10^6$ (1.62 nb)
hardest 2 fat jets HTT[JHEP1010] tagged	9679	11706 (0.21 pb)	4426 (1.07 pb)
$m_{tt} \in [1200, 1600]$ GeV	7031	2817 (0.05 pb)	978 (0.24 pb)

Table 1. Number of events and the corresponding PYTHIA8 cross section used for our analysis. The efficiencies $\varepsilon_{S,B}$ for a Z' extraction are defined as the ratio of the last to the second line in this table.

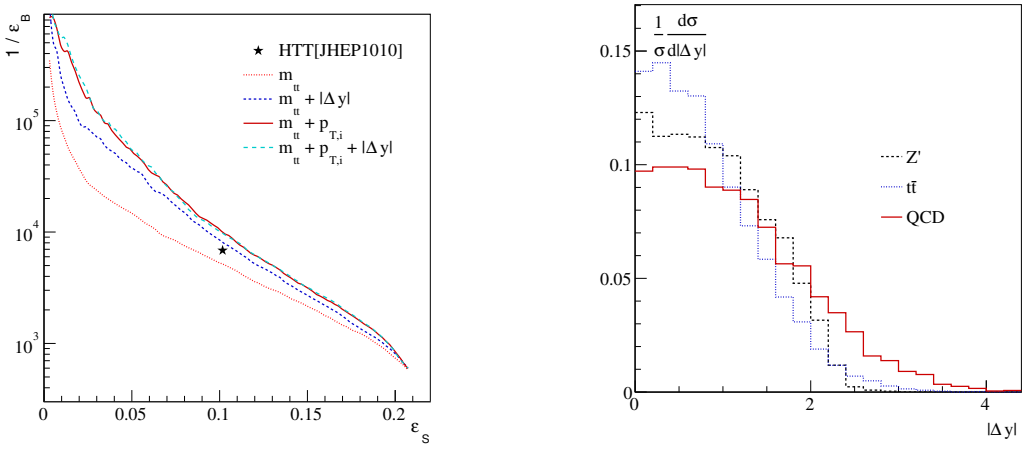

Figure 1. Left: ROC curves for the dominant QCD background vs. the Z' signal after including additional kinematic information shown in eq. (2.2). As in all figures the asterisk corresponds to the original HEPTOPTAGGER described in refs. [14, 15]. Right: $|\Delta y|$ distribution of the reconstructed top quarks for signal and backgrounds.

sentially performs like a flexible mass window and is only meant as a starting point before including further variables. In the left panel of figure 1 we first show the results as receiver operator characteristic (ROC) curve, correlating the best signal and background efficiencies based on a given set of kinematic observables. This approach has been used to improve and benchmark the general performance of the HEPTOPTAGGER [35]. Because the QCD jet background is dominant we always set up our multivariate analyses based on the Z' signal and the QCD background sample. Compared to the working point of the original public HEPTOPTAGGER tool [14, 15] with a fixed mass window $m_{tt} \in [1200, 1600]$ GeV the new HEPTOPTAGGER2 including m_{tt} in a multivariate analysis looks slightly worse. The reason is the change in the order in the algorithm described in the appendix compared to the orignal version (changed order of step 4 and step 5). It significantly reduces the background sculpting, but at the expense of background rejection for example for a constant signal efficiency [35].

For a better discrimination between signal and background we should include additional variables in our multivariate analysis. The deterministic structure of the HEPTOPTAGGER will still allow for a particularly clear separation of the actual tagging and

JHEP06(2015)203

reconstruction from a subsequent kinematic analysis based on the reconstructed top momenta. The first additional variable we include is the rapidity difference between the two reconstructed top quarks, $|\Delta y|$. The corresponding signal and background distributions are shown in the right panel of figure 1. While this variable might not be too efficient in removing the $t\bar{t}$ continuum background, events are visibly less central for QCD jets. The differences can hardly be translated into efficient kinematic cuts, but they will help as part of a multivariate analysis. In the left panel of figure 1 we show the corresponding improvement in terms of ROC curves. In particular for low signal efficiencies $\varepsilon_S < 0.1$ we find a significant reduction of the background fake rates, going beyond the working point of the first HEPTopTagger.

An obvious extension of our set of kinematic observables are the transverse momenta of the reconstructed top quarks. Note that as part of the ROC analysis we do not have to ensure that the different kinematic variables are independent of each other, which would be problematic for a combination of m_{tt} and the $p_{T,t}$ distributions. Again, the improvement from the transverse momentum spectra is shown in the left panel of figure 1. All this illustrates that the kinematic information on the tagged and reconstructed tops can increase the background rejection by 50% to 100% for fixed signal tagging efficiency. We also see that once we include the top-pair invariant mass and the transverse momenta, the additional improvement from $|\Delta y|$ vanishes, because the 2-particle final state is essentially fully described. As kinematic observables in our multivariate analysis we choose

$$\{\, m_{tt}, p_{T,t_1}, p_{T,t_2} \,\} \qquad \text{(decay kinematics)}. \tag{2.2}$$

QCD jets. In purely hadronic searches for new physics, QCD effects beyond fixed order are a major issue in trying to theoretically understand the signal and backgrounds. Before we devise strategies to deal with final-state radiation and initial-state radiation in heavy-resonance searches we can estimate their effect on the naive tagger-based analysis.

To study the impact of initial-state radiation and final-state radiation on the tagging performance, we separately remove initial-state radiation and final-state radiation from all signal events. For the QCD jets background this is not sensible, because we need both mechanisms to generate a sufficient jet multiplicity mimicing hadronic top decays. The ROC curves in figure 2 show the expected improvements in the absence of additional signal jets. We see that the leading effect spoiling the signal extraction is final-state radiation (FSR). Initial-state radiation (ISR) affects top tagging in two ways. First, the additional QCD jets can mimic for example the softer W-decay jet and degrade the tagging efficiency through combinatorics. Second, ISR jets recoil against the Z', affecting the p_T spectrum of the top quarks. In particular the tagging of the softer top decay can benefit from this recoil, which means that for large signal efficiency the results without ISR become worse than those with all jet activity included.

As a whole, the results shown in figure 2 indicate potentially significant improvements of top taggers when we target the different effects of QCD jet radiation. We will show in the following subsection how a deterministic top tagger is limited by final-state radiation and how the new HEPTopTagger2 can avoid these issues. Combinatorial problems related to initial-state radiation will then be one of the key topics in section 3.

JHEP06(2015)203

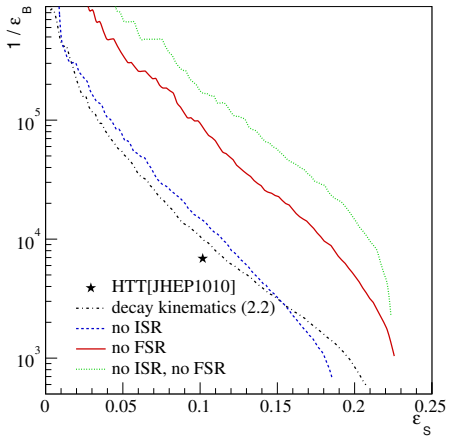

Figure 2. ROC curves for different combinations of initial-state jet radiation (ISR) and final-state jet radiation (FSR) in the Z' signal generation. The background is QCD with ISR and FSR for all curves.

Final-state radiation. Final-state radiation (FSR) turns one of the key advantages of our top tagger into a significant problem: unlike some other top tagging approaches, the HEPTopTagger returns the 4-momentum of the tagged top, including a cut on the reconstructed top mass $m_{\text{rec}} \in [150, 200]\,\text{GeV}$ [35]. This allows us to trivially reconstruct $m_{Z'}$. Final-state radiation off the top decay products will be captured by the jet clustering and contribute to the correct filtered top mass value [31]. This way it will not pose a problem as long as the Z' decays to on-shell tops.

However, if the Z' decays to slightly off-shell tops, which turn themselves into on-shell tops, this final-state radiation off the intermediate top mis-aligns the actual Z' with the Z' as reconstructed from the top quarks at the moment they decay. Because the hard radiated gluon does not enter the top reconstruction, the top tag will pass, but lead to an underestimated $m_{Z'}$ value. In the left panel of figure 3 we indeed see that the m_{tt} distribution for the top-tagged signal correctly peaks around $m_{Z'}$, but develops a sizeable asymmetric tail towards smaller m_{tt} values. While the details of this asymmetric tail from Pythia8 should be subject to a detailed Monte Carlo study, we simply confirm that turning off final-state radiation by hand gets rid of it almost entirely. The remaining slight broadening as well as a minimal tail towards smaller m_{tt} values is due to small losses in the top 4-momentum reconstruction of the tagger. At higher values of $m_{Z'}$ the asymmetric tail is further enhanced.

The problem with large asymmetric tails from final-state radiation is that they cannot simply be corrected for in a universal top tagger. The basic structure of the HEPTopTagger has to identify and reconstruct top quarks, rather than the decay products of a heavy Z' resonance. Therefore, we do not modify the actual tagger, but we account for final-state radiation through an additional set of kinematic observables.

Following the brief discussion above, including the kinematics of the fat jet in addition to the reconstructed top 4-momentum should remove the broad asymmetric tail in the reconstructed $m_{Z'}$ values. Again, we first select events with two tagged tops, including the

JHEP06(2015)203

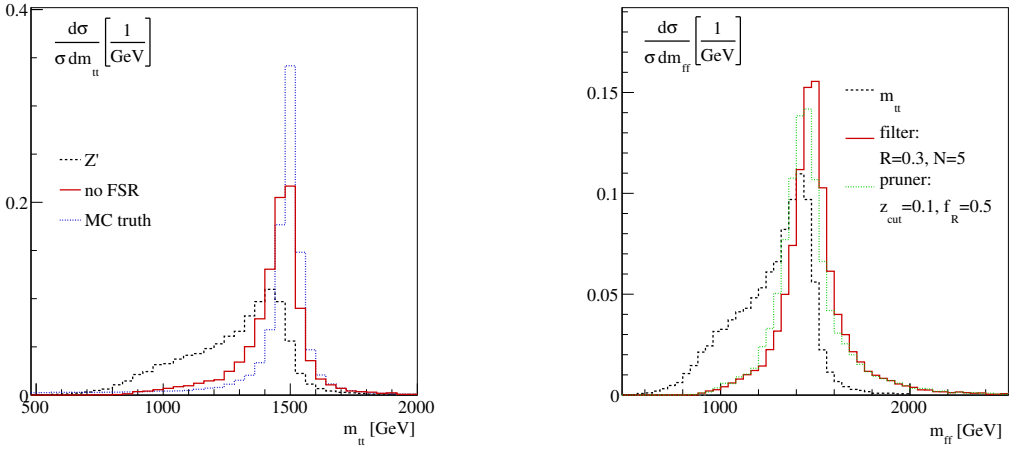

Figure 3. Effect of final-state radiation on the invariant mass of the tagged and reconstructed $t\bar{t}$ system m_{tt} for the Z' signal (left) and different approaches to reconstruct the Z' mass peak (right). Monte Carlo truth is $\sqrt{p^2_{Z'}}$ with an assumed width of 65 GeV.

top mass condition. Instead of using the 4-momenta of the tagged tops, we now reconstruct the Z' mass from the 4-momenta of the two fat jets of size $R = 1.5$, which eventually lead to the top tags. In the presence of underlying event and initial-state radiation the naive m_{ff} distribution peaks roughly at the correct Z' mass and shows symmetric tails. To use the invariant mass of the two fat jets we need to apply filtering [31]. In the right panel of figure 3 we compare the filtered invariant mass from the two fat jets [31] and its pruned value [42–44], both as implemented in FastJet [50, 51]. As a reference we also show the m_{tt} distribution from the left panel of the same figure. Unlike the reconstructed m_{tt} distribution, both, the filtered and the pruned m_{ff} distributions give symmetric peaks around the correct $m_{Z'}$ value.

To be able to use the filtered m_{ff} values in our HEPTopTagger analysis we confirm that filtering and pruning give stable numerical results for the invariant mass of the two fat jets. Results for different parameter settings are listed in table 2. We give the peak positions, which would be subject to a proper calibration, the fitted Breit-Wigner widths for the symmetric peaks, and the tagging performances for a fixed mass window $|m_{\text{ff}} - m_{Z'}| < 150\,\text{GeV}$. Replacing the Breit-Wigner width with a Gaussian would make no difference, but give a poorer modelling of the tails. Typical widths of the reconstructed Z' mass peak will range around 145 GeV, roughly twice the assumed particle width of 65 GeV. Even in the absence of detector effects, this resolution will replace the assumed particle width of 65 GeV in all of the following analysis. The constant numbers in table 2 confirm that the m_{ff} criterion is stable for different filtering parameters as well as pruning.

On the other hand, the results shown in table 2 also indicate that simply replacing the m_{tt} window by a filtered m_{ff} value will not improve the Z' extraction. In figure 4 we show that the steeply falling QCD jets background now has a maximum around $m_{\text{ff}} = 1.3\,\text{TeV}$, while for the reconstructed top quarks there exists a much more pronounced maximum around $m_{tt} = 900\,\text{GeV}$. The reason is that top tagging removes events with many hard

JHEP06(2015)203

	m_{peak} [GeV]	Γ [GeV]	$\varepsilon_{Z'}^{\pm 150}$	$1/\varepsilon_{tt}^{\pm 150}$	$1/\varepsilon_{\text{QCD}}^{\pm 150}$
$m_{tt} \in [1200, 1600]$ GeV	–	–	0.136	22	2805
unfiltered	1539	167	0.141	21	1960
$R = 0.3$, $N = 4$	1457	152	0.146	28	2218
$R = 0.3$, $N = 5$	1477	144	0.150	25	2098
$R = 0.3$, $N = 6$	1489	139	0.151	25	2052
$R = 0.3$, $N = 7$	1496	144	0.151	24	2043
$R = 0.2$, $N = 5$	1443	140	0.141	29	2329
$R = 0.3$, $N = 5$	1477	144	0.150	25	2098
$R = 0.4$, $N = 5$	1500	144	0.151	24	2030
$R = 0.5$, $N = 5$	1515	143	0.148	23	1993
pruning $z = 0.1$, $f_R = 0.5$	1443	150	0.138	26	2075

Table 2. Breit-Wigner fits and performance of different grooming approaches. The quoted efficiencies are based on a window for the invariant mass of the two filtered fat jets $|m_{\text{ff}} - m_{Z'}| < 150\,\text{GeV}$.

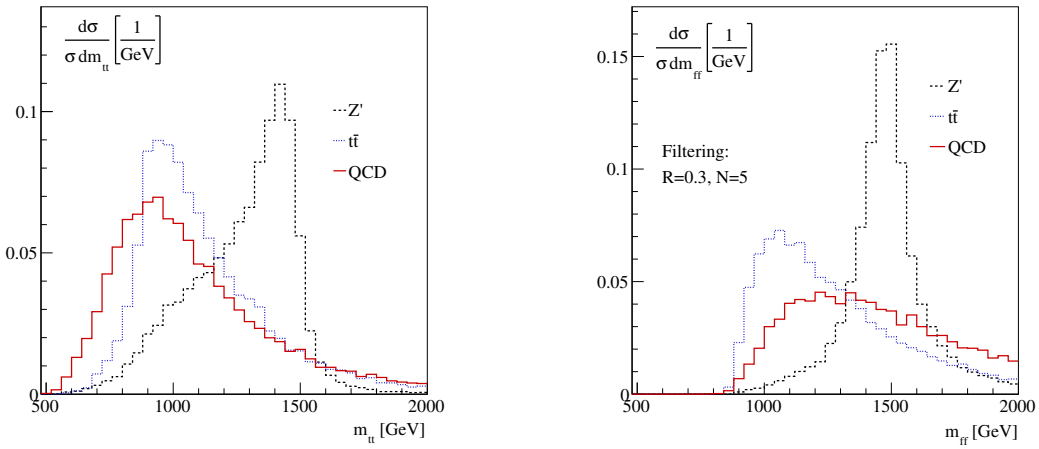

Figure 4. Reconstructed mass distribution of the Z' signal and the backgrounds based on the tagged tops (left) and the corresponding filtered fat jets (right).

QCD jets in two steps: first requiring the correct top mass value from three assumed top decay products, and second when applying the Z' mass window. If we remove the first step, the second one has to deal with larger backgrounds at high m_{ff} values.

If we want to include final-state radiation and at the same time benefit from its additional information, we need to keep m_{ff} as well as m_{tt} in our analysis, and not apply a simple mass window on the m_{tt} distribution. The kinematics of the Z'-decay is then described by

$$\{m_{tt}, m_{\text{ff}}, p_{T,t_1}, p_{T,t_2}, p_{T,\text{f}_1}, p_{T,\text{f}_2}\} \qquad \text{(filtered fat jets)}. \tag{2.3}$$

All default settings of the HEPTopTagger are listed in the appendix. We filter the fat

JHEP06(2015)203

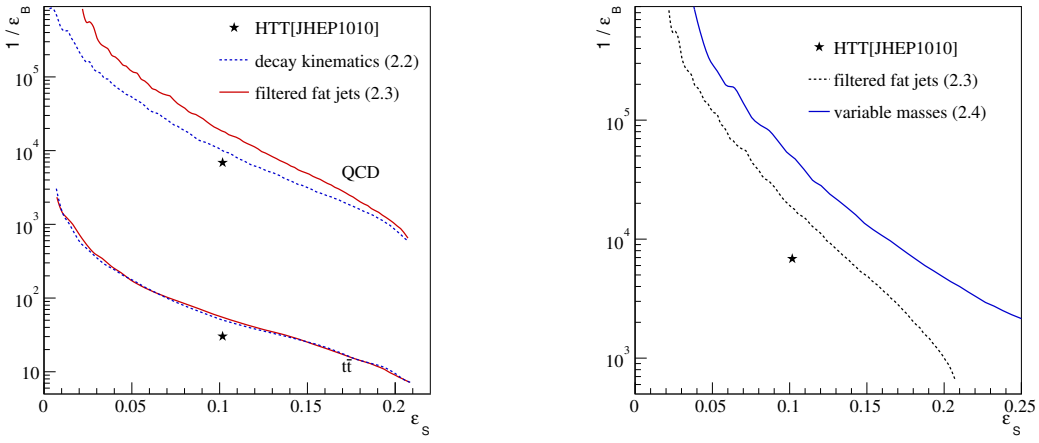

Figure 5. Left: performance of the multivariate analysis including the information on the fat jets, as given in eq. (2.2) and eq. (2.3). Only in this plot we show ROC curves both for $t\bar{t}$ and QCD backgrounds. The underlying BDT is optimized for QCD background rejection. Right: performance curve for the full analysis including eq. (2.4) only accounting for the dominant QCD jet background.

jets using $R = 0.3$ and keep the $N = 5$ hardest substructures. In the left panel of figure 5 we show the corresponding ROC curves. Unlike in the rest of the paper we study both the $t\bar{t}$ and QCD jets background. The BDTs are trained on QCD background events and then applied to the corresponding test samples. While the $t\bar{t}$ background is unchanged, the QCD background rejection benefits from the filtered fat jet information. Compared to the original HEPTopTagger we achieve an improvement of up to a factor 2 in $1/\varepsilon_B$ for constant signal efficiency. We note that for the QCD background the combination of mis-tagged top kinematics and fat jet kinematics goes beyond the description of the hard process. For example initial-state radiation, sensitive to the color structure of the signal and the background, will be captured in this combination of observables. On the other hand, because the fat jets are defined using the standard jet algorithms and show a stable filtering performance, we do not envision major experimental problems provided pile-up subtraction works as well as expected.

The set of kinematic observables listed in eq. (2.3) still relies on the deterministic HEPTopTagger output. This means that the identification of a Z' signal event is limited by the efficiency of two top tags. The choice of a working point in the top tagging algorithm will therefore limit our over-all efficiency. On the other hand, we already know that for hadronic Z' searches the QCD jets background is dominant and will only be reduced through a combination of top tags and Z' mass reconstruction.

In addition, we omit a fixed mass window for the reconstructed top mass m_{rec}. Instead, we widely open the top mass and W-mass constraints in the tagging algorithm. For each of the tops the corresponding m_{rec} value then becomes an output of the tagger. We provide the multivariate Z' analysis with the smaller and larger of these two output m_{rec} values, which we label as $m_{\text{rec}}^{\text{min}}$ and $m_{\text{rec}}^{\text{max}}$ respectively. Similarly, we avoid a fixed window for the ratio of the W-mass to the top mass, parametrized as f_W in the tagging algorithm. Its

JHEP06(2015)203

deviation from the true value is given by the value of f_{rec} defined in the appendix. In the multivariate analysis we include the maximum of the two f_{rec} values corresponding to each tagged top.

$$\{m_{tt}, m_{\text{ff}}, p_{T,t_1}, p_{T,t_2}, p_{T,\text{f}_1}, p_{T,\text{f}_2}, m_{\text{rec}}^{\text{min}}, m_{\text{rec}}^{\text{max}}, f_{\text{rec}}^{\text{max}}\} \qquad \text{(variable masses)}. \tag{2.4}$$

The result is shown in the right panel of figure 5, where the range of accessible efficiencies eventually extends to 56%. Altogether, the analysis based on the set of kinematic variables shown in eq. (2.4) gives us an improvement of up to a factor 5 in background rejection for a constant Z'-signal efficiency.

3 Updated tagger

Fat jets with a geometric size of $R = 1.5$ or even $R = 1.8$ have shown to be powerful new analysis objects at the LHC. The radius of the fat jet is directly related to the energy or boost of the heavy particles which can be captured. This means that a multi-purpose top tagger will be based on as large fat jets as possible. However, to realize their potential such large jets require additional treatment linked to their large geometric size. Without a dedicated analysis step, underlying event and pile-up will almost entirely wash out any structure inside the fat jet. Filtering [31] as an integral part of all versions of the HEPTopTagger [13–15] effectively reduces the geometric size of the fat jet used to reconstruct the top 4-momentum by introducing a second clustering stage with higher resolution. This solves the problem with underlying event and pile-up, but there remains a combinatorial problem caused for example by initial-state radiation. In particular the softer of the two subjets from the W-decay can easily be faked by a typical QCD jet inside the fat jet. This will lead to a wrong reconstruction of the top 4-momentum, which we can only counter by applying harder tagging requirements and hence reducing the tagging efficiency. These so-called type-2 tags [41], where only two of three top decay jets can be identified with a parton-level decay quark have been in the focus of HEPTopTagger studies at moderate boost [35, 41, 68]. In the reconstruction of heavy resonances we can solve the problem of (too) large fat jets by adapting the size of the fat jet to the kinematics of the tagged top. It turns out that this adaptive size of the fat jet also gives us another powerful kinematic variable for the multivariate analysis. Finally, we will show how this optimalR modification of our tagging algorithm can be further improved by including N-subjettiness variables.

OptimalR mode. There have been different attempts to adjust the size of the fat jet for example based on the transverse momentum of the fat jet [32, 57, 62], but none them lead to a dramatic effect in the performance of taggers. We instead choose a purely algorithmic way of determining the minimum size of the fat jet [69]. Assuming that three top decay jets are captured by the fat jet we can run the standard HEPToptagger algorithm to determine the top mass from the three leading subjets [35]. For a large fat jet size, typically $R = 1.5$ or $R = 1.8$, we compute a reference value of m_{rec}, which should be around the top mass. In the usual tagging algorithm, this computation of m_{rec} from filtered subjets

JHEP06(2015)203

takes into account final-state radiation off the on-shell top. We then reduce the size of the fat jet in steps of $\Delta R = 0.1$ and compute the corresponding values of $m_{\text{rec}}(R)$. In case of several possible triplets, this includes the step of choosing the one closest to the physical top mass, as described in step (5) in the appendix. As a function of the decreasing jet size R the fat jet mass $m_{\text{rec}}(R)$ will form a stable plateau, until the reduced fat jet will be too small to capture all three top decay jets. At this point $m_{\text{rec}}(R)$ will leave the plateau and show a significant drop. For $R = 1.5$, which is sufficient for the Z' mass in our study, we define this drop through

$$\frac{m_{\text{rec}}^{(1.5)} - m_{\text{rec}}(R)}{m_{\text{rec}}^{(1.5)}} > 0.2 \quad \Leftrightarrow \quad R < R_{\text{opt}}\,. \tag{3.1}$$

Once the shrinking fat jet passes this condition we go back one step to the last R value on the plateau and define this value as R_{opt}. The smallest value we allow in this study is $R_{\text{opt}} = 0.5$, but for $p_{T,t} \gtrsim 1\,\text{TeV}$ this value can be adjusted in the tagger setup. This value could be a challenge of the calorimeter resolution, so the corresponding results are subject to tests based on a full detector simulation in ATLAS and in CMS. In this paper we typically arrive around $R_{\text{opt}} = 0.6$. The tagging result for this R_{opt} value will be the output of the top tagger and the basis for all variables used in the BDT.

Measuring R_{opt} defines another useful variable for the top tagger, because we can also predict R_{opt} from the fat jet kinematics. A similar reasoning is used in the original HEPTopTagger algorithm, where a consistency condition on the reconstructed top momentum $p_{T,t} > 200\,\text{GeV}$ ensures that the reconstructed top can actually be captured in the fat jet. In the optimalR mode we first determine the transverse momentum of the filtered fat jet, $p_{T,\text{f}}$ as described in the previous section. Including up to ten hardest subjets after a filtering step with $R_{\text{filt}} = 0.2$ turns out to give the best estimate of $p_{T,\text{f}}$ for this purpose. Reducing this number to five subjets has no measurable effect on the width of the reconstructed $p_{T,\text{f}}$ distribution, but slightly shifts its maximum to smaller values [69]. The final number will be subject to an independent optimization in ATLAS and CMS.

For $p_{T,\text{f}} > 200\,\text{GeV}$ we derive a closed form by fitting a function $R_{\text{opt}}^{(\text{calc})} \propto 1/p_{T,\text{f}}$ to simulated data, as described in the appendix. The kinematic variables in our the multivariate tagger now read

$$\left\{ m_{tt}, m_{\text{ff}}, p_{T,t_1}, p_{T,t_2}, p_{T,\text{f}_1}, p_{T,\text{f}_2}, m_{\text{rec}}^{\text{min}}, m_{\text{rec}}^{\text{max}}, f_{\text{rec}}^{\text{max}}, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})} \right\} \qquad \text{(optimalR)}. \tag{3.2}$$

For this case of two top tags we choose $R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}$ as the maximum deviation of the tagged tops. In this form all subsequent kinematic variables linked to the top tags will be evaluated with the fat jet size R_{opt}. For the Z' search $R_{\text{opt}}^{(\text{calc})}$ will be strongly correlated with other kinematic variables listed in eq. (3.2). We nevertheless include it in the BDT because the general multivariate HEPTopTagger2 described in the appendix will not include the top momenta in the tagging. The increase of the tagging performance from the optimalR mode is shown in the left panel of figure 6. While for small signal efficiencies the curves for optimalR and for the variable mass setup of eq. (2.4) are identical within numerical fluctuations, we observe a significant improvement for larger signal efficiencies.

JHEP06(2015)203

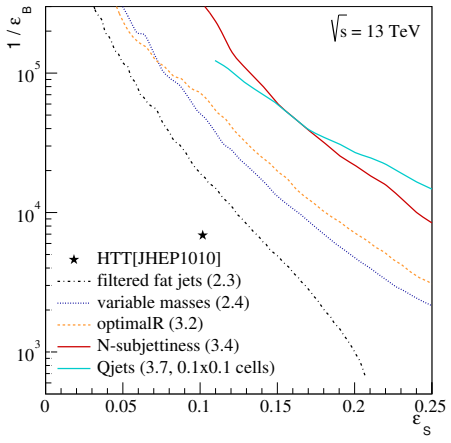

Figure 6. Performance of the optimalR mode based on the kinematic variables in eq. (3.2), including N-subjettiness variables as defined in eq. (3.4), and including QJETS. As described in the text, for QJETS we need to require a finite calorimeter resolution, while all other curves do not include any detector effects. We only consider the dominant QCD background.

N-subjettiness. The arguably simplest question we can ask as part of a top tagger is the number of hard subjets inside the fat jet with a given jet mass. This number of subjets can be defined through an observable similar to event shapes like for example thrust, called N-subjettiness [63–67]. It is based on N reference axes which are required to match the k hard substructures,

$$\tau_N = \frac{1}{R_0 \sum_k p_{T,k}} \sum_k p_{T,k} \min\left(\Delta R_{1,k}, \Delta R_{2,k}, \cdots, \Delta R_{N,k}\right), \tag{3.3}$$

where $\Delta R_{i,j}$ is the geometric separation between the axis i and the substructure j. In this form N-subjettiness parametrizes the deviation of the energy flow away from N jets not only related to an integer number of subjets, but also reflecting the color structure and the related radiation pattern.

In terms of original definition [63–65] we fix the exponent to $\beta = 1$. R_0 is an intrinsic cone size, chosen such that $\tau_N < 1$. Small values of $\tau_N \to 0$ indicate that the complete substructure is described by N axes, indicating that there are at most N relevant substructures. The ratio τ_N/τ_{N-1} will therefore become small for a fat jet with N hard subjets. For top tagging the ratio τ_3/τ_2 will be most useful and can even be used as a tagger itself. Higher τ_N values will contribute to a multivariate analysis of N-subjettiness, describing the jet radiation pattern around the assumed three partonic top decay momenta.

We will use N-subjettiness as an additional variable in our multivariate HEPTOPTAGGER. Originally, this combination did not lead to a significant improvement when added to the A-shaped cuts [35]. However, when we open the cut f_W on the reconstructed ratio m_W/m_t we observe a significant improvement for the extended set of kinematic variables. The complete set of relevant kinematic variables, now including N-subjettiness

JHEP06(2015)203

variables of fat jet i $\tau_{i,N}$ before and after filtering, is

$$\Big\{m_{tt}, m_{\rm ff}, p_{T,t_1}, p_{T,t_2}, p_{T,{\rm f}_1}, p_{T,{\rm f}_2}, m_{\rm rec}^{\rm min}, m_{\rm rec}^{\rm max}, f_{\rm rec}^{\rm max}, R_{\rm opt} - R_{\rm opt}^{\rm (calc)}, \tau_{1,N}, \tau_{1,N}^{\rm (filt)}, \tau_{2,N}, \tau_{2,N}^{\rm (filt)}\Big\} \qquad (N\text{-subjettiness}). \tag{3.4}$$

For more details on the N-subjettiness variables we refer to the appendix. As in eq. (3.2) all kinematic variables linked to the top tag will be evaluated with the fat jet size $R_{\rm opt}$. The details of implementation of the N-subjettiness variables are discussed in the appendix.

Qjets. The main limitation even of the deterministic multivariate HEPTopTagger is the aim to identify a unique set of subjets from the top decay as part of the tagging procedure, which allows us to reconstruct the 4-momentum of the tagger top and for example compare it to Monte Carlo truth. If the kinematic selection identifies a wrong set of subjets as the best candidates for the top decay products, an actual top decay can easily fail the tagging procedure. To avoid this loss in signal efficiency we can allow for more than one set of candidate subjets to be tested. One approach that not only covers several candidates of subjet combinations, but which even allows for a statistical analysis of many such assignments is Qjets [55, 56].

During the clustering of the fat jet the standard recombination algorithms combine the closest set of pre-jets according to a given measure. For the C/A algorithm this measure is the geometric separation $d_{ij} = \Delta R_{ij}^2$ of the pre-jets i and j. Qjets generalizes this deterministic choice to a likelihood measure. For each pair of pre-jets (i, j) it computes the weight

$$\omega_{ij}^{(\alpha)} = \exp\left(-\alpha\, \frac{d_{ij} - d_{ij}^{\rm min}}{d_{ij}^{\rm min}}\right), \tag{3.5}$$

and then chooses the two pre-jets to cluster according to a random number trailing the weights $\omega_{ij}^{(\alpha)}$. For this study we choose $\alpha = 0.1$, to balance the convergence of the algorithm with our aim of generating alternative subjet assignments for the top tagger. The standard jet algorithm corresponds to the limit $\alpha \to \infty$. The global weight for a clustering history is defined as

$$\Omega^{(\alpha)} = \prod_{\rm mergings} \omega_{ij}^{(\alpha)} = \left[\prod_{\rm mergings} \exp\left(-\frac{d_{ij} - d_{ij}^{\rm min}}{d_{ij}^{\rm min}}\right)\right]^{\alpha} \overset{\text{consistent}}{\longrightarrow} 1\,. \tag{3.6}$$

The universal limiting case $\Omega^{(\alpha)} \to 1$ for a perfect clustering history indicates that in searching for the largest global weight Ω the choice of α should not make a major difference. The Qjets clustering procedure can be repeated many times, where in this study we typically rely on 100 clustering histories. They can be ranked by their global weights $\Omega^{(\alpha)}$ instead of the independent local weights used by a deterministic jet algorithm. For each history we apply the unclustering and top tagging algorithm. As long as the deterministic jet algorithm picks a reasonable merging history for a signal event we expect the outcome of the deterministic tagger and the tagger acting on the clustering history with the highest global weight to be close.

JHEP06(2015)203

The first advantage of QJETS appears when during an early clustering step the deterministic measure d_{ij} identifies the wrong merging in the sense that the remaining history cannot be described well by QCD. This deterministic history will by definition receive the maximum global weight $\Omega^{(\alpha)} = 1$. However, an alternative history in better agreement with QCD could reach a similarly large global weight. Because QJETS provides many alternative clustering histories, we can search for a set of top tags with comparably large global weights. For example, we can use the two positively tagged QJETS histories with the highest global weight in the multivariate analysis. This way, a possibly misleading deterministic result is corrected. This should improve the performance in particular when we enforce high signal efficiencies, where the tagger becomes most vulnerable to a wrong clustering input. It turns out that already this simple modification gives a sizeable improvement in the signal efficiency.

The second improvement to the usual top tagger is based on HEPTOPTAGGER output for the full set of 100 clustering histories. First, we include the fraction of positive top tags based on the default HEPTOPTAGGER settings among all 100 QJETS histories, $\varepsilon_{\text{Qjets}}$, as introduced in the appendix. Next, we extract statistical information from distributions of the QJETS histories, like for example the reconstructed top mass m_{rec}. This distribution is defined for $\varepsilon_{\text{Qjets}} \times 100$ histories. Signal events will strongly peak around the top mass with a possible secondary peak around the W-mass. QCD background events will instead show a smooth decrease. The two most relevant observables in the m_{rec} distribution are the mean and the variance of this reconstructed top mass distribution with 100 entries, symbolically denoted as $\{m_{\text{rec}}^{\text{Qjets}}\}$.

Our multivariate analysis we base on the second approach. We start with the top-tagged QJETS history with the highest global weight and run the tagging algorithm of this history only. In addition, we include the statistical information of the m_{rec} distribution of the subset of the 100 QJETS histories which defines a top candidate. The complete list of observables including the QJETS information now reads

$$\Big\{ m_{tt}, m_{\text{ff}}, p_{T,t_1}, p_{T,t_2}, p_{T,\text{f}_1}, p_{T,\text{f}_2}, m_{\text{rec}}^{\min}, m_{\text{rec}}^{\max}, f_{\text{rec}}^{\max}, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}, \{\tau_N\}, \varepsilon_{\text{Qjets}}^{\min}, \{m_{\text{rec}}^{\text{Qjets}}\} \Big\} \quad \text{(QJETS)}, \tag{3.7}$$

where $\{\tau_N\}$ represents the appropriate set of filtered and unfiltered N-subjettiness variables (for example $N = 1, 2, 3$ for each of the two tops). For the two tags in the Z' analysis we choose the smaller $\varepsilon_{\text{Qjets}}$ value of the two. All variables from the tagger are evaluated for the optimized R size and the clustering history with the largest global weight.

In figure 6 we show the effect of the QJETS histories in addition to the other improvements. A key difference between the previous discussion and the QJETS approach is that we now need to include some kind of detector resolution, to limit QJETS to a manageable number of significantly different merging histories. For that reason we divide the calorimeter into $\eta \times \phi$ cells of size 0.1×0.1 and pre-cluster the entire set of calorimeter entries before applying any jet algorithm. Because this detector resolution effect is not included for the previous results, the QJETS ROC curve does not consistently exceed the N-subjettiness curve without QJETS. On the other hand, we still observe the expected im-

JHEP06(2015)203

provement towards large signal efficiencies. The moderate drop at small signal efficiencies gives us confidence that a full detector simulation will not lead to significant degradation of our results.

4 Full event information

Going back to the discussion in section 2 the remaining question is how the new HEPTopTagger2 performance compares to other approaches designed for the upcoming LHC run. The benchmark for such a comparison is event deconstruction, or more specifically the projections for a Z' resonance search [37]. As mentioned in our discussion of jet radiation in section 2 the borders between the hard process or the Z' decay on the one side and QCD jet radiation and its sensitivity to the signal and background color structure on the other side are washed out when we include for example filtered subjets or N-subjettiness information. We therefore start with a brief discussion of the additional information from jets in the entire event and then move on to the comparison with the leading benchmark in proposed Z' analyses.

Additional jets. To determine to what degree the jet structure of purely hadronic $Z' \to t\bar{t}$ events helps the extraction of the signal from the $t\bar{t}$ and QCD jets background we first study the number and kinematic distribution of small C/A jets with $R = 0.2$ and $p_{T,j} > 10\,\text{GeV}$ in addition to the fat jets fulfilling eq. (2.1). We choose these very small jets in order to test information which might be available from so-called microjets in shower deconstruction. Our discussion should not be applied to an LHC analysis one-to-one and is instead aimed at capturing as much information as possible. Without any major cuts, the number of jets will consist of three decay jets per top quark, FSR jets, and ISR jets. For an inclusive event sample, we should be able to tell apart the different processes from the number of jets and the kinematics of the individual jets [70–74].

After a first level of cuts we see in figure 7 that the Z' signal and the $t\bar{t}$ background both peak at 10 microjets, e.g. four jets from ISR and FSR combined. For the background the number is slightly larger, because we generate the scale of the hard process also through a large number of jets. We also see that the transverse momentum of the hardest jet is slightly larger for the signal. We could include these jet patterns in a multivariate analysis, but at this stage this information would be very heavily correlated with the variables from the top tagger.

In a second step we focus on the jet activity which does not contribute to the top tagging. Inside the fat jets we know that the top tagger includes information based on subjets with typically $R = 0.3$ and $p_T \gtrsim 20\,\text{GeV}$ after filtering. After two tags we then remove all calorimeter data associated with the filtered triplet of either of the top candidates and re-cluster the remnants into microjets with $R = 0.2$ and $p_{T,j} > 10\,\text{GeV}$. In the lower panels of figure 7 we see how after removing the signal decay jets the remaining number of jets peaks around two ISR or FSR jets. For the QCD background this number is higher, because it takes a larger number of equally distributed jets in the detector to fake a boosted massive top inside each fat jet. The transverse momentum of the hardest of the remaining

JHEP06(2015)203

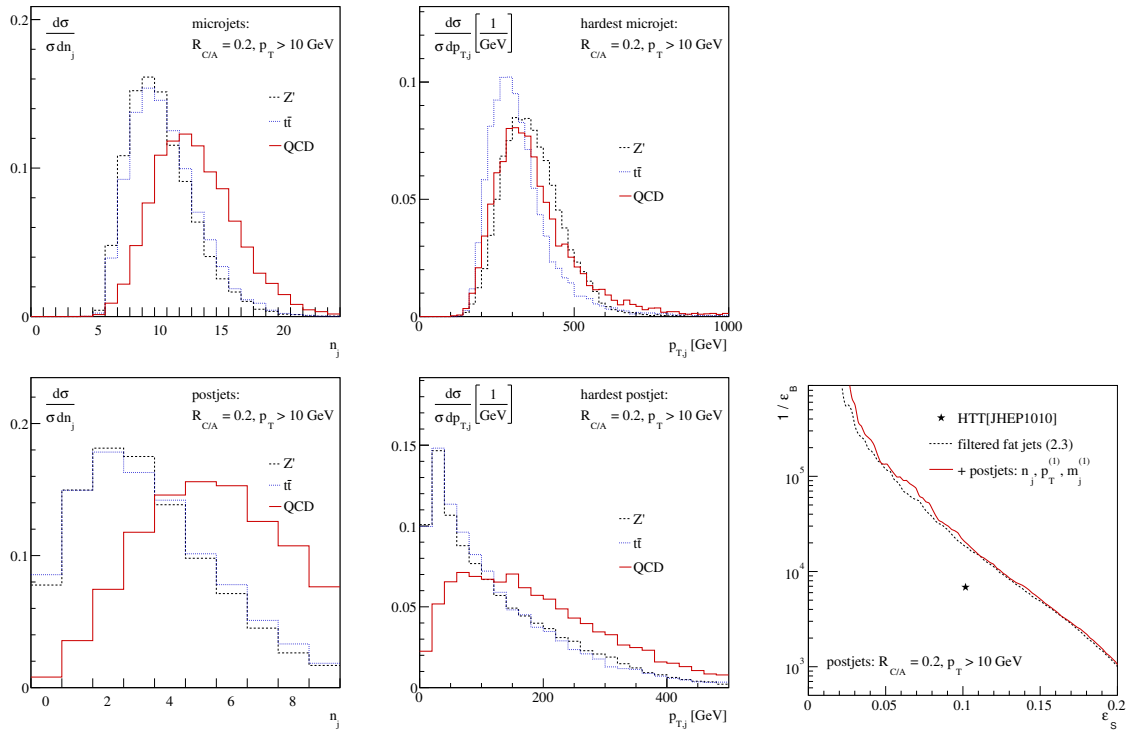

Figure 7. Information on the hardest jet before top tagging (upper row) and the hardest jet left over after top tagging (lower row). For the jets defined with $R = 0.2$ and $p_T > 10\,\text{GeV}$ we show the number of jets, the hardest jet's transverse momentum, and its mass in Z' candidate events (left to right).

QCD jets also peaks at very small values for the signal and the $t\bar{t}$ background, as one would expect for example for a small number of ISR jets. The bulk of the hardest QCD jets per event shows transverse momenta around $p_{T,j} = 50 - 200\,\text{GeV}$, still small compared to the hard scale imprinted on the multi-jet background through the kinematic selection of eq. (2.1). We should be able to use this additional information for our BDT analysis, to improve the signal extraction. In the right panel of figure 7 we see the corresponding ROC curve. It turns out that almost all of the information available through the extra jet radiation is already included in our combined analysis of top tags and subjet kinematics.

Based on this piece of information we assume that additional jet information inside and outside the fat jets hardly changes the stable results of the updated top tagger, so we can compare the new HEPTopTagger2 to other multivariate methods.

Comparison with other approaches. The most promising projections for boosted top identification and specifically searches for $t\bar{t}$ resonances during the upcoming LHC runs are available for shower deconstruction [36] or event deconstruction [37]. This method is based on a construction of likelihoods representing possible shower histories for a jet or a fat jet. The underlying objects are so-called C/A [45, 46, 50, 51] microjets with $R = 0.2$ and $p_T > 10\,\text{GeV}$ [37]. They are slightly softer and smaller than the subjets in a typical top tagger, but we have seen that the additional information from those jets should not make a big difference. Unlike general template methods, shower deconstruction relies on the

JHEP06(2015)203

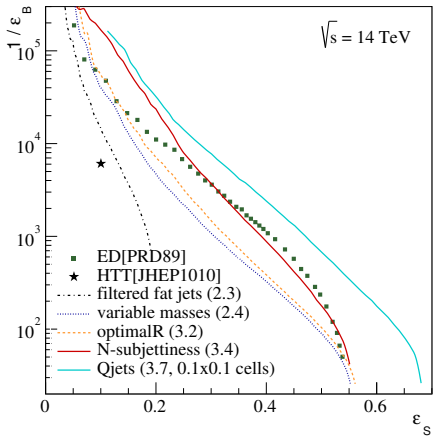

Figure 8. Comparison of the multivariate HEPTopTagger2 analysis presented in this paper with the event deconstruction approach of ref. [37]. All HEPTopTagger2 curves correspond to figure 6, but now with a collider energy of 14 TeV instead of 13 TeV, This comparison in the absence of an experimental validation should be taken as first estimate.

soft and/or collinear approximation of QCD to compute the likelihood of a given shower history in terms of splitting probabilities and Sudakov factors (non-splitting probabilities). Based on the possible shower histories the likelihood ratio of a fat jet coming from a boosted top quark or from the QCD jet background acts as a measure for the top tag. One problem with shower deconstruction, like any probabilistic approach, is that we cannot separate the identification and the reconstruction of the boosted top quark. This means we cannot for example show the quality of the reconstructed 4-momentum compared to Monte Carlo truth.

The Z' analysis using event deconstruction starts with two fat jets of size $R = 1.5$ and the acceptance cuts given in eq. (2.1). The number of microjets is limited to 9 per fat jet. In addition to the likelihood separating the top or QCD origin of each of the two fat jets, the event likelihood measure now also includes a likelihood describing the resonant or non-resonant production of the pair of fat jets given their 4-momenta. At the level of the hard process this part is not very different from the established matrix element method [75–79] and largely replaces an analysis of the m_{tt} and $p_{T,t}$ distributions defining the multivariate analysis of eq. (2.2). In ref. [37] the observable width of the m_{tt} resonance is assumed to range around 65 GeV, an assumption we follow. In our analysis the precise resolution for example after detector effects only plays a secondary role, because the resolution of the HEPTopTagger2 is limited to 145 GeV, as shown in table 2.

In figure 8 we show the performance of the analysis developed in this paper with the recent benchmark of event deconstruction. One difference to the HEPTopTagger results shown in figure 6 is that we now show Z' efficiencies up to 68%, confirming that Qjets indeed gives us a major improvement for very large signal efficiencies. Another difference is that for a direct comparison we now assume a collider energy of 14 TeV. Both, event deconstruction and the new HEPTopTagger show a comparable performance for the upcoming run. The final answer on both methods will only be given by experimental studies including data.

JHEP06(2015)203

5 Conclusion

We demonstrated how the updated HEPTopTagger2 performs in searches for Z' bosons or other heavy resonances decaying to top pairs in the upcoming LHC run. Based on the original HEPTopTagger [14, 15] we modify the tagging algorithm and add several additional kinematic variables to a multivariate analysis:

- fat jet kinematics to account for final-state radiation in resonance searches;
- algorithmically optimized size of the original fat jet combined with its prediction (optimalR mode);
- N-subjettiness probing the more general subjet structures inside the fat jet;
- Qjets with a global picture of the most likely clustering histories giving a top tag.

Each of these improvements can be added to the top tagging individually. For the specific Z' resonance search we altogether achieve an increase of the background rejection by a factor of 30 for a constant Z'-signal efficiency of 10%. Compared to the original tagger [14, 15] the background sculpting in the invariant mass of the top pair is significantly reduced [35]. These updated results are at least competitive with the leading estimates for other tagging methods.

Because the multivariate Z' analysis includes several layers of improvement, not necessarily linked to the actual top tagging, we also show in the appendix the corresponding improvements for top tagging in $t\bar{t}$ events. There, we test the updated tagger for moderate ($p_{T,t} > 200\,\text{GeV}$) and sizeable ($p_{T,t} > 600\,\text{GeV}$) boost and find a significant improvement in particular for larger boost. The limiting factor for moderate boost still is capturing all three top decay jets inside a fat jet, which has to be targeted by a dedicated low-p_T mode [41]. The corresponding HEPTopTagger2 described in the appendix will be made publicly available [14, 15, 80]. In particular for Qjets there exist different modes which need to be tested on data.

Comparing the improvement of the Z' analysis with that in the individual top tags shows that the benefits for the full Z' case are significantly larger than those just from the top tags. A lesson from this is that it is useful to consider the optimization of top tagging, not only in its own right, but also in the context of full search analyses.

Acknowledgments

T.S. acknowledges support by the IMPRS for Precision Tests of Fundamental Symmetries. GPS acknowledges partial support from ERC advanced grant Higgs@LHC.

JHEP06(2015)203

A HEPTopTagger2

In the past it has proven useful to publish details about the HEPTopTagger algorithm. We describe the new structure reflecting all changes in refs. [14, 15, 35, 41] in this appendix. Because the main body of the paper is focused on the performance in resonance searches we then present benchmark results based on purely hadronic $t\bar{t}$ events in the Standard Model. They can be directly translated for example into semi-leptonically decaying $t\bar{t}$ pairs. Finally, the enhanced capabilities of the HEPTopTagger2 have lead to enough of a complexity of the actual code that we briefly describe the run modes, the input parameters, and the available output information from the tagger.

Algorithm. The basic HEPTopTagger2 algorithm largely follows the original algorithm described in refs. [14, 15], but is based on FastJet3 [50, 51] and includes a number of new features:

1. Define a C/A fat jet with $R_{\text{fat}} = 1.8$ and determine the splitting history through the default clustering.

2. identify all hard subjets using a mass drop criterion: undo the last clustering of the jet j, into two subjets j_1, j_2 with $m_{j_1} > m_{j_2}$; require $m_{j_1} < f_{\text{drop}}\, m_j$ with $f_{\text{drop}} = 0.8$ to keep both; otherwise, keep only j_1; further decompose or add each subjet j_i to the list of relevant substructures. A global soft cutoff $m_{j_i} > m_{\text{min}} = 30\,\text{GeV}$ can be adjusted.[2]

3. Iterate through all triplets of three hard subjets: filter them with resolution $R_{\text{filt}} = \min(0.3, \Delta R_{jk}/2)$; use the $N_{\text{filt}} = 5$ hardest filtered constituents and calculate their combined jet mass; re-cluster these five subjets into three assumed top decay jets; reject all triplets outside $m_{123} \equiv m_{\text{rec}} \in [150, 200]\,\text{GeV}$; keep the event if at least one such triplet exists. For the multivariate analysis this window is opened to $m_{\text{rec}} < 1\,\text{TeV}$, which allows us to use m_{rec} as a kinematic output of the tagger.

   This set of re-clustering and filtering steps by default uses the C/A jet algorithm [45, 46]. However, to guarantee infrared safety and enhance the performance at large boosts [41] it can be switched to k_T jets [47–49].

4. Order the three subjets j_1, j_2, j_3 by p_T; if the masses (m_{12}, m_{13}, m_{23}) satisfy one of the following three criteria, accept them as a top candidate:

$$0.2 < \arctan\frac{m_{13}}{m_{12}} < 1.3 \quad \text{and} \quad R_{\text{min}} < \frac{m_{23}}{m_{123}} < R_{\text{max}} \tag{A.1}$$

$$R^2_{\text{min}}\left(1+\left(\frac{m_{13}}{m_{12}}\right)^2\right) < 1 - \left(\frac{m_{23}}{m_{123}}\right)^2 < R^2_{\text{max}}\left(1+\left(\frac{m_{13}}{m_{12}}\right)^2\right) \text{ and } \frac{m_{23}}{m_{123}} > 0.35$$

$$R^2_{\text{min}}\left(1+\left(\frac{m_{12}}{m_{13}}\right)^2\right) < 1 - \left(\frac{m_{23}}{m_{123}}\right)^2 < R^2_{\text{max}}\left(1+\left(\frac{m_{12}}{m_{13}}\right)^2\right) \text{ and } \frac{m_{23}}{m_{123}} > 0.35$$

[2] We have checked that replacing the mass drop criterion with a soft drop criterion [33] does not improve the performance of the tagger noticeably.

JHEP06(2015)203

where $R_{\text{min,max}} = (1 \mp f_W) m_W/m_t$ defines the parameter f_W, by default set to $f_W = 0.15$. The soft cutoff $m_{23} > 0.35\, m_{123}$ as well as the limits $[0.2, 1.3]$ in the first line can be adjusted. All kinematic cuts are listed in table 5 and can be adapted in a multivariate approach. In the multivariate case we open the W-mass window to $f_W = 0.3$. The ratio of the W-mass to the top mass can then be used as a kinematic output defined as

$$f_{\text{rec}} = \min_{ij} \left| \frac{\frac{m_{ij}}{m_{123}}}{\frac{m_W}{m_t}} - 1 \right| . \tag{A.2}$$

5. Of all triplets passing the above criteria in a given fat jet choose the one with $m_{123} \equiv m_{\text{rec}}$ closest to m_t. This selection has shown to be the most efficient, and applying it after all kinematic cuts minimizes the background sculpting. The m_{rec} and f_{rec} values supplied to the multivariate analysis are those corresponding to this triplet.

6. For consistency, require the reconstructed $p_{T,t}$ to exceed 200 GeV.

7. In the low-p_T mode [41] reduce this threshold to $p_{T,t} > 150\,\text{GeV}$; compute the Fox-Wolfram moments [52–54]

$$\begin{aligned} H_\ell^x &= \sum_{i,j=1}^{N} W_{ij}^x \, P_\ell(\cos\Omega_{ij}) \\ \text{with} \quad W_{ij}^T &= \frac{p_{Ti}\, p_{Tj}}{\left(\sum p_{Ti}\right)^2} \qquad \text{and} \qquad W_{ij}^U = \frac{1}{N^2} , \end{aligned} \tag{A.3}$$

   of the subjets relative to each other and relative to the reconstructed top momentum. This mode is not part of the usual tagger and relies on external GSL libraries [81] for Legendre polynomials.

8. In the optimalR mode repeat steps 1 to 3 with a decreasing fat jet radius in steps of $\Delta R = 0.1$; based on the condition $m_{\text{rec}}^{(1.8)} - m_{\text{rec}} > 0.2 m_{\text{rec}}^{(1.8)}$ determine the minimum radius $R_{\text{opt}} > 0.5$; follow steps 4 to 6 with this modified fat jet. We also parametrize the expected value for R_{opt} in terms of $p_{T,\text{f}}$ based on the numerical simulation of the top decay kinematics illustrated in figure 9

$$R_{\text{opt}}^{(\text{calc})} = \frac{327}{p_{T,\text{f}}} . \tag{A.4}$$

9. In the N-subjettiness mode [35] compute the τ_j [63–65] as defined in eq. (3.3) from the filtered and unfiltered subjets, as described below. Again, this mode is not part of our tagger code and relies on the FastJet Contrib [50, 51, 80] add-on for N-subjettiness [63–65].

10. In the Qjets mode replace the deterministic output of step 1 by a set of possible histories defined in eq. (3.6); run the tagger for each of them, giving a set of clustering histories with global weights Ω, and a positive or negative tagging result.

JHEP06(2015)203

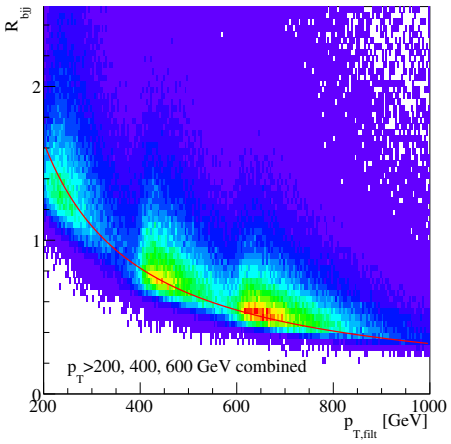

Figure 9. $R_{\text{opt}}^{(\text{calc})}$ fit based on Standard Model $t\bar{t}$ samples with $p_{T,t} > 200, 400, 600\,\text{GeV}$ for the parton level distance of decay products R_{bjj}. The fat jets are filtered with $R = 0.2$, $N = 10$. The functional form of the fit curve is given in eq. (A.4).

Following this description the low-p_T (7) and N-subjettiness (9) modes simply add kinematic observables to the tagger output. These observables can be included in a multivariate analysis or can be cut on in the deterministic top tagging decision. The improvement in the low-p_T mode is illustrated in detail in ref. [35] while the impact of N-subjettiness variables on the resonance search is illustrated in figure 6.

In contrast, the optimalR mode and the QJETS mode modify the clustering histories (1) underlying the mass drop search (2). Depending on the modified fat jet size or on the QJETS weight they return a set of tagging outputs. For the optimalR mode it is straightforward to choose the smallest reasonable fat jet size R_{opt} for the actual tagging. The QJETS histories can be evaluated in a range of possible ways.

Performance. The main body of this paper focuses on $t\bar{t}$ resonance searches using the HEPTOPTAGGER described above. While the combination of tagged top kinematics and fat jet kinematics in section 2 does not directly translate into to a universal top tagger, the multivariate aspects discussed in section 3, namely optimalR, N-subjettiness, and QJETS do. Here, we show efficiencies for extracting $t\bar{t}$ events from the QCD multi-jet background.

Our analyses are based on fully hadronic $t\bar{t}$ signal and QCD dijet background samples generated with PYTHIA8 [59]. For the general top tagger analysis in this appendix we include underlying event in the event generation and mimic the limited detector resolution by clustering the hadronic activity into $\eta \times \phi$ cells of size 0.1×0.1, similar to the QJETS results shown in figure 6. Instead of the hard acceptance cuts in eq. (2.1) we now allow for softer fat jets. Two multivariate BDT analyses focus on $t\bar{t}$ samples with

$$p_{T,\text{fat}} > 200\,\text{GeV} \qquad |y_{\text{fat}}| < 2.5 \qquad p_{T,t} > 200, 600\,\text{GeV}\,, \tag{A.5}$$

where the top momenta are evaluated on the Monte Carlo truth level. We select events with fat C/A jets of radius $R_{\text{fat}} = 1.8$ and $|y_{\text{fat}}| < 2.5$ constructed with FASTJET.

Background efficiencies ε_B are defined as relative to the number of those fat jets. For the signal efficiencies we require that the fat jets can be matched to a parton level top

JHEP06(2015)203

quark within $\Delta R < 0.8$. Using the original version of the HEPTopTagger [14, 15] we find for the $p_T > 600\,\text{GeV}$ samples a signal efficiency of $\varepsilon_S = 35.6\%$ and a mis-tagging rate $\varepsilon_B = 2.7\%$. The first change in the algorithm addresses the signal efficiency and background sculpting. In the original algorithm the triplet of subjets closest to the true top mass is selected and only later the mass plane cuts are applied. Therefore, the tagger will fail if this triplet does not pass the mass plane constraints and no alternative triplet is analyzed. To eliminate this limitation, we first apply the mass plane constraints and then pick the triple closest to the top mass, as described above.

As in the main text we study further improvements of the tagger based on ROC curves. To allow for such improvements we loosen the cuts of the tagger to $m_\text{rec} < 1\,\text{TeV}$ and $f_W = 0.3$. The initial set of BDT parameters in analogy to eq. (2.4) is

$$\{m_\text{rec}, f_\text{rec}\} \qquad \text{(variable masses)}. \tag{A.6}$$

The large cone size of $R = 1.8$ is not always appropriate, so the optimalR mode optimizes the radius of each fat jet. Starting from the initial cone size we stepwise reduce the size of the fat jet until the criterion eq. (3.1) indicates that we miss a top decay jet. For the last stable R size we run the usual tagging algorithm. We can calculate the expected value R_opt^calc for the critical radius based on the transverse momentum of the filtered fat jet. For a fat jet originating from a top decay this prediction should agree with the measured value, while for a background fat jet the two are only strongly correlated when the entire subjet kinematics is a perfect match to a top decay. For the optimalR mode we set up a BDT analysis with the observables

$$\left\{m_\text{rec}, f_\text{rec}, R_\text{opt} - R_\text{opt}^{(\text{calc})}\right\} \qquad \text{(optimalR)}. \tag{A.7}$$

All tagging observables are evaluated for a fat jet with size R_opt. In figure 10 we show the improvement from the optimized size of the fat jet. Obviously, it is more impressive for larger boost, while for $p_{T,t} > 200\,\text{GeV}$ the optimalR mode hardly leads to a reduction in fat jet size.

The N-subjettiness variables are best applied independently for fat jets which would pass and would not pass the initial tagging criterion. The optimalR working point

$$m_\text{rec} \in [150, 200]\,\text{GeV} \qquad f_\text{rec} < 0.175 \qquad R_\text{opt} - R_\text{opt}^{(\text{calc})} < 0.3\,, \tag{A.8}$$

which corresponds to the signal efficiency $\varepsilon_S = 0.22(0.27)$ in figure 10, defines these two categories. Fat jets passing eq. (A.8) can be assumed to include a complete set of top decay products and are filtered with $R_\text{filt}^{(1)} = 0.2$ and $N_\text{filter}^{(1)} = 5$; fat jets failing this criterion are instead filtered with $R_\text{filt}^{(0)} = 0.3$ and $N_\text{filter}^{(0)} = 3$. The unfiltered N-subjettiness variables τ_i defined in eq. (3.3) and their filtered counter parts $\tau_i^{(0)}, \tau_i^{(1)}$ are included up to $i \leq 3$. The

JHEP06(2015)203

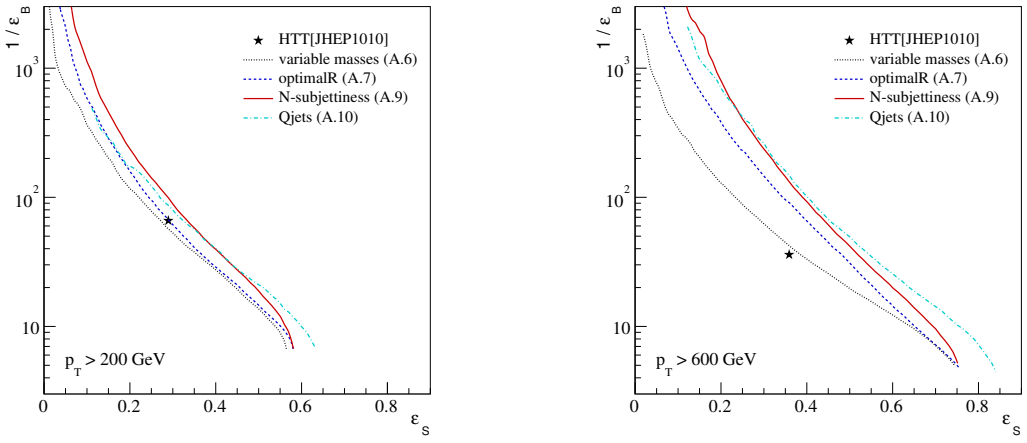

Figure 10. Performance of the HEPTopTagger2 for $t\bar{t}$ production in the Standard Model. We show the incremental improvements from the extended multivariate analyses for top quarks with $p_{T,t} > 200\,\text{GeV}$ and $p_{T,t} > 600\,\text{GeV}$.

reference axes are chosen as k_T-axes. We then set up two independent BDTs with

$$\begin{aligned} &\Big\{m_{\text{rec}}, f_{\text{rec}}, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}, m_{\text{fat}}^{(1)}, \tau_3^{(1)}, \tau_3^{(1)}/\tau_2^{(1)}, \tau_2^{(1)}/\tau_1^{(1)}, \\ &\qquad \tau_2, \tau_3/\tau_2, \tau_2/\tau_1\Big\} \qquad (N\text{-subjettiness, pass}) \\ &\Big\{m_{\text{rec}}, f_{\text{rec}}, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}, m_{\text{fat}}^{(0)}, \tau_3^{(0)}, \tau_3^{(0)}/\tau_2^{(0)}, \tau_2^{(0)}/\tau_1^{(0)}, \\ &\qquad \tau_1, \tau_3/\tau_2, \tau_2/\tau_1\Big\} \qquad (N\text{-subjettiness, fail}), \end{aligned} \tag{A.9}$$

and later combine them into one ROC curve. This precise condition is represented by the more generic eq. (3.4). In figure 10 we show the corresponding ROC curves for a successively improved tagger.

Finally, we can replace the deterministic clustering history from the usual jet algorithm with a set of QJets histories with large global weights $\Omega^{(\alpha)}$ defined in eq. (3.6) for $\alpha = 0.1$. This way we avoid cases where the deterministic clustering history entering the top tagging algorithm is misled during the independent evaluation of splittings in the usual jet algorithm. When defining jets as analysis objects for a hard process this does not pose a problem, but for subjet analyses it can have an effect.

Our analysis is based on 100 QJets histories per fat jet. In table 3 we show their signal and background efficiency if required to lead to individual top tags. As the reference value we use the default HEPTopTagger with fixed mass windows. Based on 100 QJets histories we then define the fraction $\varepsilon_{\text{Qjets}}$ of histories which lead to a top tag with the default tagging setup. We see that for moderately boosted tops the deterministic signal tagging efficiency can be reproduced by requiring 30% of the QJets histories to deliver a positive tag. The corresponding mis-tag probability is slightly reduced compared to the deterministic tagger. For harder tops the corresponding value is around $\varepsilon_{\text{Qjets}} > 20\%$, with no improvement in the background rejection.

As discussed in section 3 QJets offers two strategies to improve the top tagger. To maximize the improvement in the tagging performance and to limit the CPU time we

JHEP06(2015)203

	$t\bar{t}$	QCD
default HTT	0.337	0.0212
$\varepsilon_{\text{Qjets}} > 0.1$	0.435	0.0318
$\varepsilon_{\text{Qjets}} > 0.2$	0.384	0.0231
$\varepsilon_{\text{Qjets}} > 0.3$	0.341	0.0174
$\varepsilon_{\text{Qjets}} > 0.4$	0.298	0.0123
$\varepsilon_{\text{Qjets}} > 0.5$	0.250	0.0089
$\varepsilon_{\text{Qjets}} > 0.6$	0.212	0.0064
$\varepsilon_{\text{Qjets}} > 0.7$	0.163	0.0036
$\varepsilon_{\text{Qjets}} > 0.8$	0.118	0.0021
$\varepsilon_{\text{Qjets}} > 0.9$	0.064	0.0007

	$t\bar{t}$	QCD
default HTT	0.465	0.0489
$\varepsilon_{\text{Qjets}} > 0.1$	0.524	0.0661
$\varepsilon_{\text{Qjets}} > 0.2$	0.447	0.0461
$\varepsilon_{\text{Qjets}} > 0.3$	0.388	0.0342
$\varepsilon_{\text{Qjets}} > 0.4$	0.336	0.0245
$\varepsilon_{\text{Qjets}} > 0.5$	0.281	0.0168
$\varepsilon_{\text{Qjets}} > 0.6$	0.236	0.0118
$\varepsilon_{\text{Qjets}} > 0.7$	0.181	0.0062
$\varepsilon_{\text{Qjets}} > 0.8$	0.133	0.0032
$\varepsilon_{\text{Qjets}} > 0.9$	0.069	0.0009

Table 3. Tagging efficiencies for $p_T > 200\,\text{GeV}$ (left) and $p_T > 600\,\text{GeV}$ (right). $\varepsilon_{\text{Qjets}}$ is defined as the number of QJETS tags per number of QJETS runs. For this table we test 10.000 fat jets with 100 QJETS iterations.

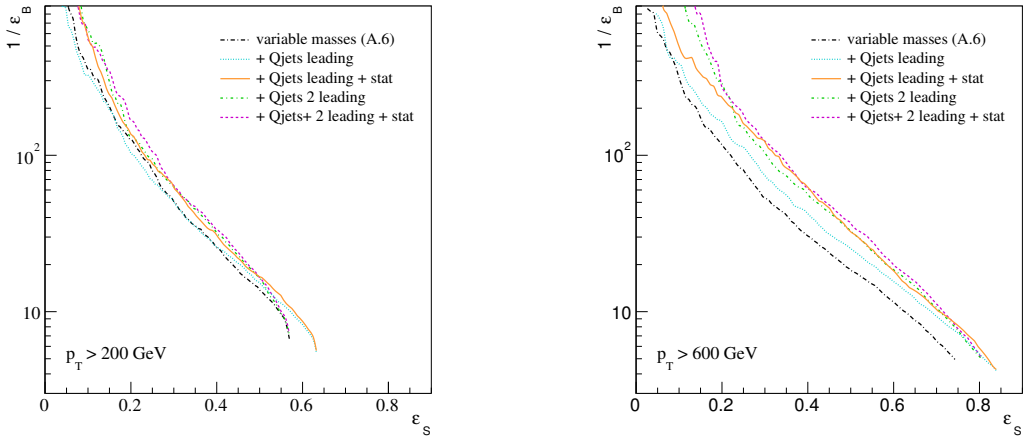

Figure 11. Performance of the HEPTOPTAGGER2 for $t\bar{t}$ production in the Standard Model. For $p_{T,t} > 200\,\text{GeV}$ and $p_{T,t} > 600\,\text{GeV}$ we focus on different QJETS setups, based on a more basic multivariate tagger without optimalR and N-subjettiness.

base the multivariate analysis on the tagged history with the largest global weight. As additional parameters we include the value of $\varepsilon_{\text{Qjets}}$ as well as the mean and variance of the m_{rec} distribution with the 100 QJETS entries, symbolically denoted as $\{m_{\text{rec}}^{\text{Qjets}}\}$. For the BDT analysis the variables are

$$\left\{m_{\text{rec}}, f_{\text{rec}}, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}, m_{\text{fat}}, \tau_N, \tau_N^{(\text{filt})}, \varepsilon_{\text{Qjets}}, \{m_{\text{rec}}^{\text{Qjets}}\}\right\} \qquad (\text{QJETS}) \qquad (\text{A.10})$$

As usual, all variables from the tagger are evaluated for the optimized R size and the clustering history with the largest global weight. The additional improvement is shown in figure 10.

Because QJETS offers a variety of improvements to the tagger, we study different setups based on the stage with multivariate mass windows in figure 11. We start by replacing

JHEP06(2015)203

name	description
`EARLY_MASSRATIO_SORT_MASS`	apply the 2D mass plane requirements, then select the candidate which minimizes $\lvert m_{\text{cand}} - m_t \rvert$
`LATE_MASSRATIO_SORT_MASS`	select the candidate which minimizes $\lvert m_{\text{cand}} - m_t \rvert$
`EARLY_MASSRATIO_SORT_MODDJADE`	apply the 2D mass plane requirements, then select the candidate with the highest modified Jade distance
`LATE_MASSRATIO_SORT_MODDJADE`	select the candidate with the highest modified Jade distance
`TWO_STEP_FILTER`	only analyze the candidate built with the highest $p_{T,t}$ after unclustering

Table 4. HEPTopTagger working modes.

the deterministic C/A output with the most likely Qjets history and including $\varepsilon_{\text{Qjets}}$ in the multivariate analysis. This leads to a moderate improvement of the tagger at large transverse momenta and at large signal efficiencies. Adding the statistical information from the $\varepsilon_{\text{Qjets}} \times 100$ entries in the m_{rec} information leads to a sizeable improvement over a wide range of signal efficiencies. This is the mode we use for the Z' analysis as well as in figure 10.

Next, we add the second-best Qjets history to the tagger, such that the multivariate tagger (including $\varepsilon_{\text{Qjets}}$) is free to construct a criterion based on one or two tags in the two best Qjets histories. For most of the ROC curves this comparably simple approach is as successful as the full statistical information. Finally, adding the statistical information on the m_{rec} distribution leads to a mild improvement.

Interface. To apply the HEPTopTagger algorithm to a fat C/A jet constructed with FastJet3 [50, 51], the only necessary steps are executing the default constructor `HEPTopTagger(fastjet::PseudoJet jet)` followed by running the tagger using `run()`. This will analyze the fat jet using the optimalR procedure with the default settings given in table 5. The available operation modes are shown in table 4. All configurable parameters are listed in table 5. Functions to retrieve results are presented in table 6. `QHTT()` sets up the Qjets mode. It is applied to a fully configured HEPTopTagger by `void run(HEPTopTagger htt)`. All configurable parameters are given in table 7. A list of functions to access the results is presented in table 8.

In addition, we provide a framework for the calculation of Fox-Wolfram moments that relies on an existing installation of GSL [81]. While the constructor `FWM(vector<fastjet::PseudoJet> jets)` allows the calculation of Fox-Wolfram moments for a given set of jets, `FWM(HEPTopTagger htt, unsigned selection)` uses the b, W_1, and W_2 momenta from the HEPTopTagger run and calculates the Fox-Wolfram moments in the top rest frame. The boost axis $\vec{a}$ itself can be included [35]. Subsets of these four vectors can be set via `unsigned selection`, as a sequence of 0 or 1 in the order abW_1W_2. In table 9 we show how to extract the Fox-Wolfram moment of a given order of the Legendre polynomials.

Finally, we include an example class `LowPt()` for a fixed low-p_T mode working point returning a tagging decision including the set low-p_T mode by `is_tagged(HEPTopTagger)`.

JHEP06(2015)203

name	default	description
general:		
`do_optimalR(bool)`	`true`	use optimalR approach
unclustering:		
`set_mass_drop_threshold(double)`	0.8	mass drop threshold
`set_max_subjet_mass(double)`	30	max subjet mass for unclustering
filtering:		
`set_filtering_R(double)`	0.3	max subjet distance for filtering
`set_filtering_n(unsigned)`	5	max subjet number for filtering
`set_filtering_minpt_subjet(double)`	0.	min subjet p_T for filtering
`set_filtering_jetalgorithm(fastjet::JetAlgorithm)`	`cambridge_algorithm`	jet algorithm for filtering
reclustering:		
`set_reclustering_jetalgorithm(fastjet::JetAlgorithm)`	`cambridge_algorithm`	jet algorithm for reclustering
candidate selection:		
`set_mode(enum)`	`EARLY_MASSRATIO_SORT_MASS`	run mode, see Tab. 4
`set_mt(double)`	172.3	true top mass
`set_mw(double)`	80.4	true W mass
`set_top_mass_range(double, double)`	150, 200	top mass window
`set_fw(double)`	0.15	width of A-shaped bands f_W
`set_mass_ratio_range(double, double)`	$(1-f_W)\,m_W/m_t = 0.397$ $(1+f_W)\,m_W/m_t = 0.537$	width of cut in 2D mass plane
`set_mass_ratio_cut(double, double, double)`	0.35, 0.2, 1.3	boundaries in 2D mass plane
`set_top_minpt(double)`	200	min $p_{T,t}$ consistency cut
pruning:		
`set_pruning_zcut(double)`	0.1	z_{cut} for pruned mass m_{prune}
`set_pruning_rcut_factor(double)`	0.5	r_{cut} for pruned mass m_{prune}
optimalR:		
`set_optimalR_max(double)`	size of the input fat jet	max jet size
`set_optimalR_min(double)`	0.5	min jet size
`set_optimalR_step(double)`	0.1	step size (multiple of 0.1)
`set_optimalR_threshold(double)`	0.2	optimalR mass threshold
calculation of $R_{\text{opt}}^{(\text{calc})}$:		
`set_filtering_optimalR_calc_R(double)`	0.2	max subjet distance for filtering
`set_filtering_optimalR_calc_n(unsigned)`	10	max subjet number for filtering
`set_optimalR_calc_fun(double (*f)(double))`	$327/p_{T,\text{filt}}$	dependency of $R_{\text{opt}}^{(\text{calc})}$ on $p_{T,\text{filt}}$
optimalR type:		
`set_optimalR_type_top_mass_range(double, double)`	150. 200.	mass range for optimalR type 1
`set_optimalR_type_f_rec(double)`	0.175	max f_{rec} for optimalR type 1
`set_optimalR_type_max_diff(double)`	0.3	max $R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}$ for optimalR type 1
N-subjettiness:		
`set_filtering_optimalR_pass_R(double)`	0.2	R_{filt} for optimalR type 1
`set_filtering_optimalR_pass_n(unsigned)`	5	N_{filt} optimalR type 1
`set_filtering_optimalR_fail_R(double)`	0.3	R_{filt} for optimalR type 0
`set_filtering_optimalR_fail_n(unsigned)`	3	N_{filt} for optimalR type 0

Table 5. Additional parameters of the HEPTopTagger algorithm. All functions have a return type of `void`.

JHEP06(2015)203

name	description
`bool is_maybe_top()`	top mass window requirement passed?
`bool is_masscut_passed()`	2D mass plane requirements passed?
`bool is_minptcut_passed()`	candidate $p_{T,t}$ threshold passed?
`bool is_tagged()`	top mass window, 2D mass plane requirement, and $p_{T,t}$ threshold passed?
`double delta_top()`	$\lvert m_{\rm rec} - m_t \rvert$
`double djsum()`	modified Jade distance
`double pruned_mass()`	pruned top mass
`double unfiltered_mass()`	mass of the triplet of subjets after unclustering before filtering
`double f_rec()`	minimal $\lvert (m_{ij}/m_{\rm rec})/(m_W/m_t) - 1 \rvert$
`const PseudoJet & t()`	top candidate 4-vector
`const PseudoJet & b()`	subjet corresponding to the b
`const PseudoJet & W()`	combined subjets corresponding to the W
`const PseudoJet & W1()`	leading subjet from the W
`const PseudoJet & W2()`	sub-leading subjet from the W
`const std::vector<PseudoJet> & top_subjets()`	three subjets from the top, ordered: b, W_1, W_2
`const PseudoJet & j1()`	leading subjet
`const PseudoJet & j2()`	sub-leading subjet
`const PseudoJet & j3()`	sub-sub-leading subjet
`const std::vector<PseudoJet> & top_hadrons()`	all top constituents
`const std::vector<PseudoJet> & hardparts()`	hard subtructures after unclustering, sorted by p_T
`const PseudoJet & fat_inital()`	original fat jet (after QJETS reclustering)
`const PseudoJet & fat_Ropt()`	fat jet reduced to $R_{\rm opt}$
`void get_setting()`	print settings to `stdout`
`void get_info()`	print tagger information to `stdout`
`HEPTopTagger HTTagger(unsigned i)`	HEPTOPTAGGER candidate for a distance parameter $R = i/10$. By default all functions above return values at $R = R_{\rm opt}$. This function accesses candidates for different values of R.
`double Ropt()`	$R_{\rm opt}$
`double Ropt_calc()`	$R_{\rm opt}^{\rm (calc)}$
`int optimalR_type()`	result of set optimalR working point. 1 = pass, 0 = fail
`double nsub_unfiltered(int order, fastjet::contrib::Njettiness::AxesMode axes = fastjet::contrib::Njettiness::kt_axes, double beta = 1., double R0 = 1.);`	N-subjettiness τ_i for the unfiltered fat jet
`double nsub_filtered(int order, fastjet::contrib::Njettiness::AxesMode axes = fastjet::contrib::Njettiness::kt_axes, double beta = 1., double R0 = 1.);`	N-subjettiness $\tau_i^{\rm (filt)}$ for the fat jet after filtering depending on `optimalR_type()`.
`double q_weight()`	weight of used QJETS history

Table 6. Functions to retrieve results of the HEPTOPTAGGER algorithm.

JHEP06(2015)203

name	default	description
`set_iterations(unsigned)`	100	number of QJETS iterations
`set_q_zcut(double)`	0.1	z_{cut} for pruning in QJETS
`set_q_dcut_fctr(double)`	0.5	D_{cut} factor for pruning in QJETS
`set_q_exp(double a, double b)`	0., 0. (C/A)	set distance measure for QJETS $d_{ij} = \min(p_{T,i}, p_{T,j})^a \max(p_{T,i}, p_{T,j})^b R_{ij}^2$
`set_q_rigidity(double)`	0.1	rigidity α for QJETS
`set_q_truncation_fctr(double)`	0.	threshold for merging probability ω_{ij} in QJETS

Table 7. Parameters of the QJETS frame for the HEPTOPTAGGER. All functions have a return type of `void`.

name	description
`HEPTopTagger leading()`	HEPTopTagger with leading tagged history
`HEPTopTagger subleading()`	HEPTopTagger with subleading tagged history
`double weight_leading()`	QJETS weight of the leading tagged history
`double weight_subleading()`	QJETS weight of the subleading tagged history
`double eps_q()`	$\varepsilon_{\text{Qjets}}$
`double m_mean()`	$\langle m \rangle$ for the tagged histories
`double m2_mean()`	$\langle m^2 \rangle$ for the tagged histories

Table 8. Functions to retrieve results of the QJETS frame.

name	description
`double U(unsigned)`	FWM of given order with unit weight
`double Pt(unsigned, fastjet::PseudoJet=`$(0., 0., 1., 0.)$`)`	FWM of given order with p_T weight relative to the given reference vector.

Table 9. Functions to retrieve Fox-Wolfram moments.

JHEP06(2015)203

References

[1] D.E. Morrissey, T. Plehn and T.M.P. Tait, *Physics searches at the LHC*, *Phys. Rept.* **515** (2012) 1 [arXiv:0912.3259] [INSPIRE].

[2] K. Agashe, A. Belyaev, T. Krupovnickas, G. Perez and J. Virzi, *LHC Signals from Warped Extra Dimensions*, *Phys. Rev.* **D 77** (2008) 015003 [hep-ph/0612015] [INSPIRE].

[3] V. Barger, T. Han and D.G.E. Walker, *Top Quark Pairs at High Invariant Mass: A Model-Independent Discriminator of New Physics at the LHC*, *Phys. Rev. Lett.* **100** (2008) 031801 [hep-ph/0612016] [INSPIRE].

[4] B. Lillie, L. Randall and L.-T. Wang, *The Bulk RS KK-gluon at the LHC*, *JHEP* **09** (2007) 074 [hep-ph/0701166] [INSPIRE].

[5] U. Baur and L.H. Orr, *High p_T Top Quarks at the Large Hadron Collider*, *Phys. Rev.* **D 76** (2007) 094012 [arXiv:0707.2066] [INSPIRE].

[6] U. Baur and L.H. Orr, *Searching for $t\bar{t}$ Resonances at the Large Hadron Collider*, *Phys. Rev.* **D 77** (2008) 114001 [arXiv:0803.1160] [INSPIRE].

[7] P. Fileviez Perez, R. Gavin, T. McElmurry and F. Petriello, *Grand Unification and Light Color-Octet Scalars at the LHC*, *Phys. Rev.* **D 78** (2008) 115017 [arXiv:0809.2106] [INSPIRE].

[8] ATLAS collaboration, *Search for resonances decaying into top-quark pairs using fully hadronic decays in pp collisions with ATLAS at $\sqrt{s} = 7$ TeV*, *JHEP* **01** (2013) 116 [arXiv:1211.2202] [INSPIRE].

[9] ATLAS collaboration, *Performance of boosted top quark identification in 2012 ATLAS data*, ATLAS-CONF-2013-084 (2013).

[10] ATLAS collaboration, *Search for Resonances Decaying into Top Quark Pairs Using Fully Hadronic Decays in pp Collisions with ATLAS at $\sqrt{s} = 7$ TeV*, ATLAS-CONF-2012-102 (2012).

[11] G Piacquadio *Identification of b-jets and investigation of the discovery potential of a Higgs boson in the $WH \to \ell\nu b\bar{b}$ channel with the ATLAS experiment*, Ph.D. Thesis, Universität Freiburg, Freiburg, Germany, CERN-THESIS-2010-027 (2010).

[12] G. Kasieczka, *Search for Resonances Decaying into Top Quark Pairs Using Fully Hadronic Decays in pp Collisions with ATLAS at $\sqrt{s} = 7$ TeV*, Ph.D. Thesis, Universisität Heidelberg, Heidelberg, Germany (2013), http://www.ub.uni-heidelberg.de/archiv/14941.

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

[14] T. Plehn, M. Spannowsky, M. Takeuchi and D. Zerwas, *Stop Reconstruction with Tagged Tops*, *JHEP* **10** (2010) 078 [arXiv:1006.2833] [INSPIRE].

[15] http://www.thphys.uni-heidelberg.de/~plehn/index.php?show=heptoptagger&visible=tools.

[16] L.G. Almeida, S.J. Lee, G. Perez, I. Sung and J. Virzi, *Top Jets at the LHC*, *Phys. Rev.* **D 79** (2009) 074012 [arXiv:0810.0934] [INSPIRE].

JHEP06(2015)203

[17] L.G. Almeida, S.J. Lee, G. Perez, G.F. Sterman, I. Sung and J. Virzi, *Substructure of high-p_T Jets at the LHC*, *Phys. Rev.* **D 79** (2009) 074017 [arXiv:0807.0234] [INSPIRE].

[18] L.G. Almeida, S.J. Lee, G. Perez, G. Sterman and I. Sung, *Template Overlap Method for Massive Jets*, *Phys. Rev.* **D 82** (2010) 054034 [arXiv:1006.2035] [INSPIRE].

[19] M. Backović and J. Juknevich, *TemplateTagger v1.0.0: A Template Matching Tool for Jet Substructure*, *Comput. Phys. Commun.* **185** (2014) 1322 [arXiv:1212.2978] [INSPIRE].

[20] 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].

[21] T. Plehn and M. Spannowsky, *Top Tagging*, *J. Phys.* **G 39** (2012) 083001 [arXiv:1112.4441] [INSPIRE].

[22] 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].

[23] S. Schaetzel, *Boosted Top Quarks and Jet Structure*, arXiv:1403.5176 [INSPIRE].

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

[25] 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].

[26] W. Skiba and D. Tucker-Smith, *Using jet mass to discover vector quarks at the LHC*, *Phys. Rev.* **D 75** (2007) 115010 [hep-ph/0701247] [INSPIRE].

[27] B. Holdom, *t-prime at the LHC: The physics of discovery*, *JHEP* **03** (2007) 063 [hep-ph/0702037] [INSPIRE].

[28] M. Gerbush, T.J. Khoo, D.J. Phalen, A. Pierce and D. Tucker-Smith, *Color-octet scalars at the CERN LHC*, *Phys. Rev.* **D 77** (2008) 095003 [arXiv:0710.3133] [INSPIRE].

[29] G. Brooijmans, *High p_T Hadronic Top Quark Identification. Part I: Jet mass and YSplitter*, ATL-PHYS-CONF-2008-008 (2008).

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

[31] 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].

[32] 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].

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

[34] M. Stoll, *Vetoed jet clustering: The mass-jump algorithm*, *JHEP* **04** (2015) 111 [arXiv:1410.4637] [INSPIRE].

[35] C. Anders, C. Bernaciak, G. Kasieczka, T. Plehn and T. Schell, *Benchmarking an even better top tagger algorithm*, *Phys. Rev.* **D 89** (2014) 074047 [arXiv:1312.1504] [INSPIRE].

JHEP06(2015)203

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

[37] D.E. Soper and M. Spannowsky, *Finding physics signals with event deconstruction*, *Phys. Rev.* **D 89** (2014) 094005 [arXiv:1402.1189] [INSPIRE].

[38] M.R. Buckley, T. Plehn and M. Takeuchi, *Buckets of Tops*, *JHEP* **08** (2013) 086 [arXiv:1302.6238] [INSPIRE].

[39] M.R. Buckley, T. Plehn, T. Schell and M. Takeuchi, *Buckets of Higgs and Tops*, *JHEP* **02** (2014) 130 [arXiv:1310.6034] [INSPIRE].

[40] T. Plehn and M. Takeuchi, *W+Jets at CDF: Evidence for Top Quarks*, *J. Phys.* **G 38** (2011) 095006 [arXiv:1104.4087] [INSPIRE].

[41] T. Plehn, M. Spannowsky and M. Takeuchi, *How to Improve Top Tagging*, *Phys. Rev.* **D 85** (2012) 034029 [arXiv:1111.5034] [INSPIRE].

[42] 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].

[43] 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].

[44] C.K. Vermilion, *Jet Substructure at the Large Hadron Collider: Harder, Better, Faster, Stronger*, arXiv:1101.1335 [INSPIRE].

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

[46] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross-sections in deep inelastic scattering*, hep-ph/9907280 [INSPIRE].

[47] S. Catani, Y.L. Dokshitzer, M. Olsson, G. Turnock and B.R. Webber, *New clustering algorithm for multi-jet cross-sections in e^+e^- annihilation*, *Phys. Lett.* **B 269** (1991) 432 [INSPIRE].

[48] 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].

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

[50] 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].

[51] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J.* **C 72** (2012) 1896 [arXiv:1111.6097] [INSPIRE], http://fastjet.fr.

[52] G.C. Fox and S. Wolfram, *Observables for the Analysis of Event Shapes in e^+e^- Annihilation and Other Processes*, *Phys. Rev. Lett.* **41** (1978) 1581 [INSPIRE].

[53] C. Bernaciak, M.S.A. Buschmann, A. Butter and T. Plehn, *Fox-Wolfram Moments in Higgs Physics*, *Phys. Rev.* **D 87** (2013) 073014 [arXiv:1212.4436] [INSPIRE].

[54] C. Bernaciak, B. Mellado, T. Plehn, P. Schichtel and X. Ruan, *Improving Higgs plus Jets analyses through Fox-Wolfram Moments*, *Phys. Rev.* **D 89** (2014) 053006 [arXiv:1311.5891] [INSPIRE].

JHEP06(2015)203

[55] S.D. Ellis, A. Hornig, T.S. Roy, D. Krohn and M.D. Schwartz, *Qjets: A Non-Deterministic Approach to Tree-Based Jet Substructure*, *Phys. Rev. Lett.* **108** (2012) 182003 [arXiv:1201.1914] [INSPIRE].

[56] S.D. Ellis, A. Hornig, D. Krohn and T.S. Roy, *On Statistical Aspects of Qjets*, *JHEP* **01** (2015) 022 [arXiv:1409.6785] [INSPIRE].

[57] Y.-T. Chien, *Telescoping jets: Probing hadronic event structure with multiple R's*, *Phys. Rev.* **D 90** (2014) 054008 [arXiv:1304.5240] [INSPIRE].

[58] F. Kling, T. Plehn and M. Takeuchi, *Tagging single Tops*, *Phys. Rev.* **D 86** (2012) 094029 [arXiv:1207.4787] [INSPIRE].

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

[60] A. Höcker et al., *TMVA — Toolkit for Multivariate Data Analysis*, PoS(ACAT)040 [physics/0703039] [INSPIRE].

[61] P. Speckmayer, A. Höcker, J. Stelzer and H. Voss, *The toolkit for multivariate data analysis, TMVA 4*, *J. Phys. Conf. Ser.* **219** (2010) 032057 [INSPIRE], http://tmva.sourceforge.net.

[62] D. Krohn, J. Thaler and L.-T. Wang, *Jets with Variable R*, *JHEP* **06** (2009) 059 [arXiv:0903.0392] [INSPIRE].

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

[64] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, *JHEP* **02** (2012) 093 [arXiv:1108.2701] [INSPIRE].

[65] I.W. Stewart, F.J. Tackmann and W.J. Waalewijn, *N-Jettiness: An Inclusive Event Shape to Veto Jets*, *Phys. Rev. Lett.* **105** (2010) 092002 [arXiv:1004.2489] [INSPIRE].

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

[67] A.J. Larkoski, I. Moult and D. Neill, *Building a Better Boosted Top Tagger*, *Phys. Rev.* **D 91** (2015) 034035 [arXiv:1411.0665] [INSPIRE].

[68] M. Freytsis, T. Volansky and J.R. Walsh, *Tagging Partially Reconstructed Objects with Jet Substructure*, arXiv:1412.7540 [INSPIRE].

[69] T. Strebler, *Substructure studies in preparation for a measurement of ttH*, MSc Thesis, ETH Zürich, Switzerland (2014).

[70] C. Englert, T. Plehn, P. Schichtel and S. Schumann, *Jets plus Missing Energy with an Autofocus*, *Phys. Rev.* **D 83** (2011) 095009 [arXiv:1102.4615] [INSPIRE].

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

[72] E. Gerwick, T. Plehn, S. Schumann and P. Schichtel, *Scaling Patterns for QCD Jets*, *JHEP* **10** (2012) 162 [arXiv:1208.3676] [INSPIRE].

[73] E. Gerwick, S. Schumann, B. Gripaios and B. Webber, *QCD Jet Rates with the Inclusive Generalized kt Algorithms*, *JHEP* **04** (2013) 089 [arXiv:1212.5235] [INSPIRE].

[74] Z. Bern et al., *Next-to-Leading Order $W + 5$-Jet Production at the LHC*, *Phys. Rev.* **D 88** (2013) 014025 [arXiv:1304.1253] [INSPIRE].

JHEP06(2015)203

[75] K. Kondo, *Dynamical Likelihood Method for Reconstruction of Events With Missing Momentum. 1: Method and Toy Models*, *J. Phys. Soc. Jap.* **57** (1988) 4126 [INSPIRE].

[76] D. Atwood and A. Soni, *Analysis for magnetic moment and electric dipole moment form-factors of the top quark via $e^+e^- \to t\bar{t}$*, *Phys. Rev.* **D 45** (1992) 2405 [INSPIRE].

[77] D0 collaboration, V.M. Abazov et al., *Helicity of the W boson in lepton + jets $t\bar{t}$ events*, *Phys. Lett.* **B 617** (2005) 1 [hep-ex/0404040] [INSPIRE].

[78] CDF collaboration, A. Abulencia et al., *Measurement of the top quark mass with the dynamical likelihood method using lepton plus jets events with b-tags in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$-TeV.*, *Phys. Rev.* **D 73** (2006) 092002 [hep-ex/0512009] [INSPIRE].

[79] D0 collaboration, V.M. Abazov et al., *A precision measurement of the mass of the top quark*, *Nature* **429** (2004) 638 [hep-ex/0406031] [INSPIRE].

[80] http://fastjet.hepforge.org/contrib/.

[81] http://www.gnu.org/software/gsl/.

JHEP06(2015)203