Eur. Phys. J. C manuscript No.
(will be inserted by the editor)

Towards an Understanding of the Correlations Within Jet Substructure

Report of BOOST2013, hosted by the University of Arizona, 12^{th}-16^{th} of August 2013.

D. Adams[1], A. Arce[2], L. Asquith[3], M. Backovic[4], T. Barillari[5], P. Berta[6], D. Bertolini[2], A. Buckley[8], J. Butterworth[9], R. C. Camacho Toro[10], J. Caudron[9], Y.-T. Chien[11], J. Cogan[12], B. Cooper[9], D. Curtin[17], C. Debenedetti[18], J. Dolen[9], M. Eklund[22], S. El Hedri[22], S. D. Ellis[22], T. Embry[22], D. Ferencek[23], J. Ferrando[24], S. Fleischmann[16], M. Freytsis[25], M. Giulini[21], Z. Han[27], D. Hare[4], P. Harris[4], A. Hinzmann[4], R. Hoing[4], A. Hornig[22], M. Jankowiak[4], K. Johns[28], G. Kasieczka[23], T. Knight[24], G. Kasieczka[29], R. Kogler[30], W. Lampl[4], A. J. Larkoski[4], C. Lee[31], R. Leone[31], P. Loch[31], D. Lopez Mateos[27], H. K. Lou[27], M. Low[27], P. Maksimovic[32], I. Marchesini[32], S. Marzani[32], L. Masetti[33], R. McCarthy[32], S. Menke[32], D. W. Miller[35], K. Mishra[36], B. Nachman[32], P. Nef[4], F. T. O'Grady[24], A. Ovcharova[23], A. Picazio[37], C. Pollard[38], B. Potter Landua[29], C. Potter[29], S. Rappoccio[39], J. Rutherfoord[40], G. P. Salam[10,11], J. Schabinger[23], A. Schwartzman[4], M. D. Schwartz[27], B. Shuve[43], P. Sinervo[44], D. Soper[45], D. E. Sosa Corral[45], M. Spannowsky[32], E. Strauss[34], M. Swiatlowski[4], J. Thaler[34], C. Thomas[34], E. Thompson[1], N. V. Tran[36], J. Tseng[36], E. Usai[36], L. Valery[36], J. Veatch[23], M. Vos[23], W. Waalewijn[4], and C. Young[47]

[1] *Columbia University, Nevis Laboratory, Irvington, NY 10533, USA*
[2] *Duke University, Durham, NC 27708, USA*
[3] *Argonne National Laboratory, Lemont, IL 60439, USA*
[4] *SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[5] *Deutsches Elektronen-Synchrotron, DESY, D-15738 Zeuthen, Germany*
[6] *Cornell University, Ithaca, NY 14853, USA*
[7] *Lund University, Lund, SE 22100, Sweden*
[8] *University of Edinburgh, EH9 3JZ, UK*
[9] *University College London, WC1E 6BT, UK*
[10] *LPTHE, UPMC Univ. Paris 6 and CNRS UMR 7589, Paris, France*
[11] *CERN, CH-1211 Geneva 23, Switzerland*
[12] *CAFPE and U. of Granada, Granada, E-18071, Spain*
[13] *McGill University, Montreal, Quebec H3A 2T8, Canada*
[14] *Iowa State University, Ames, Iowa 50011, USA*
[15] *Rutgers University, Piscataway, NJ 08854, USA*
[16] *Bergische Universitaet Wuppertal, Wuppertal, D-42097, Germany*
[17] *YITP, Stony Brook University, Stony Brook, NY 11794-3840, USA*
[18] *University of Manchester, Manchester, M13 9PL, UK*
[19] *UNESP - Universidade Estadual Paulista, Sao Paulo, 01140-070, Brazil*
[20] *INFN and University of Naples, IT80216, Italy*
[21] *University of Geneva, CH-1211 Geneva 4, Switzerland*
[22] *University of Washington, Seattle, WA 98195, USA*
[23] *Instituto de Física Corpuscular, IFIC/CSIC-UVEG, E-46071 Valencia, Spain*
[24] *University of Glasgow, Glasgow, G12 8QQ, UK*
[25] *Berkeley National Laboratory, University of California, Berkeley, CA 94720, USA*
[26] *Universidad de Buenos Aires, AR-1428, Argentina*
[27] *Harvard University, Cambridge, MA 02138, USA*
[28] *Weizmann Institute, 76100 Rehovot, Israel*
[29] *Universitaet Hamburg, DE-22761, Germany*
[30] *Universitaet Heidelberg, DE-69117, Germany*
[31] *University of Arizona, Tucson, AZ 85719, USA*
[32] *IPPP, University of Durham, Durham, DH1 3LE, UK*
[33] *Universitaet Mainz, DE 55099, Germany*
[34] *MIT, Cambridge, MA 02139, USA*
[35] *University of Chicago, IL 60637, USA*
[36] *Fermi National Accelerator Laboratory, Batavia, IL 60510, USA*
[37] *Indiana University, Bloomington, IN 47405, USA*
[38] *University of California, Davis, CA 95616, USA*
[39] *Johns Hopkins University, Baltimore, MD 21218, USA*
[40] *INFN and University of Pisa, Pisa, IT-56127, Italy*
[41] *Texas A & M University, College Station, TX 77843, USA*
[42] *INFN and University of Calabria, Rende, IT-87036, Italy*
[43] *Brown University, Richmond, RI 02912, USA*
[44] *Yale University, New Haven, CT 06511, USA*
[45] *CEA Saclay, Gif-sur-Yvette, FR-91191, France*
[46] *University of Illinois, Chicago, IL 60607, USA*
[47] *University of California, Berkeley, CA 94720, USA*

Abstract Abstract for BOOST2013 report

Keywords boosted objects · jet substructure · beyond-the-Standard-Model physics searches · Large Hadron Collider

1 Introduction

Jet substructure has been around a while now, and it's time to study the correlations between the plethora of observables that have been developed and used. Previous BOOST reports [1,2,3] studied some of these things.

2 Monte Carlo Samples

2.1 Quark/gluon and W taggging

Samples were generated the $\sqrt{s} = 8$ TeV for QCD dijets and W^+W^- pairs decaying hadronically off a (pseudo)scalar resonance. The QCD events were split into subsamples of gg and $q\bar{q}$ events, allowing for tests of both W and quark-gluon discrimination.

For events at $\sqrt{s} = 8$ TeV, individual quark and gluon samples were produced at leading order (LO) using MadGraph5, while W^+W^- samples were generated using the JHU Generator to allow for separation of longitudinal and transverse polarizations. Both were produced in p_T bins of 100 TeV. and generated using CTEQ6L1 PDFs. These were then showered using Pythia8 (version 8.176) using the default tune 4C.

The analysis relies on clustering with FastJet 3.03. Events were clustered using the anti-k_t algorithm with jet radii of $R = 0.4, 0.8, 1.2$. A cut on the jet p_T is once again applied after showering/clustering, to ensure similar p_T spectra for signal and background in each bin.

2.2 Top tagging

Samples were generated at $\sqrt{s} = 14$ TeV. Standard Model dijet and top pair samples were produced with Sherpa 2.0.0, with matrix elements with up to two extra partons matched to the shower. The top samples included only hadronic decays and were generated in exclusive p_T bins of width 100 GeV, taking as slicing parameter the maximum of the top/anti-top p_T. The QCD samples were generated with a cut on the leading parton-level jet p_T, where parton-level jets are clustered with the anti-k_t algorithm with jet radius $R = 1.2$. The matching scale is selected to be $Q_{\text{cut}} = 40$, 60, 80 GeV for the $p_{T\,\text{min}} = 600$, 1000, and 1500 GeV bins, respectively.

The analysis relies on FastJet 3.0.3 for jet clustering and calculation of jet substructure observables. Events were clustered using the anti-k_t algorithm with jet radii of $R = 0.4$, 0.8, and 1.2. A cut on the jet p_T is once again applied after showering/clustering, to ensure similar p_T spectra for signal and background in each bin.

3 Jet Algorithms and Grooming Approaches

Describe the jet algorithms and grooming approaches that we will use in the report. Give the nomenclature that we will use to refer to e.g. the groomed mass in the rest of the report.

4 Substructure Variables/Taggers

Describe the specific substructure variables and tagging approaches that we will be using in this report e.g. n-subjettiness, Q-jets, HTT, JH tagger. Give the nomenclature that we will use to refer to these variables/taggers in the rest of the report.

5 Quark-Gluon Discrimination

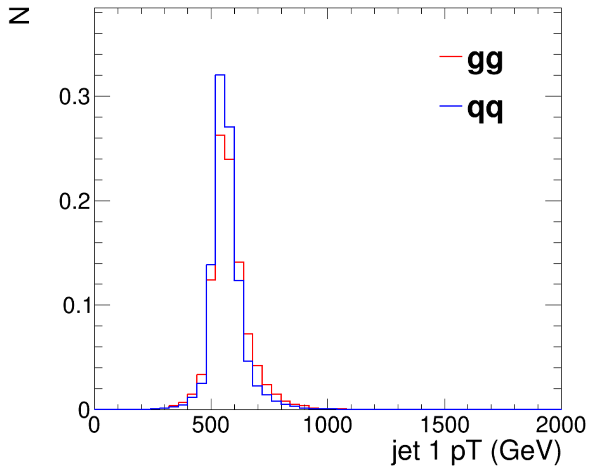
In this section we examine the differences between quark and gluon initiated jets in terms of the substructure variables, and to what extent these variables are correlated. Along the way, we attempt to provide some theoretical understanding of these observations. The motivation for these studies comes not only from the desire to "tag" a jet as being quark or gluon initiated, but also from the point of view of understanding the quark and gluon components to the QCD background to boosted boson and boosted top tagging.

5.1 Methodology

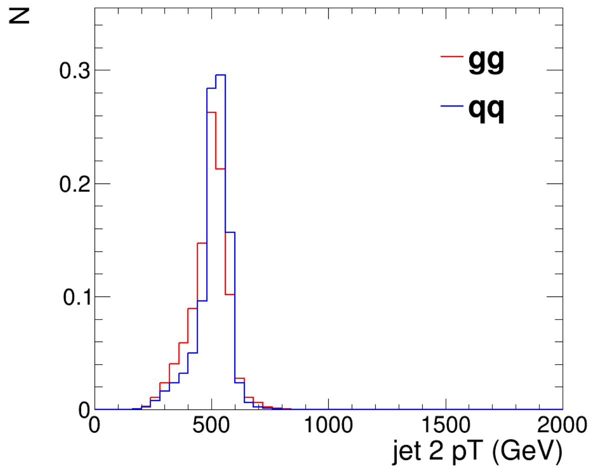
These studies use the qq and gg samples, described previously in Section 2.

Jets are reconstructed using the anti-k_{T} algorithm, and have various jet grooming approaches applied, as described in Section 3. The following event selection is then applied to these samples....(presumably this will vary depending on which kinematic bin is used, as will the actual samples used - maybe summarize in a table).

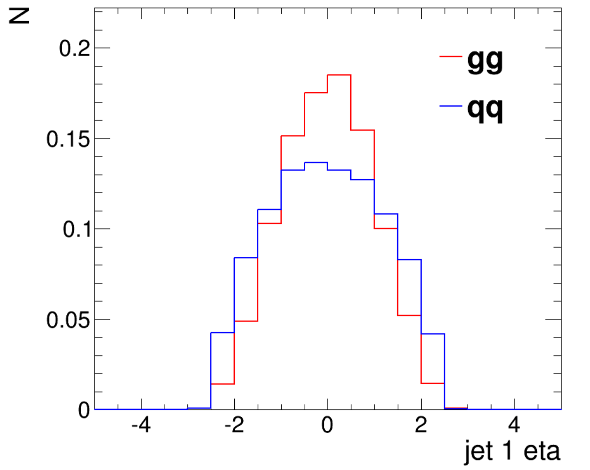
Go on to explain how we produce the ROC curves, how the BDT training is done etc.



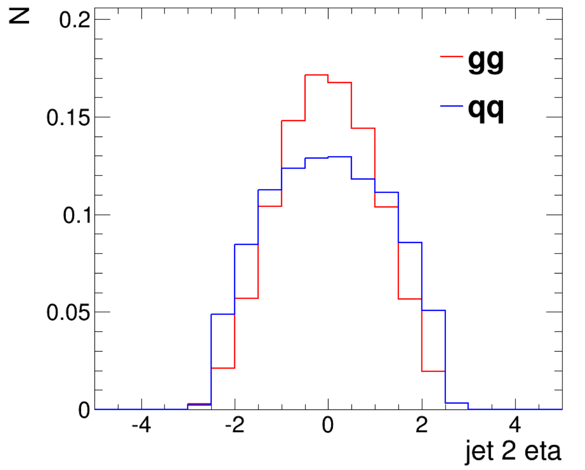
(a) Leading jet p_{T}



(b) Sub-leading jet p_{T}



(c) Leading jet η



(d) Sub-leading jet η

Fig. 1 Comparisons of quark and gluon distributions in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm: basic kinematic distributons.

Figure 1 shows a comparison of the quark and gluon samples in some basic kinematic distributions.

5.2 Single Variable Discrimination

Figure 2 the compares the quark and gluon samples in the mass distributions for the different groomers, and Figure 3 in the different substructure variables.

Figure 4 shows the single variable ROC curves in the p_{T} 500 GeV bin for the anti-k_{T} R=0.8 algorithm, compared to the ROC curve for a BDT combination of all the variables. Only the ungroomed mass is shown. One can see that the single most discriminant variables are $\mathrm{n}_{\mathrm{constits}}$ and $C_1^{\beta=0}$.

We want to look also at:

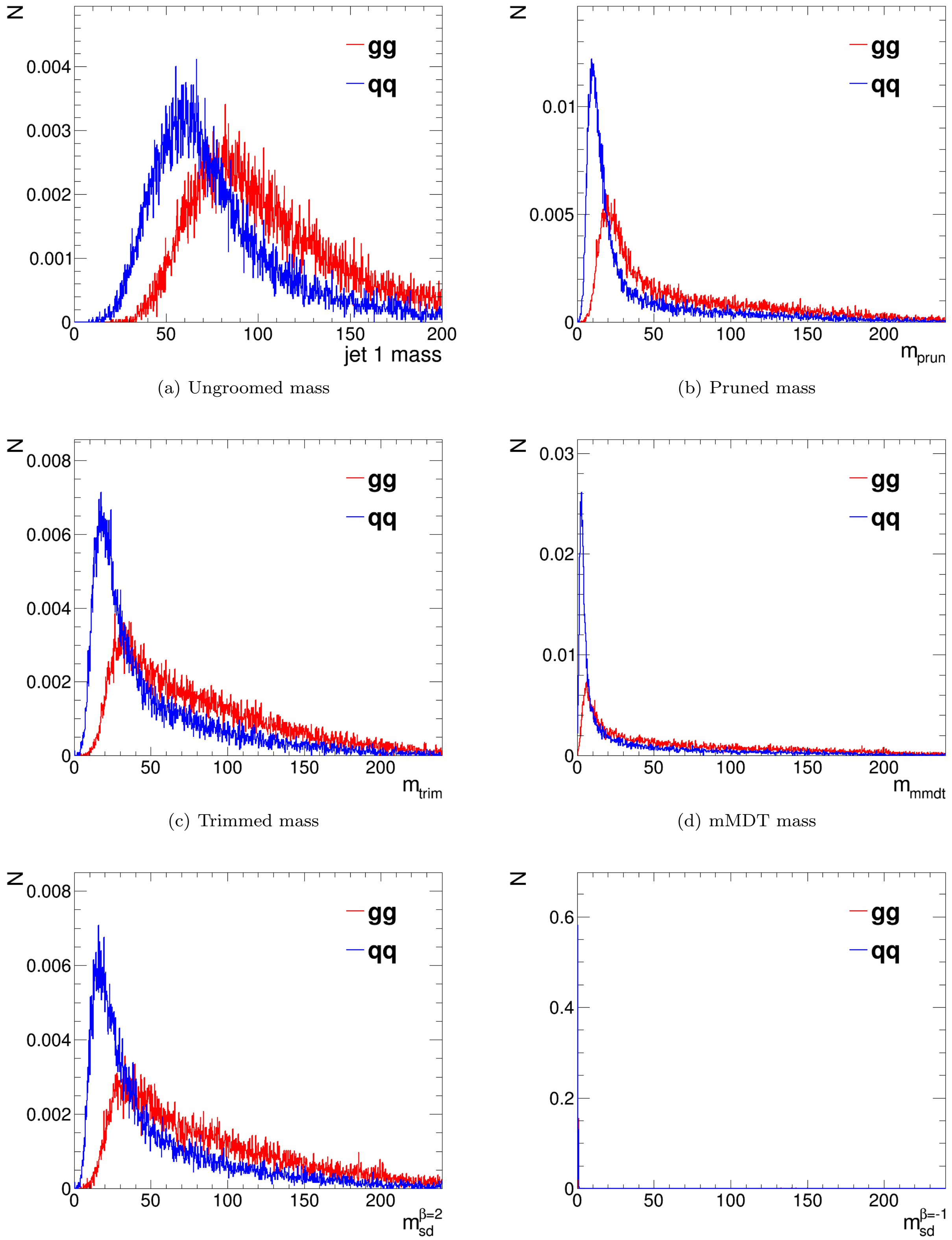
- *Dependence on R.*
- *Dependence on pT.*

5.3 Correlations

Put in 2-D plots of correlations between variables (see theory discussions below)

5.4 Combined Performance of Quark-Gluon Tagging

Put in ROC curves of BDT combination of variables



(a) Ungroomed mass

(b) Pruned mass

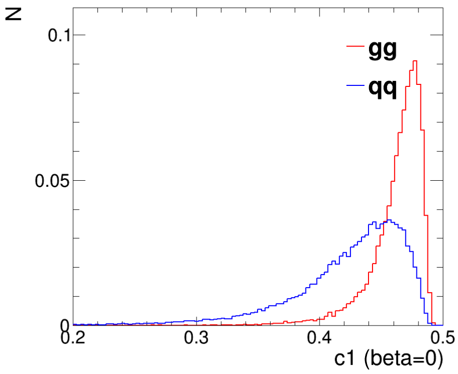
(c) Trimmed mass

(d) mMDT mass

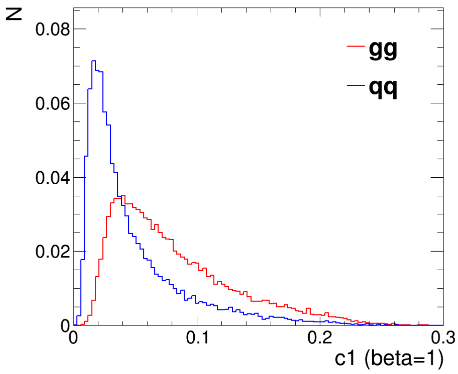
(e) Soft-drop $\beta = 2$ mass

(f) Soft-drop $\beta = -1$ mass

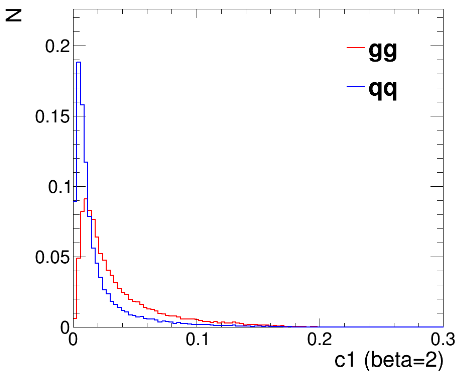
Fig. 2 Comparisons of quark and gluon distributions in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm: leading jet mass distributions.



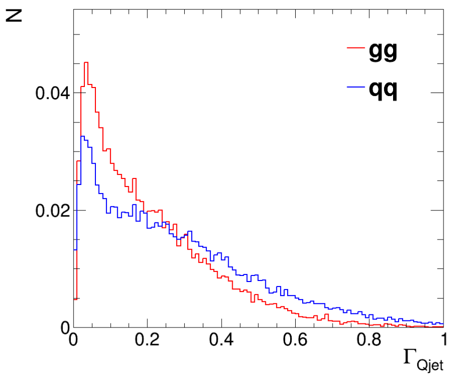
(a) $C_1^{\beta=0}$



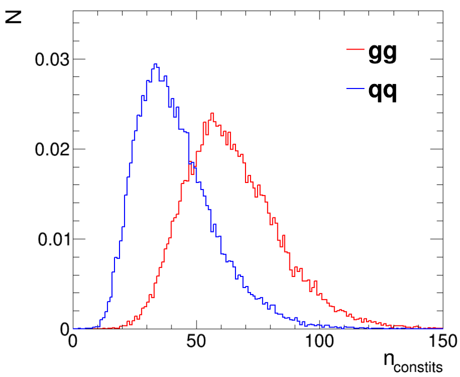
(b) $C_1^{\beta=1}$



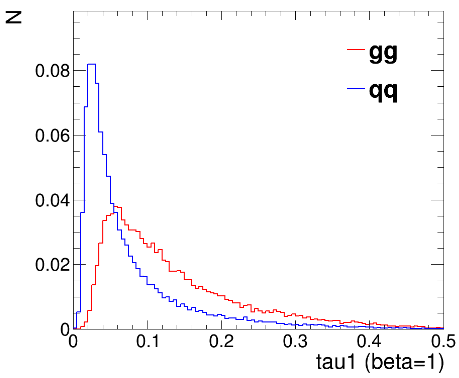
(c) $C_1^{\beta=2}$



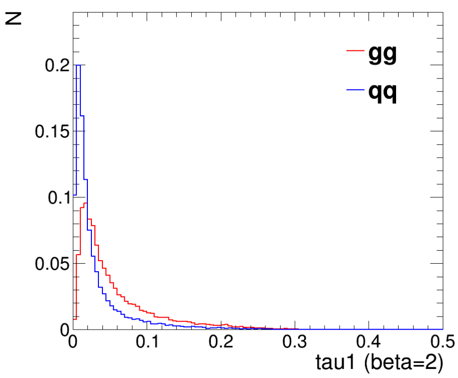
(d) Γ_{Qjet}



(e) $\mathrm{n_{constits}}$



(f) $\tau_1^{\beta=1}$



(g) $\tau_1^{\beta=2}$

Fig. 3 Comparisons of the quark and gluon distributions in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm: substructure variables.

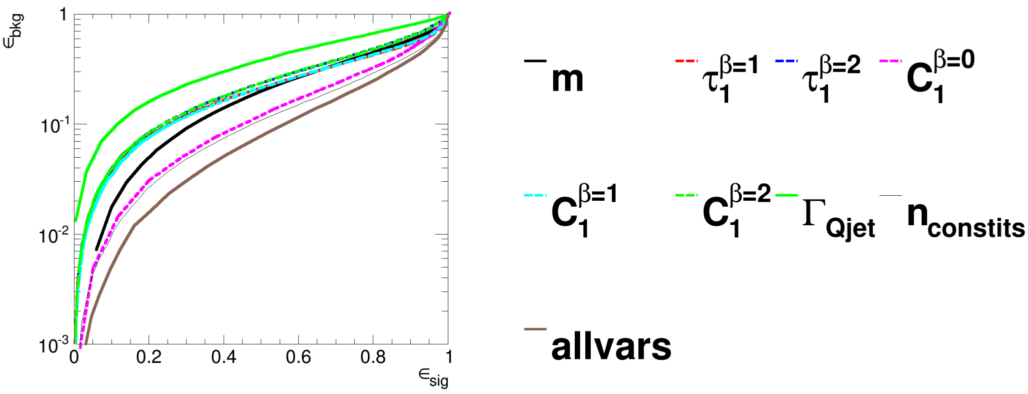


Fig. 4 The ROC curve for all single variables considered for quark-gluon discrimination in the p_T 500 GeV bin using the anti-k_T R=0.8 algorithm.

5.5 QJets Volatility and $p_T D$ ($C_1^{(\beta=0)}$)

Simple explanation of correlation, or why does combining volatility and $p_T D$ improve quark versus gluon discrimination. $p_T D$ ($C_1^{(\beta=0)}$) takes small (large) values for a jet with near-democratic energy sharing between particles and large (small) values when the energy of the jet is contained in a few particles. Because we expect gluons to radiate more particles, we expect that $p_T D_g < p_T D_q$ (or ${C_1^{(\beta=0)}}_g > {C_1^{(\beta=0)}}_q$). Now, we expect the volatility of gluon jets to be in general smaller than that of quark jets because there is a greater probability (by a factor of about $C_A/C_F = 9/4$) that there was a relatively hard emission in a jet that is not groomed away. By measuring both volatility and $p_T D$, we are sensitive to both regions of phase space: where a relatively hard emission dominates the mass of the jet as well as the region where many soft emissions set the jet mass.

The following is Steve's discussion of volatility difference between quarks and gluons:

Here is the (qualitative) thinking: typical QCD jet mass distributions look as illustrated on slide 17, although you should really be thinking in terms of plot versus m/p_T, since p_T is what sets the scale in the plot. Qualitatively there is a (very) large peak for $m/p_T \lesssim 0.1$ and you should think of these jets as having masses that arise from multiple soft emissions, some of which are at substantial angles. It is these components of the jet that are operated on by pruning (reducing the mass dramatically) and that yield the large volatility tail for QCD jets. For larger m/p_T values there is typically a shoulder (my description is clearest on a semi-log plot) that runs out to about $m/pT \sim 0.40.5$ (where the distribution decreases rapidly). These are the QCD jets (a small fraction of the total in a given p_T bin) that contain a hard, relatively large angle emission, which supplies the bulk of the jet mass. Such jets are effected only slightly by pruning and should exhibit much smaller volatility than the jets in the (smaller mass) peak region.

With that picture in mind and recalling that the size of the shoulder is given by low order perturbation theory (the probability of the one hard emission), we expect that the shoulder will be higher for gluons than for quarks (essentially by the usual C_A/C_F color charge factor), as suggested by the lower right plot on slide 17. Since the shoulder presumably plays a more important role for gluons (since it is larger), one would expect that the volatility distribution for gluons is narrower than quarks, as suggested in the upper left plot on slide 17. Am I making sense?

On the other hand, the volatility distribution plot indicates that the Q vs G distributions for your cuts are not really very different, which is presumably why it is not a very good discriminant by itself. But I expect this to depend it detail on where we are operating on the m/pT distributions. This leads to my request above. Your p_T bin is pretty broad and I dont expect the q and g samples to have the same shape within the bin. Of course, this may not be an issue, but I would like to check.

5.6 Comparison of Groomed Jet Masses

6 Boosted W-Tagging

In this section we study the performance of various groomed jet masses, substructure variables, and BDT combinations of groomed mass and substructure, in terms of the identification of a boosted hadronically decaying W signal aginst a gluon-gluon background. We produce

Receiver Operating Characteristic (ROC) curves that elucidate the performance of the various groomed mass and substructure variables that are capable of providing discrimination between signal and background. A range of different distance parameter settings for the anti-k_{T} jet algorithm are explored, in a variety of kinematic regimes (lead jet p_T 200-300 GeV, 500-600 GeV, 1.0-1.1 TeV), to explore the performance as a function of jet radius and jet boost, and to see where substructure approaches may break down. The groomed mass and substructure variables are then combined in a Boosted Decision Tree (BDT), and the performance of the resulting BDT discriminant explored through ROC curves to understand the degree to which variables are correlated and exploiting the same information, and how this changes with jet boost and jet radius.

6.1 Methodology

These studies use the $X \to WW$ samples as signal and the gg samples to model the QCD background, described previously in Section 2. Whilst only gluonic backgrounds are explored here, the conclusions as to the dependence of the performance and correlations on the jet boost and radius have been verified to hold also for qq backgrounds. *To be checked!*

Jets are reconstructed using the anti-k_{T} algorithm, and have various jet grooming approaches applied, as described in Section 3. The following event selection is then applied to these samples....(presumably this will vary depending on which kinematic bin is used, as will the actual samples used - maybe summarize in a table).

Figure 5 shows a comparison of the leading jet p_T for the signal and background in the p_T 300-400 GeV bin, for the two different anti-k_{T} jet algorithm distance parameters explored in this bin (R=0.8 and R=1.2). Figures 6 and 7 show the same for the p_T 500-600 GeV bin and p_T 1.0-1.1 TeV bin respectively, where for the p_T 1.0-1.1 TeV bin the distance parameter R=0.4 is also explored.

Go on to explain how we produce the ROC curves, how the BDT training is done etc.

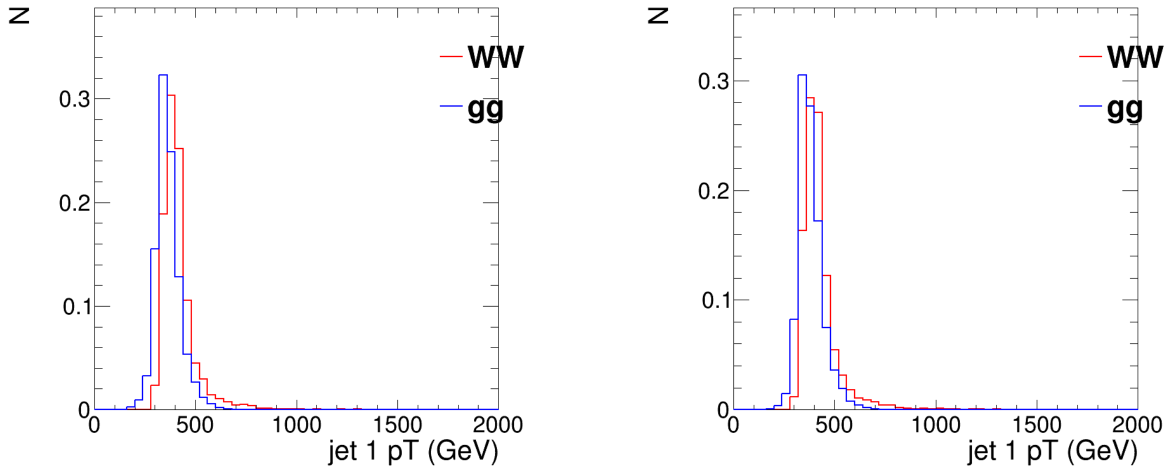
6.2 Single Variable Performance

In this section we will explore the performance of the various groomed jet mass and substructure variables in terms of discriminating signal and background, and how this performance changes depending on the kinematic bin and jet radius considered.

Figure 8 the compares the signal and background in terms of the different groomed masses explored for the anti-k_{T} R=0.8 algorithm in the p_T 500-600 bin. One can clearly see that in terms of separating signal and background the groomed masses will be significantly more performant than the ungroomed anti-k_{T} R=0.8 mass. *Need to comment on the soft drop B=-1 mass here* Figure 9 compares signal and background in the different substructure variables explored for the same jet radius and kinematic bin.

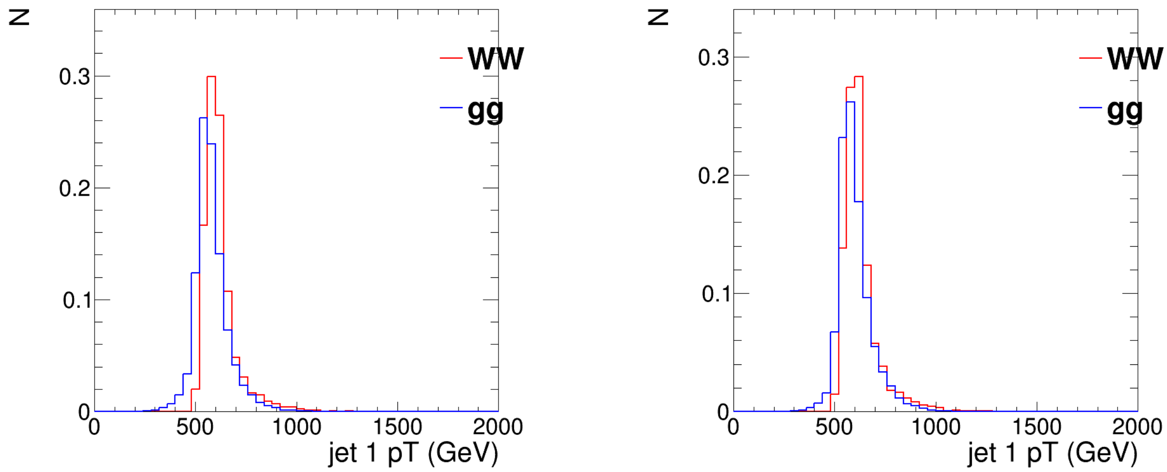
Figures 10,11 and 12 show the single variable ROC curves compared to the ROC curve for a BDT combination of all the variables (labelled “allvars”), for each of the anti-k_{T} distance parameters considered in each of the kinematic bins. One can see that, in all cases, the “allvars” option is considerably more performant than any of the individual single variables considered, indicating that there is considerable complimentarity between the variables, that will be explored further in the next section. The best performant individual variables for a reasonable signal efficiency are the groomed masses, which all have a similar level of performance that is superior to that of any of the substructure variables considered.

Although the ROC curves give all the relavant information, it is hard to compare performance quantitatively. In Figures 13,14 and 15 matrices are shown which give the background rejection for a signal efficiency of 50% when two variables (that on the x-axis and that on the y-axis) are combined in a BDT. Thus, the diagonal of these plots can be examined to see quantitatively the individual single variable performance. Because we have not attempted to optimise the grooming parameter settings of each grooming algorithm, we do not want to place too much emphasis here on the relative performance of the groomed masses, but instead look at the trends veresus p_T and R. One can see clearly that the background rejection power of the groomed mass variables increases as the p_T is increased. Within a p_T bin, one can also see that the groomed mass performance is rather invariant to changes in the jet radius. In contrast, the substructure variable performance varies considerably as the jet radius is changed. In general, the background rejection power of individual jet substructure variables gets worse as the jet radius is increased. The only exception to this is in the highest p_T bin, where the background rejection power of $C_2^{\beta=1}$ improves when going from jet radius R=0.4 to R=0.8, but then gets worse again as we go to R=1.2. *Insert some nice discussion/explanation of why jet substructure power generally gets worse as we go to large jet radius, but groomed mass performance does not*



(a) anti-k_{T} R=0.8 (b) anti-k_{T} R=1.2

Fig. 5 Comparisons of the leading jet p_T spectrum of the gg background to the WW signal in the p_T 300-400 GeV bin using the different anti-k_{T} jet distance parameters explored.



(a) anti-k_{T} R=0.8 (b) anti-k_{T} R=1.2

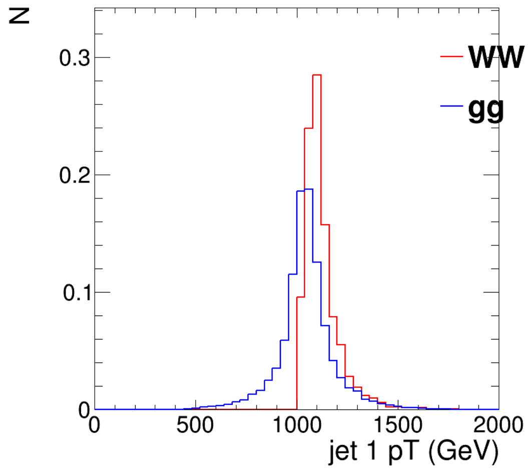
Fig. 6 Comparisons of the leading jet p_T spectrum of the gg background to the WW signal in the p_T 500-600 GeV bin using the different anti-k_{T} jet distance parameters explored.

6.3 Combined Performance

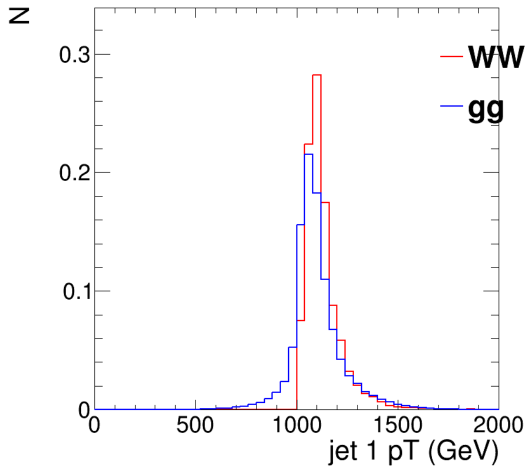
Mass + X Performance

Figure 16 shows the background efficiency for a fixed signal efficiency (50%) of each BDT combination of each pair of variables considered, in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm. One can see that the best background rejection is achieved using combinations of the groomed mass variables with other substructure variables (with the exception of the soft drop mass with $\beta = -1$). Combinations of the mass variables themselves are not particularly powerful, but are interesting for understanding the correlations between the masses (see Section 6.3). Equally, combination of the substructure variables, without using a mass, are not powerful.

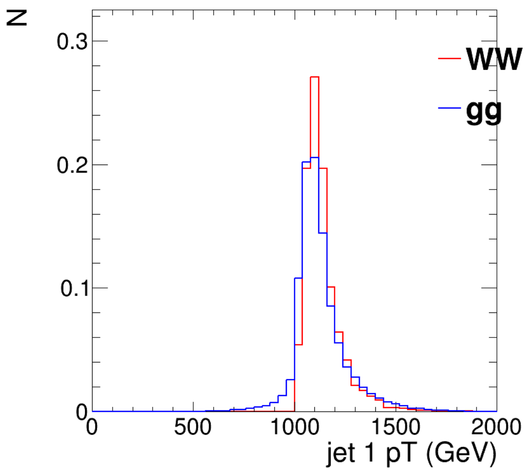
Figure 17 shows the actual ROC curves of the BDT combinations of each mass variable with every other variable considered in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm. *Can we drop the combinations of mass + mass from these plots to make them clearer? Also would be good to put the single variable mass curve on these plots, so you can see how much*



(a) anti-k_{T} R=0.4



(b) anti-k_{T} R=0.8



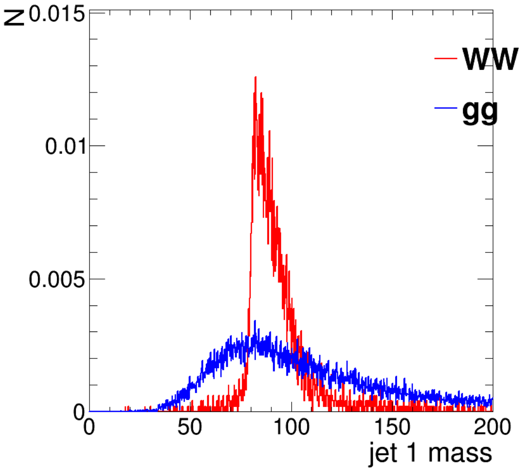
(c) anti-k_{T} R=1.2

Fig. 7 Comparisons of the leading jet p_T spectrum of the gg background to the WW signal in the p_T 1.0-1.1 TeV bin using the different anti-k_{T} jet distance parameters explored.

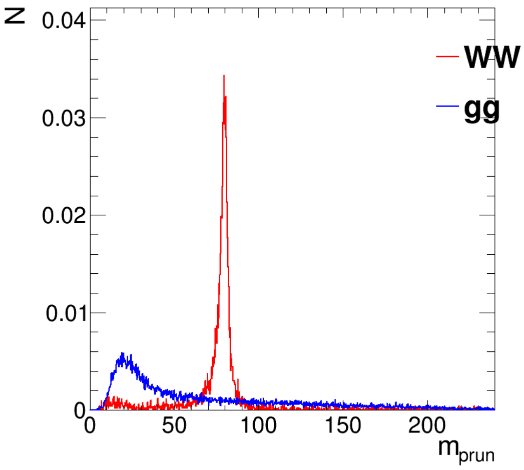
improvement the combination gives, and the "all variables" curve.

No combination with other variables can recover the poor performance of the ungroomed mass and the soft drop mass with $\beta = -1$. Figures 16 and 17 show that the other groomed/filtered masses are all most improved by combination with the $C_2^{\beta=1}$ energy correlation function. Figure 18 shows the 2-D correlation plots between the mMDT mass and the $C_2^{\beta=1}$, Γ_{Qjet} and $\tau_{21}^{\beta=1}$ variables. One can clearly see that there is substantially less correlation between the mass and $C_2^{\beta=1}$ than the other variables. Similar results are seen for the other groomed masses.

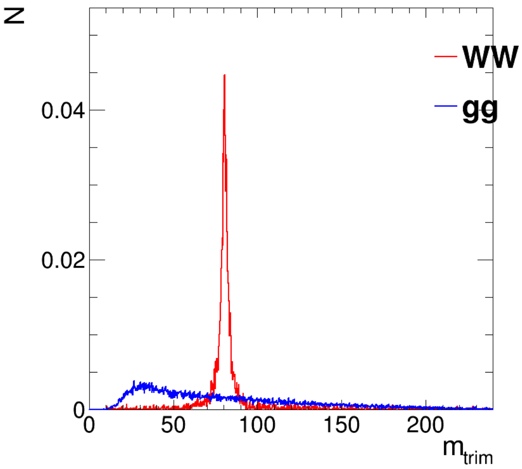
Figure 19 shows the background efficiency for a fixed signal efficiency (50%) of each BDT combination of each pair of variables considered, in the p_T 500 GeV bin, now using the anti-k_{T} R=1.2 algorithm. Compared to Figure 16, the overall trends are similar, but there are clear differences in the relative power of the mass + X combinations. Interestingly, the groomed masses are now all most improved by combination with the $\tau_{21}^{\beta=1}$ variable, in contrast with $C_2^{\beta=1}$ which performed best for the smaller radius of R=0.8. Figure 20 shows the actual ROC curves for the BDT combinations of the best performant groomed masses with every other variable considered in the p_T 500 GeV bin using the anti-k_{T} R=1.2 algorithm. One can see from Figure **??** that the single variable discrimination of $\tau_{21}^{\beta=1}$ and $C_2^{\beta=1}$ changes quite markedly when the distance parameter R is varied, although in both cases $C_2^{\beta=1}$ is a better single variable discriminant (except for very high signal efficiencies). Figure 21 shows how the actual distribu-



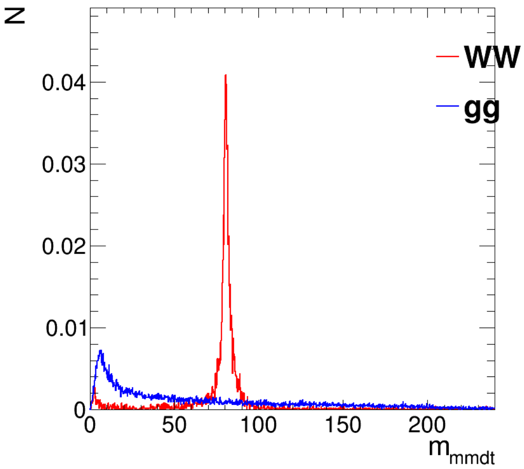
(a) Ungroomed mass



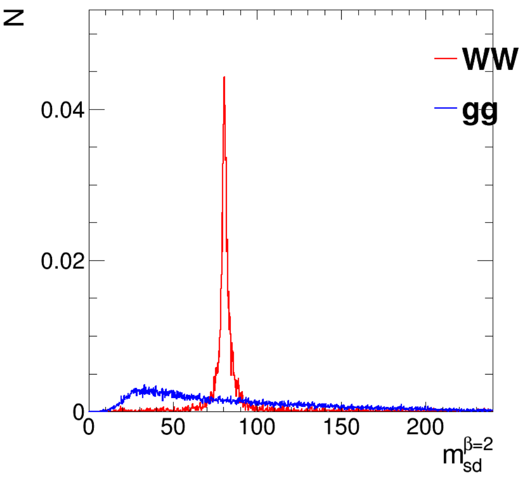
(b) Pruned mass



(c) Trimmed mass

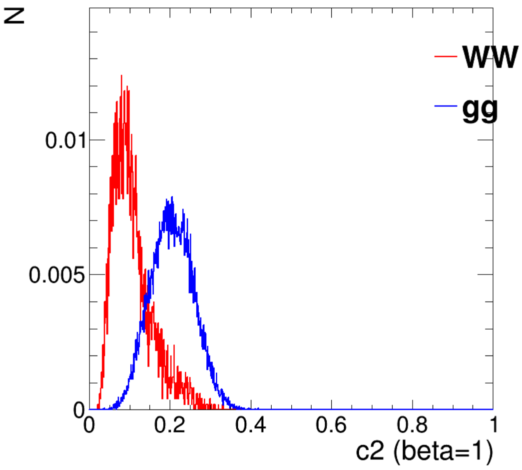


(d) mMDT mass

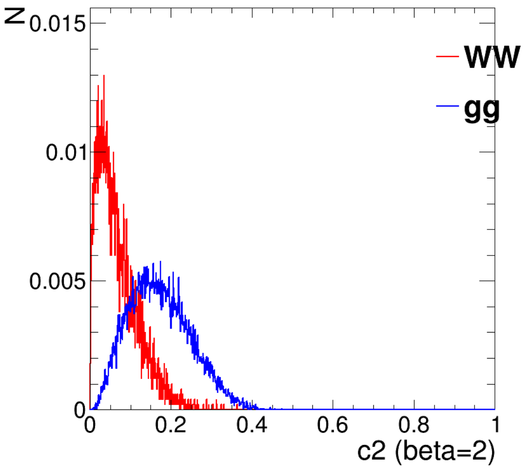


(e) Soft-drop $\beta = 2$ mass

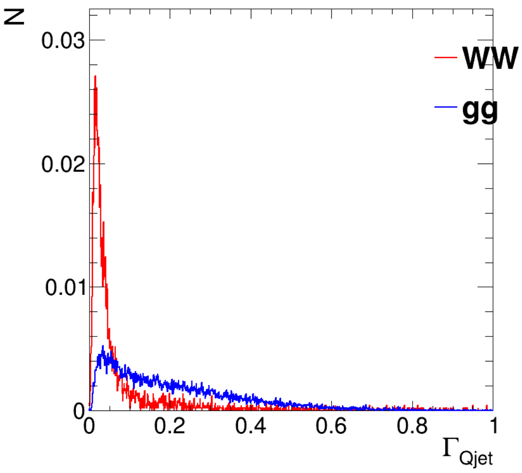
Fig. 8 Comparisons of the QCD background to the WW signal in the p_T 500-600 GeV bin using the anti-k_{T} R=0.8 algorithm: leading jet mass distributions.



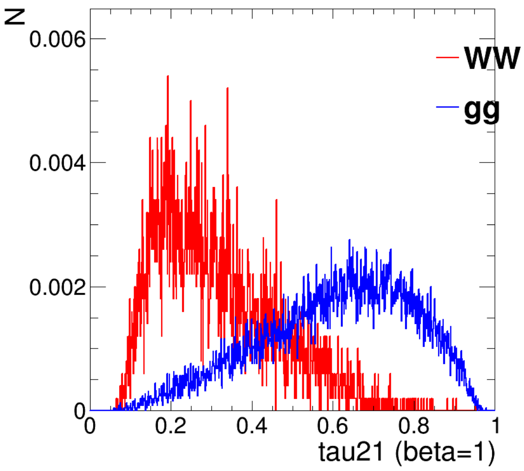
(a) $C_2^{\beta=1}$



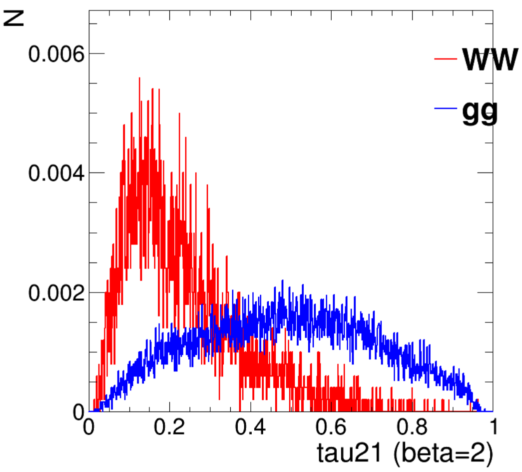
(b) $C_2^{\beta=2}$



(c) Γ_{Qjet}

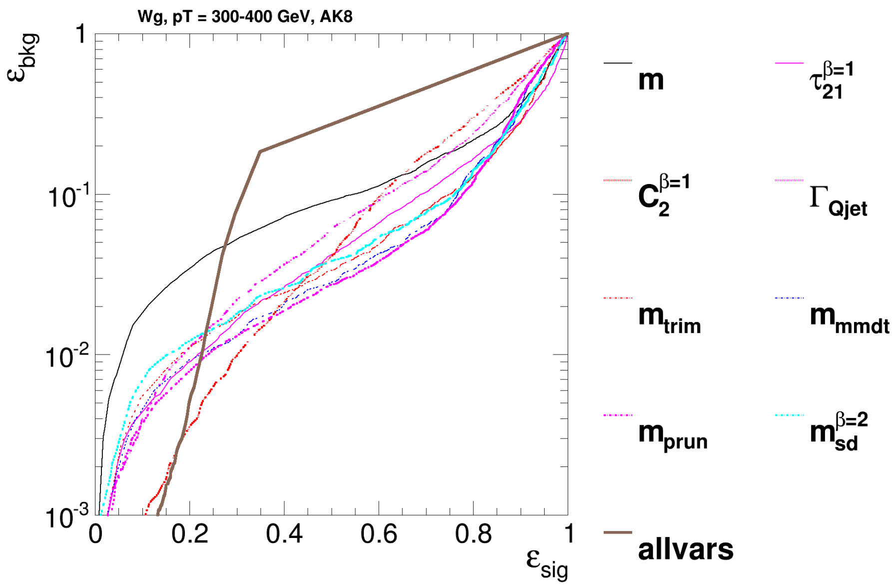


(d) $\tau_{21}^{\beta=1}$

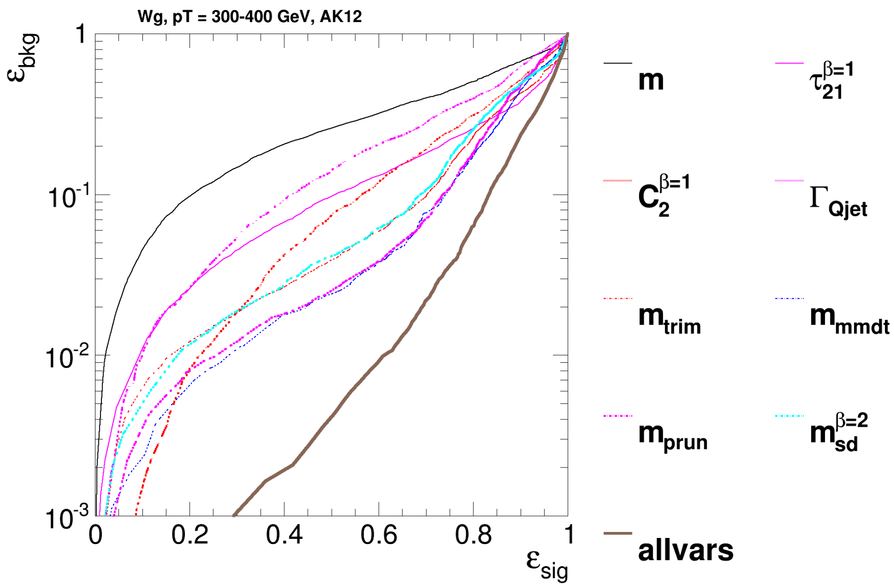


(e) $\tau_{21}^{\beta=2}$

Fig. 9 Comparisons of the QCD background to the WW signal in the p_T 500-600 GeV bin using the anti-k_{T} R=0.8 algorithm: substructure variables.



(a) anti-k_T R=0.8, p_T 300-400 GeV bin

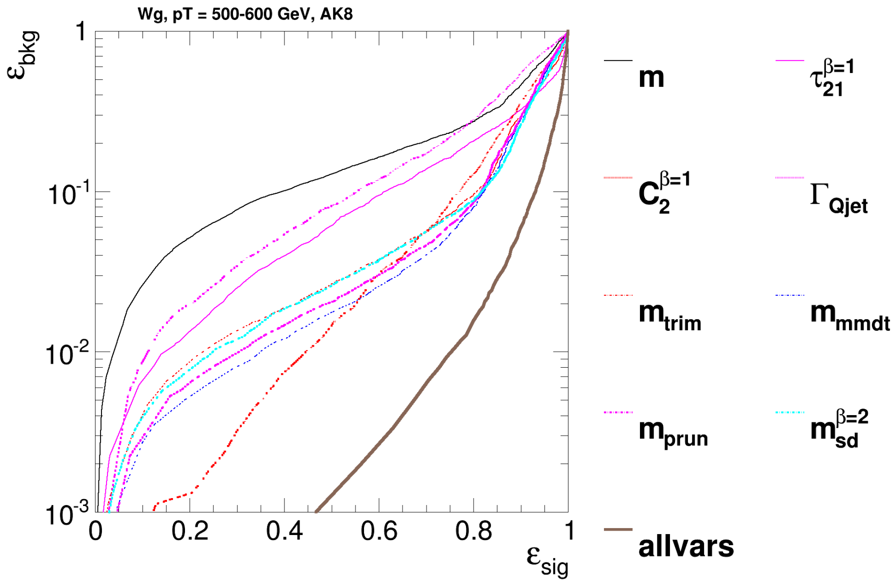


(b) anti-k_T R=1.2, p_T 300-400 GeV bin

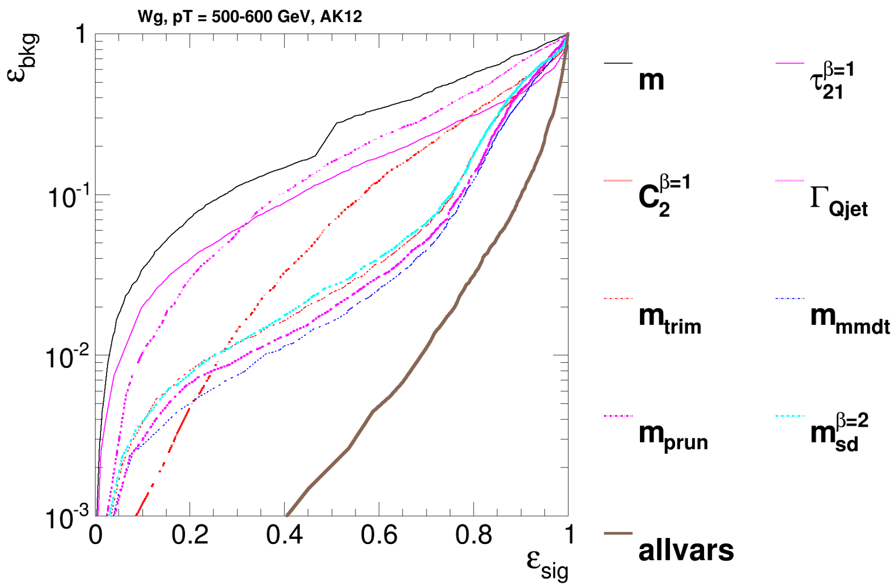
Fig. 10 The ROC curve for all single variables considered for W tagging in the p_T 300-400 GeV bin using the anti-k_T R=0.8 algorithm (top) and R=1.2 algorithm (bottom).

tions of the $C_2^{\beta=1}$ and $\tau_{21}^{\beta=1}$ change when we change the distance parameter. Figure 22 shows the 2-D correlation plots between between the mMDT mass and the $C_2^{\beta=1}$, Γ_{Qjet} and $\tau_{21}^{\beta=1}$ variables for the R=1.2 case. It is hard to see a substantial difference in the correlations here versus Figure 18, but perhaps $C_2^{\beta=1}$ is marginally more correlated with the mass for R=1.2 compared to R=0.8. *Andrew to add his explanation of why discrimination power of C2 versus tau21 gets worse when we go to larger jet radii (email 0606/2014)*

Now show a plot which compares on one plot the best combined performance for each groomed mass + X for both R=0.8 and 1.2 cases e.g. mass + $C_2^{\beta=1}$ *for R=0.8 and mass +* $\tau_{21}^{\beta=1}$ *for R=1.2, and draw on also the all variables curve for both R=0.8,1.2. Then we can see if there is much dependence on choice of mass once you combine with another variable, and compare directly the two distance parameters. This plot is just for one kinematic bin, we should make the same plot for others.*



(a) anti-k_T R=0.8, p_T 500-600 GeV bin



(b) anti-k_T R=1.2, p_T 500-600 GeV bin

Fig. 11 The ROC curve for all single variables considered for W tagging in the p_T 500-600 GeV bin using the anti-k_T R=0.8 algorithm (top) and R=1.2 algorithm (bottom).

Repeat these studies for different R and different kinematic bins. Finally make plots which compare best combined performance for different R and kinematics.

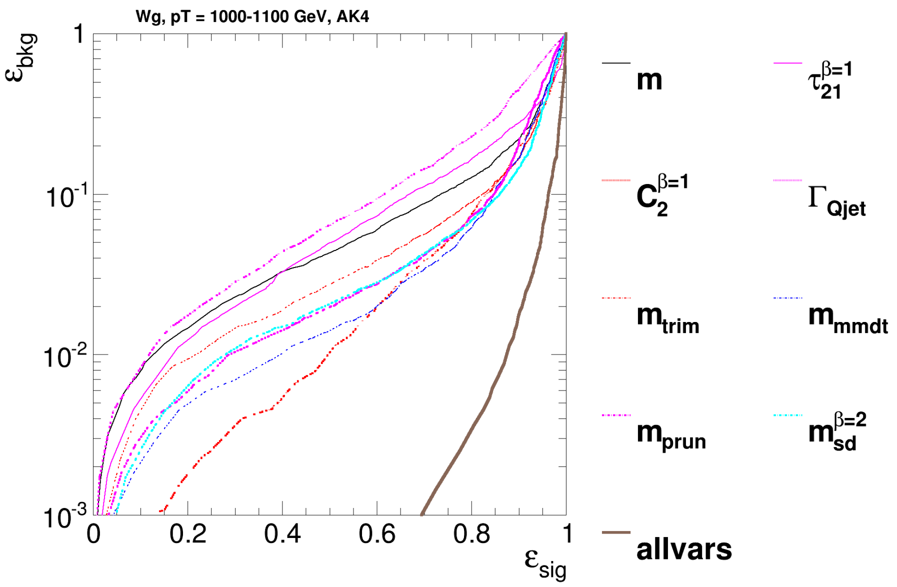
Do we want to look at other combinations of variables which don't involve mass? Practically I think we will always be making mass + X though.

Mass + Mass Performance

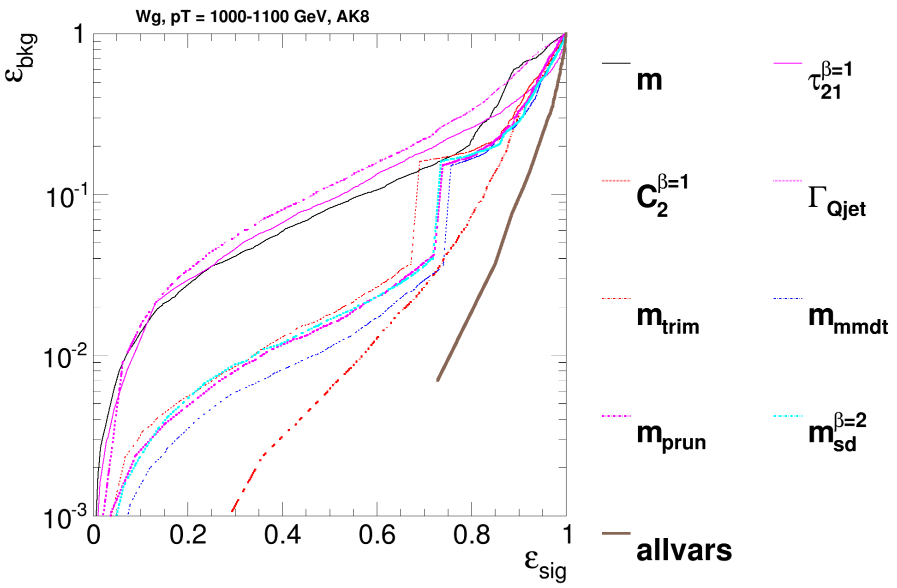
It's interesting also to study and understand how the different groomed masses relate to each other and how they are correlated.

Figures 23 and Figures 24 shows 2-D correlation plots of the different types of groomed mass in the p_T 500 GeV bin using the anti-k_T R=0.8 algorithm.

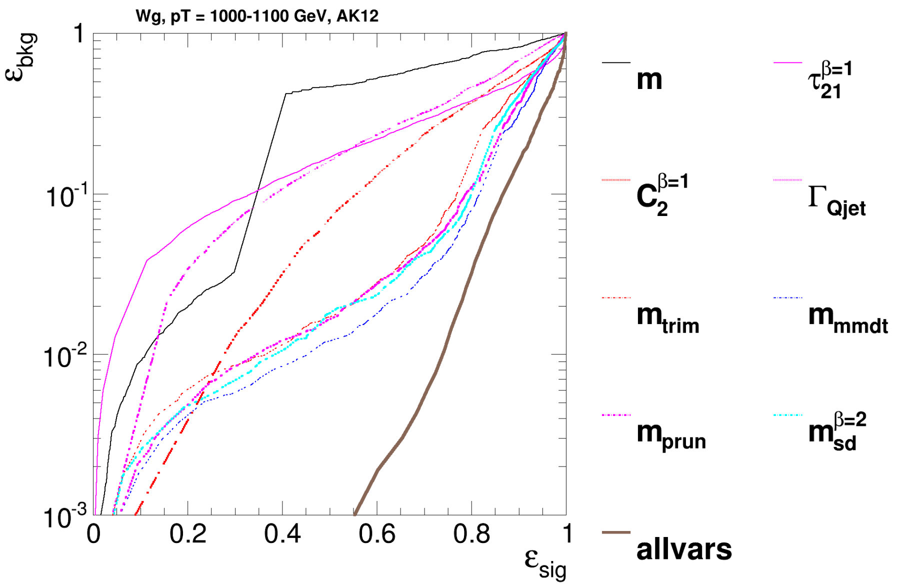
Worth also showing some ROC curves for mass + mass combinations?



(a) anti-k_{T} R=0.4, p_T 1.0-1.1 TeV bin

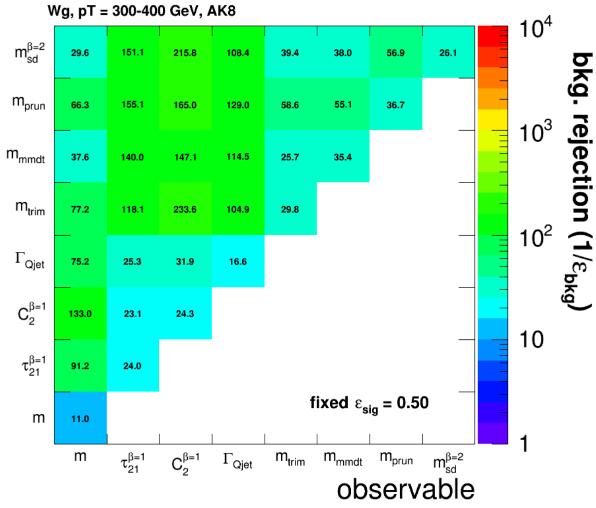


(b) anti-k_{T} R=0.8, p_T 1.0-1.1 TeV bin

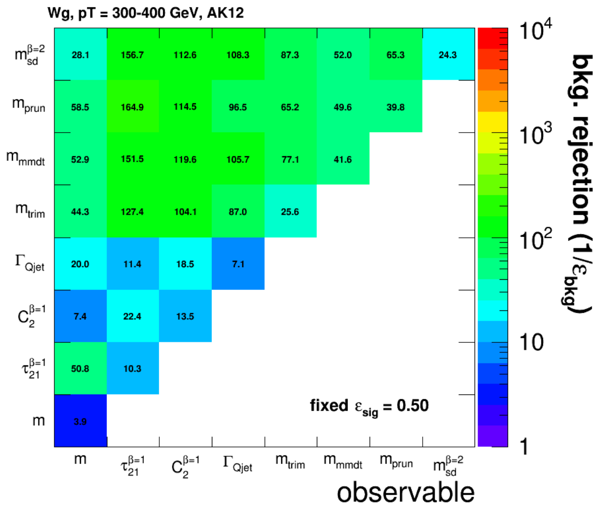


(c) anti-k_{T} R=1.2, p_T 1.0-1.1 TeV bin

Fig. 12 The ROC curve for all single variables considered for W tagging in the p_T 1.0-1.1 TeV bin using the anti-k_{T} R=0.4 algorithm (top), anti-k_{T} R=0.8 algorithm (middle) and R=1.2 algorithm (bottom).



(a) anti-k_T R=0.8, p_T 300-400 GeV bin

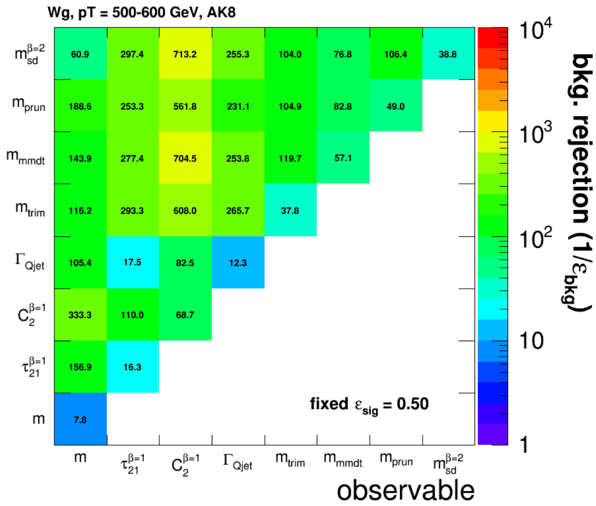


(b) anti-k_T R=1.2, p_T 300-400 GeV bin

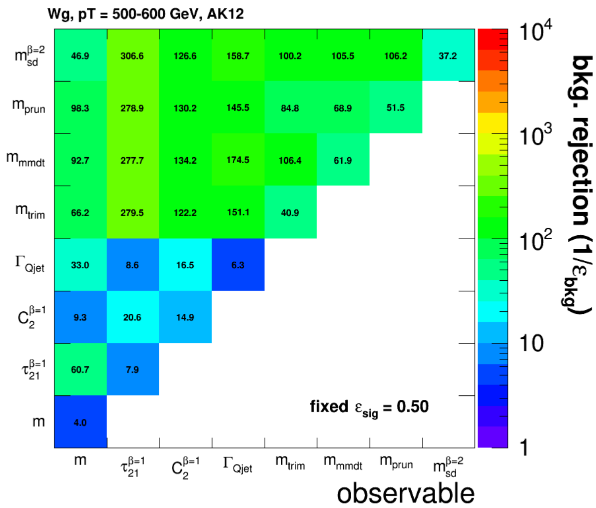
Fig. 13 The background rejection for a fixed signal efficiency (50%) of each BDT combination of each pair of variables considered, in the p_T 300-400 GeV bin using the anti-k_T R=0.8 algorithm (top) and R=1.2 algorithm (bottom).

Table 1 Action of various groomers on the jet mass distribution in the different phase space regions. For pruning, $a_{\rm prune} = z_{\rm cut} R_0$ and for trimming $a_{\rm trim} = \sqrt{z_{\rm cut}} R_{\rm sub}$.

Action	Pruning	Trimming	mMDT	SD ($\beta > 0$)
$m > \sqrt{z_{\rm cut}} R_0 p_T$	–	–	–	–
$m < \sqrt{z_{\rm cut}} R_0 p_T$ $m > a_x p_T$	cuts soft & soft-collinear	cuts soft & soft-collinear	cuts soft & soft-collinear	cuts soft & partially (β) on soft-collinear
$m < a_x p_T$	cuts partially on both soft & soft-collinear	–	cuts soft & soft-collinear	cuts soft & partially (β) on soft-collinear

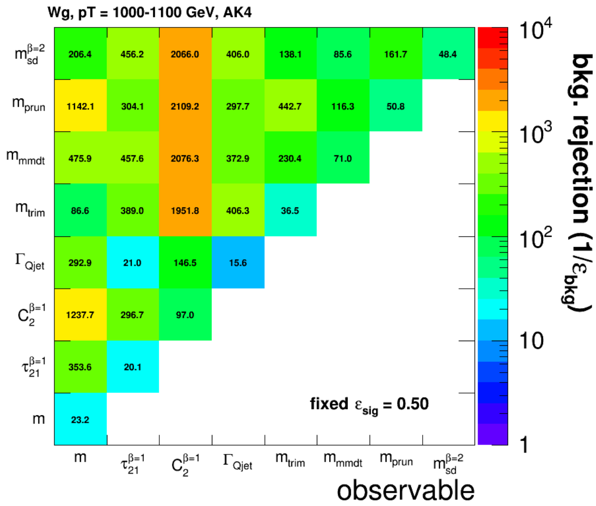


(a) anti-k_{T} R=0.8, p_T 500-600 GeV bin

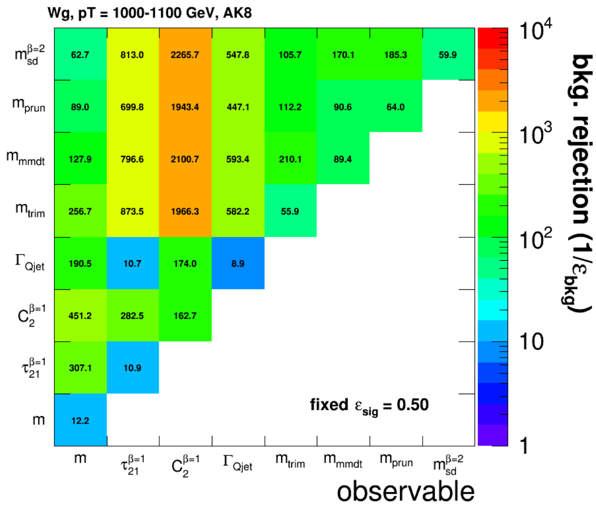


(b) anti-k_{T} R=1.2, p_T 500-600 GeV bin

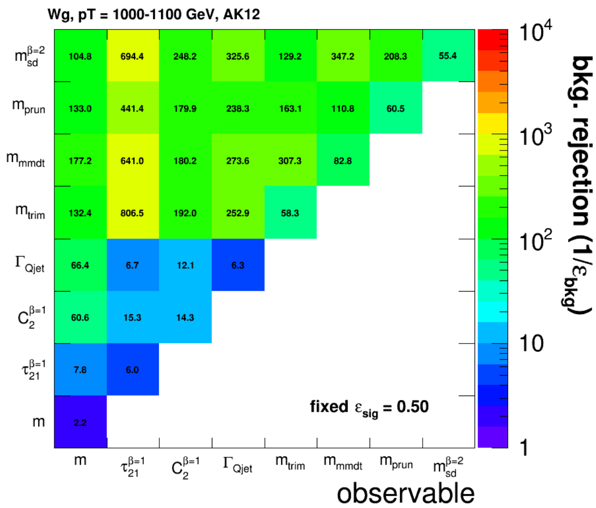
Fig. 14 The background rejection for a fixed signal efficiency (50%) of each BDT combination of each pair of variables considered, in the p_T 500-600 GeV bin using the anti-k_{T} R=0.8 algorithm (top) and R=1.2 algorithm (bottom).



(a) anti-k_T R=0.4, p_T 1.0-1.1 TeV bin



(b) anti-k_T R=0.8, p_T 1.0-1.1 TeV bin



(c) anti-k_T R=1.2, p_T 1.0-1.1 TeV bin

Fig. 15

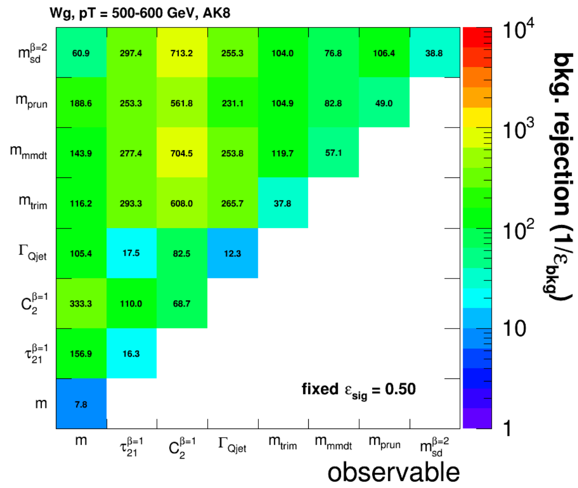
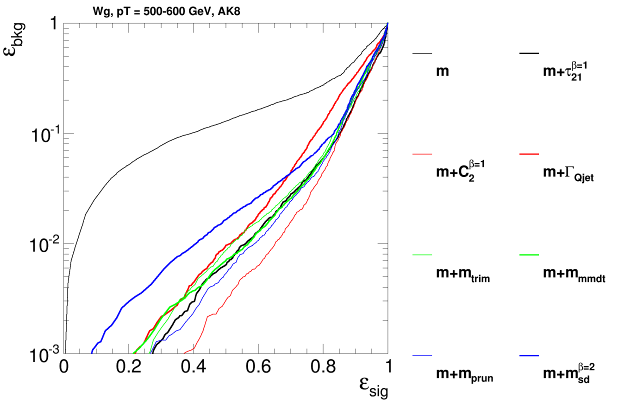
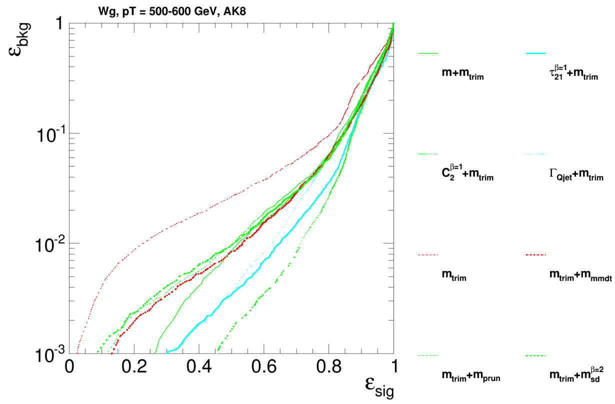


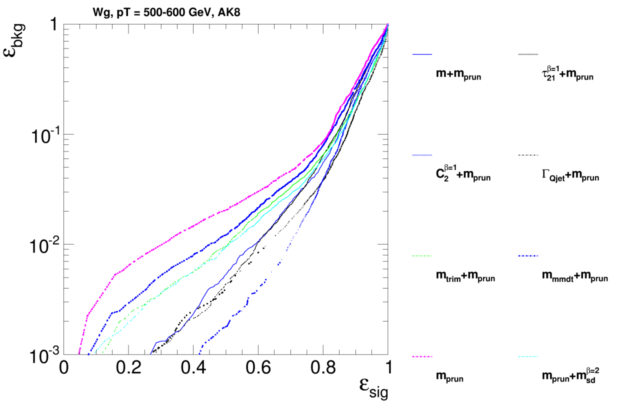
Fig. 16 The background efficiency for a fixed signal efficiency (50%) of each BDT combination of each pair of variables considered, in the p_T 500 GeV bin using the anti-k_T R=0.8 algorithm.



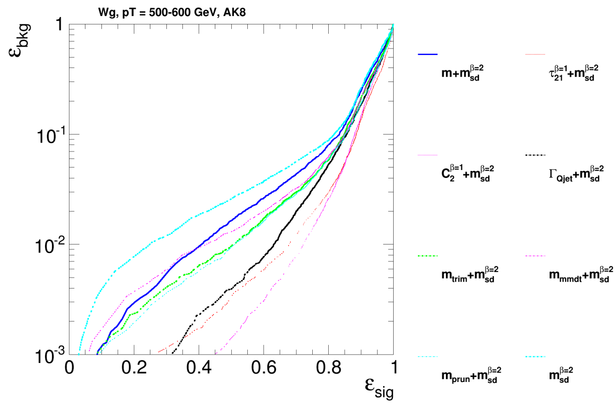
(a) Ungroomed mass + X



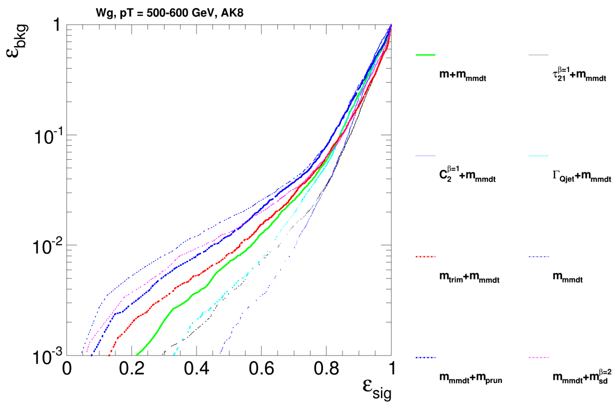
(b) Trimmed mass + X



(c) Pruned mass + X

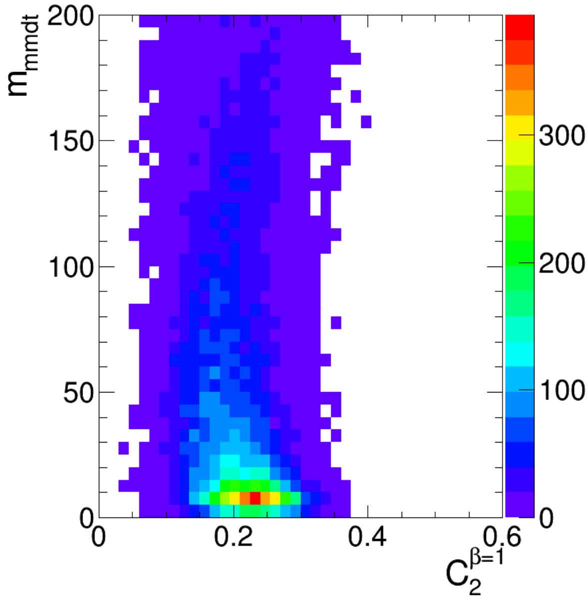


(d) Soft drop mass ($\beta = 2$) + X

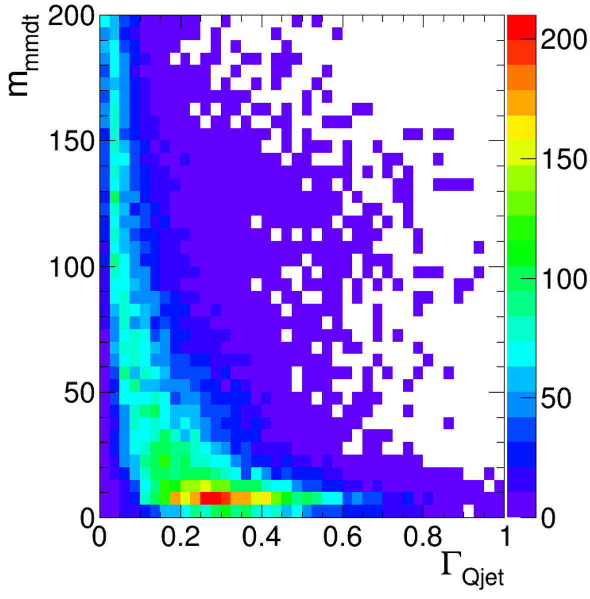


(e) mMDT mass + X

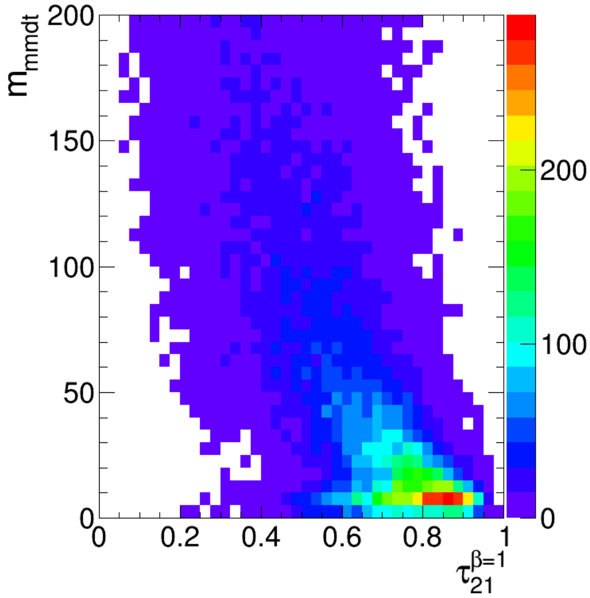
Fig. 17 The BDT combinations of each mass variable with every other variable considered in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm.



(a) mMDT mass vs $C_2^{\beta=1}$



(b) mMDT mass vs Γ_{Qjet}



(c) mMDT mass vs $\tau_{21}^{\beta=1}$

Fig. 18 2-D plots showing the correlation between mMDT mass and various substructure variables in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm in the gg sample.

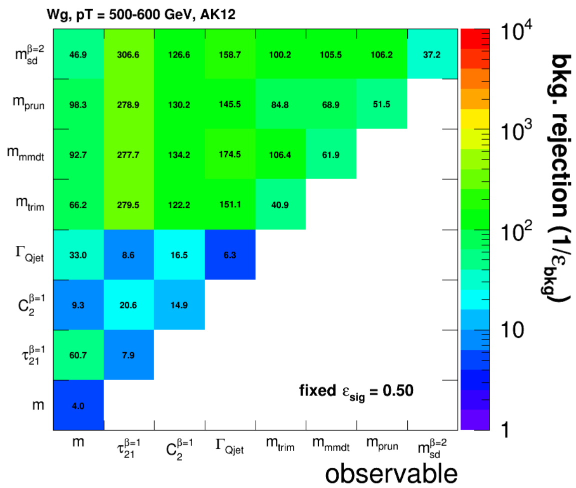
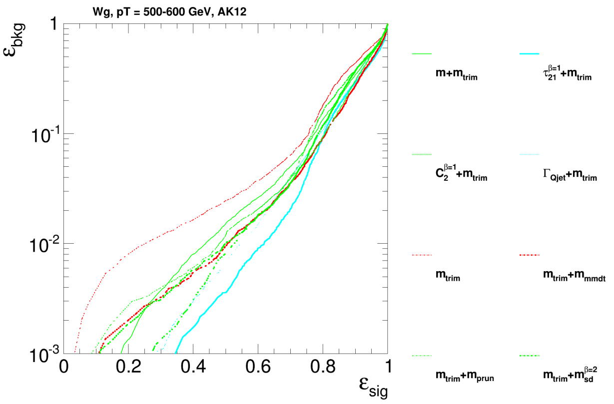
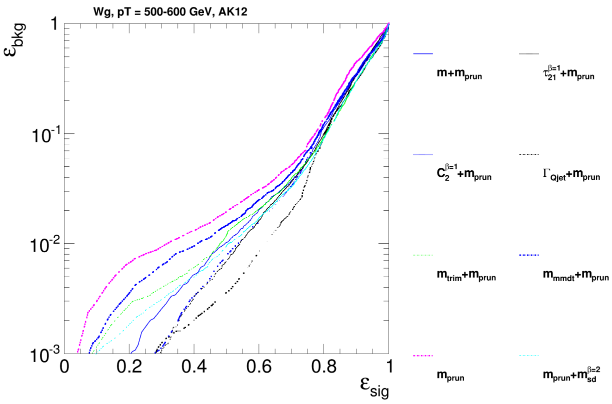


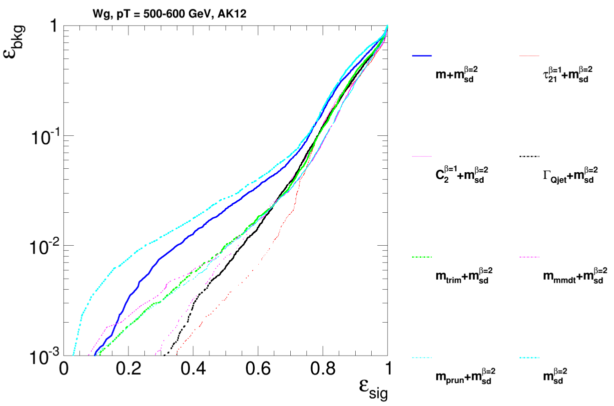
Fig. 19 The background efficiency for a fixed signal efficiency (50%) of each BDT combination of each pair of variables considered, in the p_T 500 GeV bin using the anti-k_{T} R=1.2 algorithm.



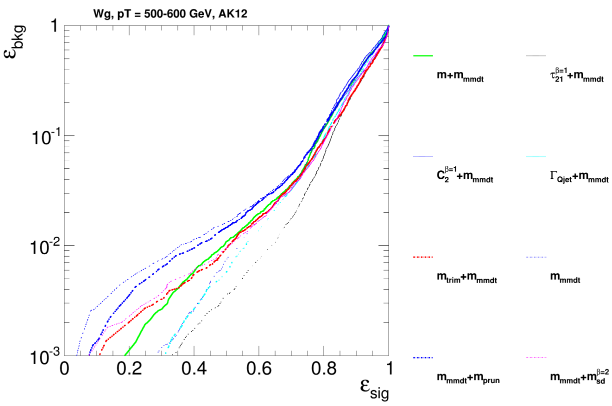
(a) Trimmed mass + X



(b) Pruned mass + X

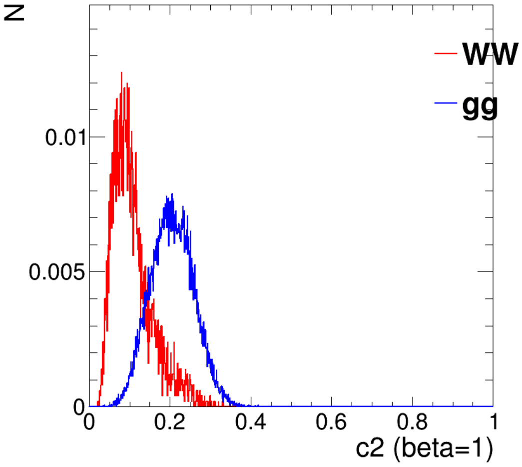


(c) Soft drop mass ($\beta = 2$) + X

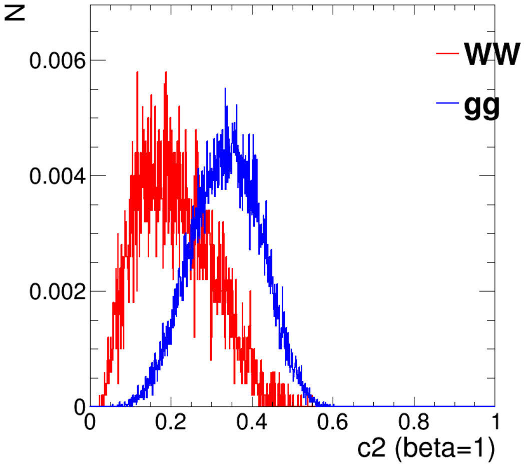


(d) mMDT mass + X

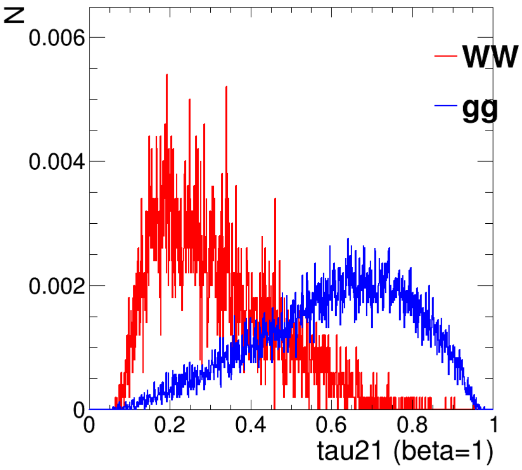
Fig. 20 The BDT combinations of each mass variable with every other variable considered in the p_T 500 GeV bin using the anti-k_{T} R=1.2 algorithm.



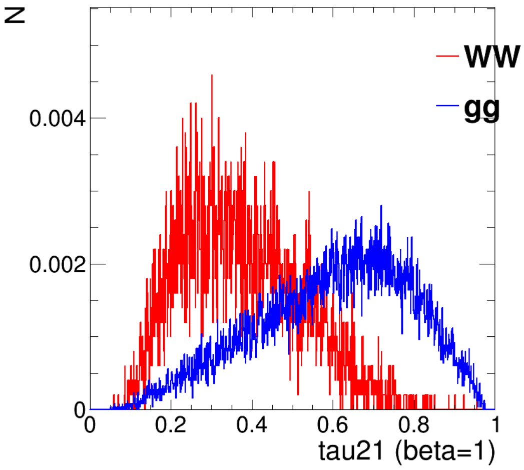
(a) $C_2^{\beta=1}$, R=0.8



(b) $C_2^{\beta=1}$, R=1.2

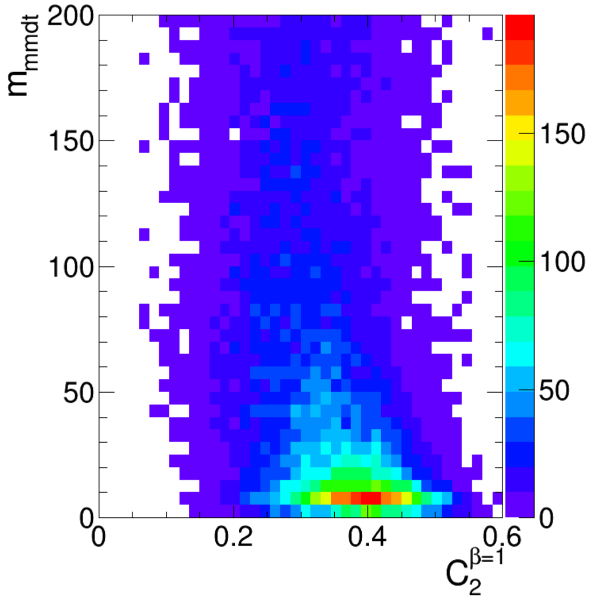


(c) $\tau_{21}^{\beta=1}$, R=0.8

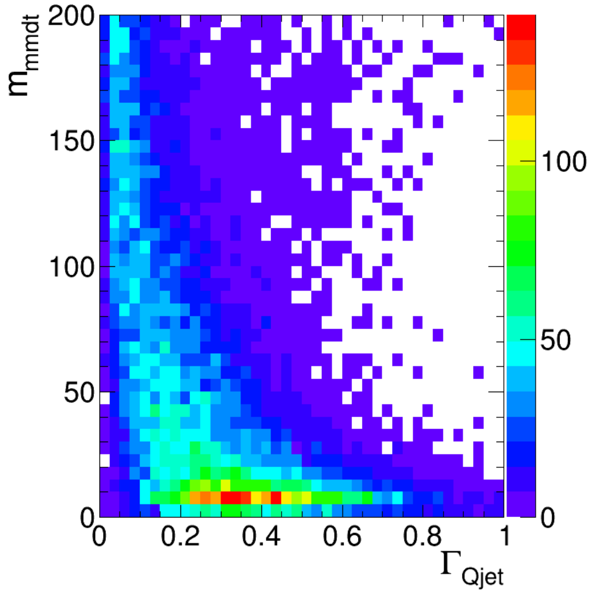


(d) $\tau_{21}^{\beta=1}$, R=1.2

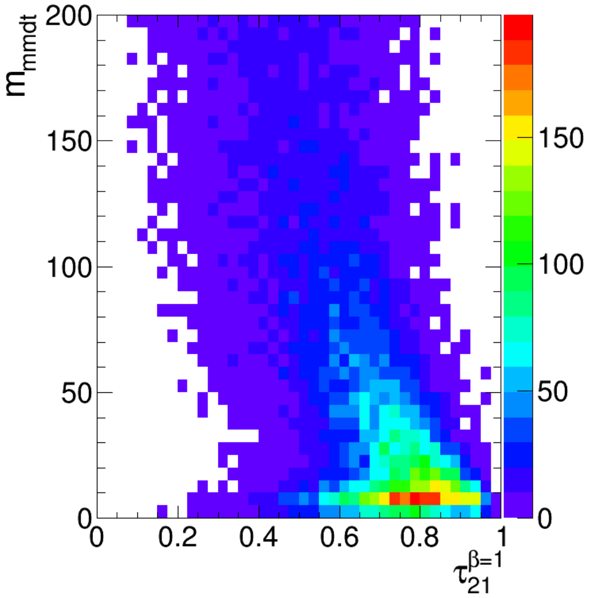
Fig. 21 Comparisons of the QCD background to the WW signal in the p_T 500 GeV bin for $C_2^{\beta=1}$ and $\tau_{21}^{\beta=1}$ variables and using the R=0.8 and R=1.2 anti-$k_{\rm T}$ distance parameters.



(a) mMDT mass vs $C_2^{\beta=1}$

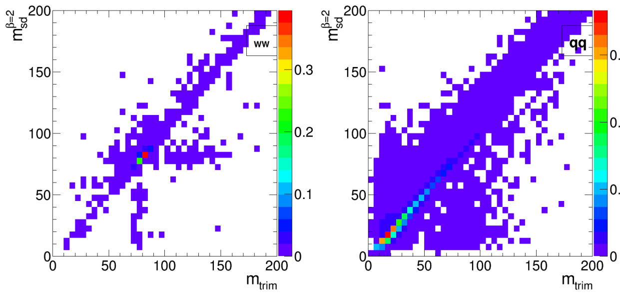


(b) mMDT mass vs Γ_{Qjet}

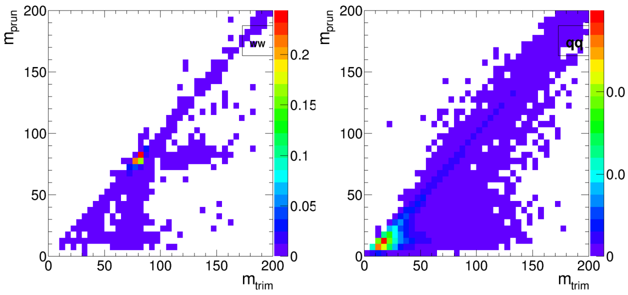


(c) mMDT mass vs $\tau_{21}^{\beta=1}$

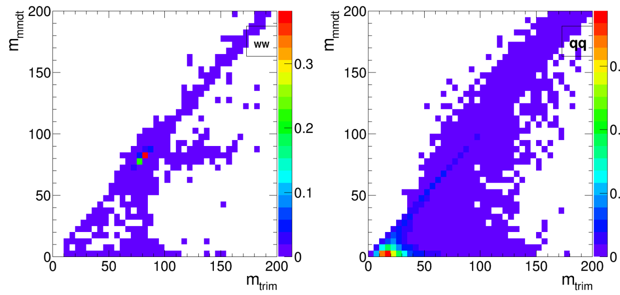
Fig. 22 2-D plots showing the correlation between mMDT mass and various substructure variables in the p_T 500 GeV bin using the anti-k_{T} R=1.2 algorithm in the gg sample.



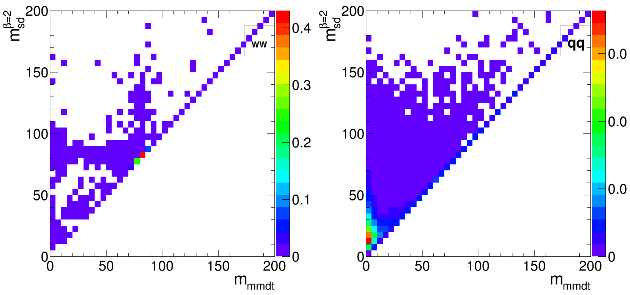
(a) Trimmed mass vs Soft drop mass ($\beta = 2$)



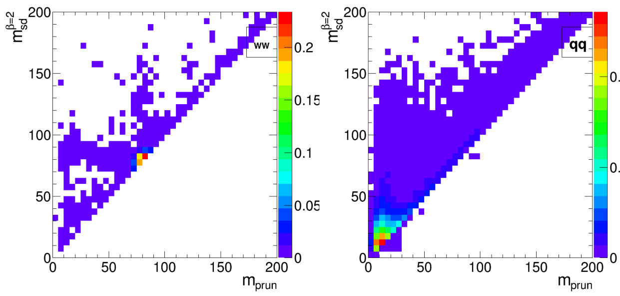
(b) Trimmed mass vs Pruned mass



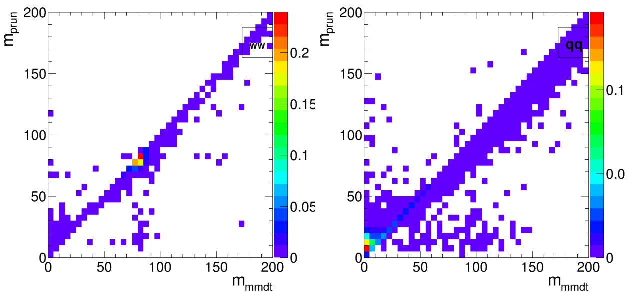
(c) Trimmed mass vs mMDT mass



(d) mMDT mass vs Soft drop mass ($\beta = 2$)

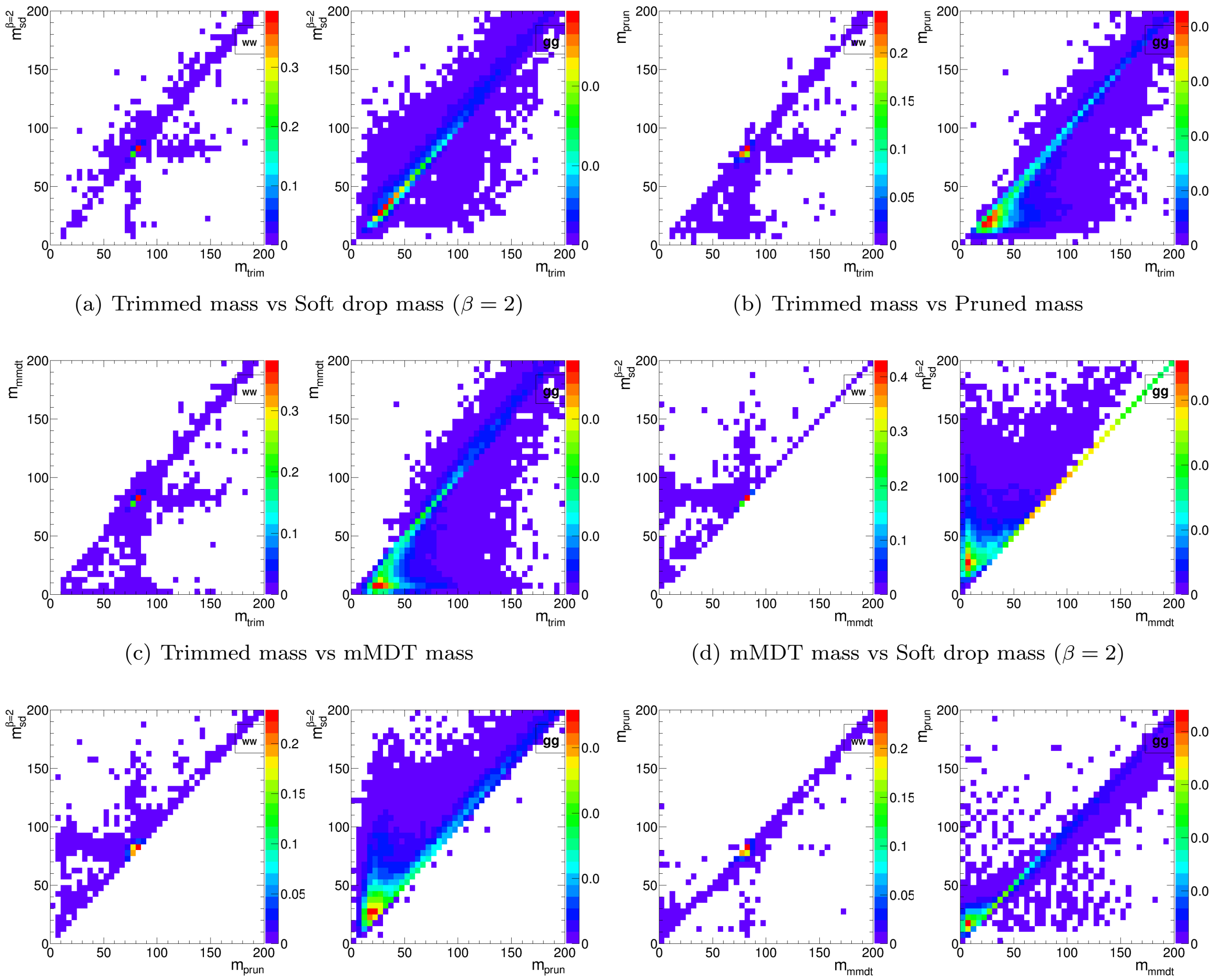


(e) Pruned mass vs Soft drop mass ($\beta = 2$)



(f) mMDT mass vs Pruned mass

Fig. 23 2-D plots showing the correlation between different types of groomed mass in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm, separately for the jets in the $X \to WW$ sample and the jets in the quark-quark sample.



(a) Trimmed mass vs Soft drop mass ($\beta = 2$)

(b) Trimmed mass vs Pruned mass

(c) Trimmed mass vs mMDT mass

(d) mMDT mass vs Soft drop mass ($\beta = 2$)

(e) Pruned mass vs Soft drop mass ($\beta = 2$)

(f) mMDT mass vs Pruned mass

Fig. 24 2-D plots showing the correlation between different types of groomed mass in the p_T 500 GeV bin using the anti-k_{T} R=0.8 algorithm, separately for the jets in the $X \to WW$ sample and the jets in the gluon-gluon sample.

7 Top Tagging

In this section, we study the identification of boosted top quarks at Run II of the LHC. Boosted top quarks result in large-radius jets with complex substructure, containing a b-subjet and a boosted W. The additional kinematic handles coming from the reconstruction of the W mass and b-tagging allows a very high degree of discrimination of top quark jets from QCD backgrounds.

We consider top quarks with moderate boost (600-1000 GeV), and perhaps most interestingly, at high boost ($\gtrsim$ 1500 GeV). Top tagging faces several challenges in the high-p_T regime. For such high-p_T jets, the b-tagging efficiencies are no longer reliably known. Also, the top jet can also accompanied by additional radiation with $p_T \sim m_t$, leading to combinatoric ambiguities of reconstructing the top and W, and the possibility that existing taggers or observables shape the background by looking for subjet combinations that reconstruct m_t/m_W. To study this, we examine the performance of both mass-reconstruction variables, as well as shape observables that probe the three-pronged nature of the top jet and the accompanying radiation pattern.

7.1 Methodology

We study a number of top-tagging strategies, in particular:

1. HEPTopTagger
2. Johns Hopkins Tagger (JH)
3. Trimming
4. Pruning

The top taggers have criteria for reconstructing a top and W candidate, while the grooming algorithms (trimming and pruning) do not incorporate a W-identification step. For a level playing field, we construct a W candidate from the three leading subjets by taking the pair of subjets with the smallest invariant mass; in the case that only two subjets are reconstructed, we take the mass of the leading subjet. All of the above taggers and groomers incorporate a step to remove pile-up and other soft radiation.

We also consider the performance of jet shape observables. In particular, we consider the N-subjettiness ratios $\tau_{32}^{\beta=1}$ and $\tau_{21}^{\beta=1}$, energy correlation function ratios $C_3^{\beta=1}$ and $C_2^{\beta=1}$, and the Qjet mass volatility Γ. In addition to the jet shape performance, we combine the jet shapes with the mass-reconstruction methods listed above to determine the optimal combined performance.

To quantify the performance of each set of variables, we combine the relevant tagger output observables and/or jet shapes into a boosted decision tree (BDT), which determines the optimal multivariable cut. Additionally, because each tagger has two inputs (list, or maybe refer back to Section 3), we scan over reasonable values of the inputs to determine the optimal value for each top tagging signal efficiency. This allows a direct comparison of the optimized version of each tagger.

7.2 Single-observable performance

We start by investigating the behavior of individual jet substructure observables. Because of the rich, three-pronged structure of the top decay, it is expected that combinations of masses and jet shapes will far outperform single observables in identifying boosted tops. However, a study of the top-tagging performance of single variables facilitates a direct comparison with the W tagging results in Section 6, and also allows a straightforward examination of the performance of each observable for different p_T and jet radius.

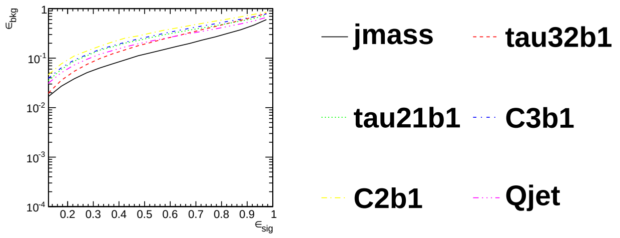
Fig. 25 shows the ROC curves for each of the top-tagging observables, with the bare jet mass also plotted for comparison. Unlike W tagging, the jet shape observables perform more poorly than jet mass. *(Check reasoning: this argument due to Andrew Larkoski).* As an example illustrating why this is the case, consider N-subjettiness. The W is two-pronged and the top is three-pronged; therefore, we expect τ_{21} and τ_{32} to be the best-performant N-subjettiness ratio, respectively. However, τ_{21} also contains an implicit cut on the denominator, τ_1, which is strongly correlated with jet mass. Therefore, τ_{21} combines both mass and shape information to some extent. By contrast, and as is clear in Fig.25(a), the best shape for top tagging is τ_{32}, which contains no information on the mass. Therefore, it is unsurprising that the shapes most useful for top tagging are less sensitive to the jet mass, and under-perform relative to the corresponding observables for W tagging.

Of the two top tagging algorithms, the Johns Hopkins (JH) tagger out-performs the HEPTopTagger in its signal-to-background separation of both the top and W candidate masses, with larger discrepancy at higher p_T and larger jet radius. In Fig. 26, we show the histograms for the top mass output from the JH and HEPTopTagger for different p_T and R, optimized at a signal efficiency of 30%. The likely reason for this behavior is that, in the HEPTopTagger algorithm, the jet is filtered to select the five hardest subjets, and then three subjets are chosen which reconstruct the top mass. This requirement tends to shape a peak in the QCD background around m_t for the HEPTopTagger, while the JH tagger has no such requirement. It has been suggested by Anders *et al.* [4] that performance in the HEPTopTagger may be improved by selecting the three subjets reconstructing the top only among those that pass the W mass constraints, which somewhat reduces the shaping of the background. *Maybe try this out with my code to see if it helps?*

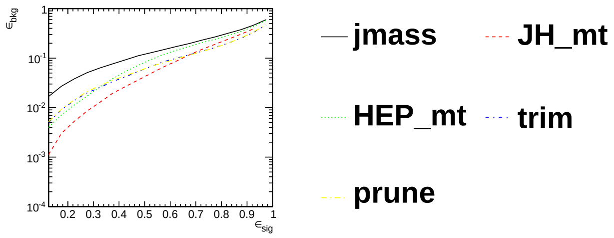
We also directly compare each variable's performance for different jet p_T and radius. The results are shown in Figs. 27-29 for different p_T bins and Figs. 30-32 for different R values. The input parameters of the taggers, groomers, and shape variables are separately optimized for each p_T and radius. If we only optimize the tagger inputs for one value of p_T and R, the ROC curve behavior does not change substantially from one where the inputs are optimized at each p_T and R value; however, not all signal efficiencies are possible for every choice of tagger input, since the baseline selection efficiency might be too low.

7.3 Performance of multivariable combinations

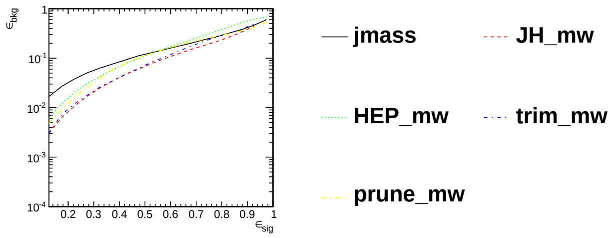
7.4 Performance at Sub-Optimal Working Points



(a) Jet shapes

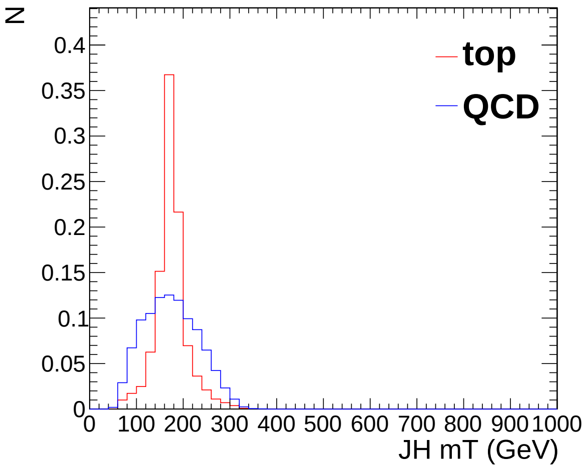


(b) Top mass

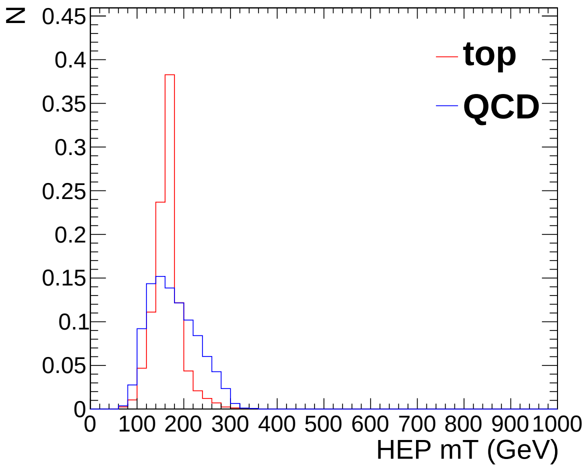


(c) W mass

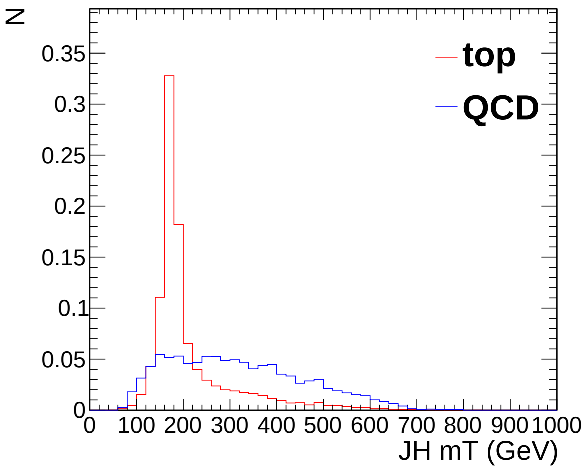
Fig. 25 Comparison of single-variable top-tagging performance in the p_T 1000-1100 GeV bin using the anti-k_{T}, R=0.8 algorithm.



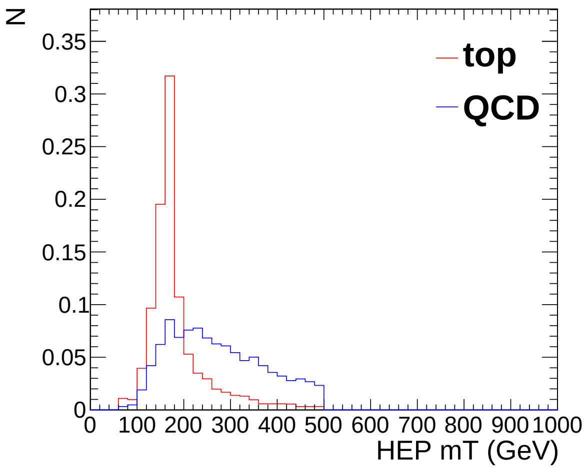
(a) Johns Hopkins Tagger, $p_T = 600 - 700$ GeV, $R = 0.8$



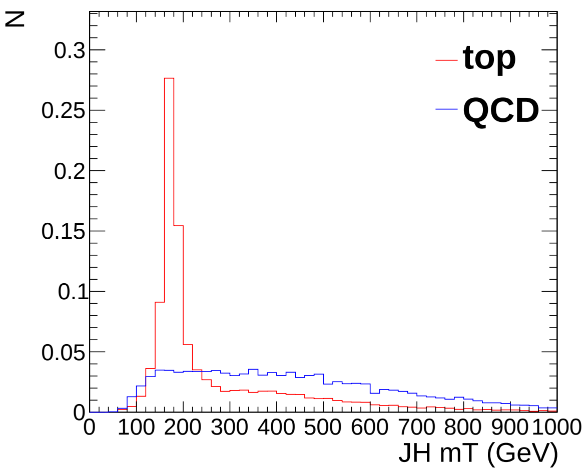
(b) HEPTopTagger, $p_T = 600 - 700$ GeV, $R = 0.8$



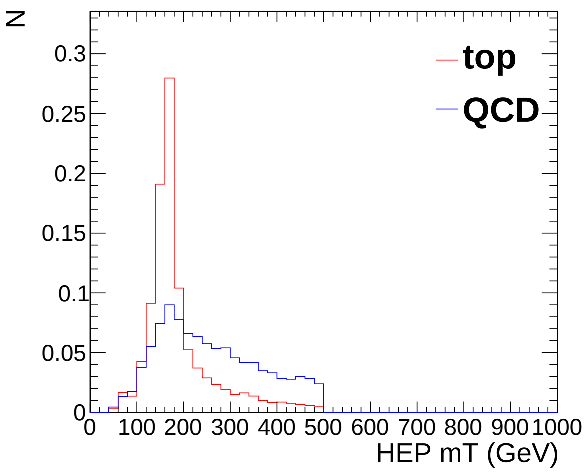
(c) Johns Hopkins Tagger, $p_T = 1500 - 1600$ GeV, $R = 0.8$



(d) HEPTopTagger, $p_T = 1500 - 1600$ GeV, $R = 0.8$

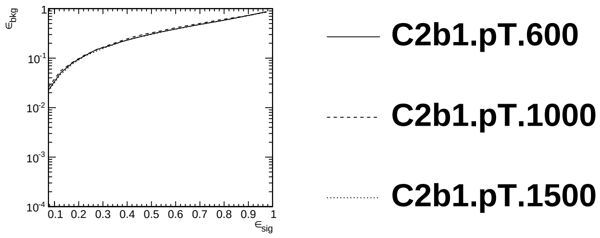


(e) Johns Hopkins Tagger, $p_T = 1500 - 1600$ GeV,$R = 1.2$

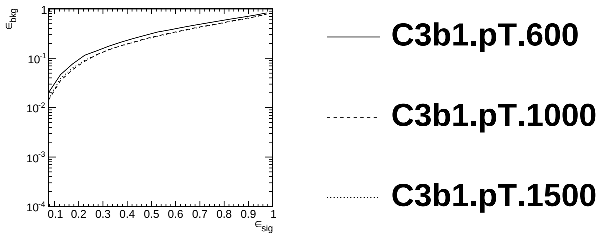


(f) HEPTopTagger, $p_T = 1500 - 1600$ GeV, $R = 1.2$

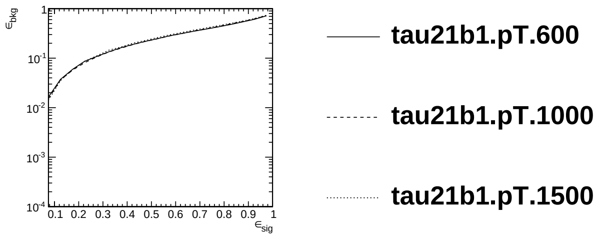
Fig. 26 Comparison of individual jet shape performance at different p_T using the anti-k_{T} R=0.8 algorithm.



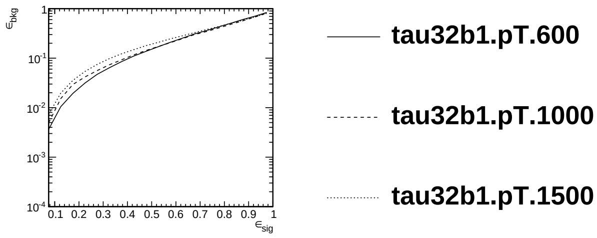
(a) C2



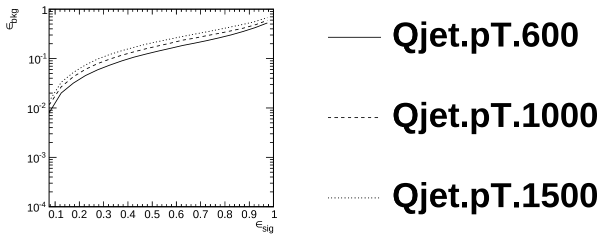
(b) C3



(c) τ_{21}

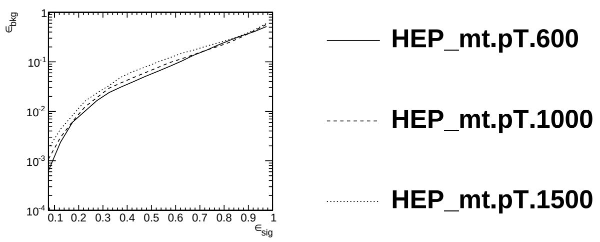


(d) τ_{32}

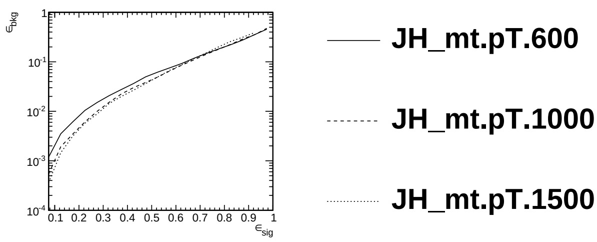


(e) Mass volatility

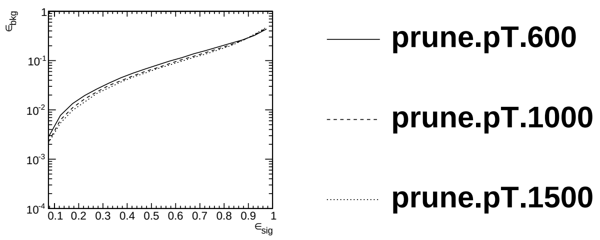
Fig. 27 Comparison of individual jet shape performance at different p_T using the anti-k_T R=0.8 algorithm.



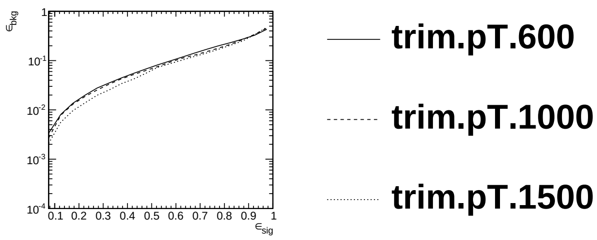
(a) HEPTopTagger m_t



(b) Johns Hopkins Tagger m_t

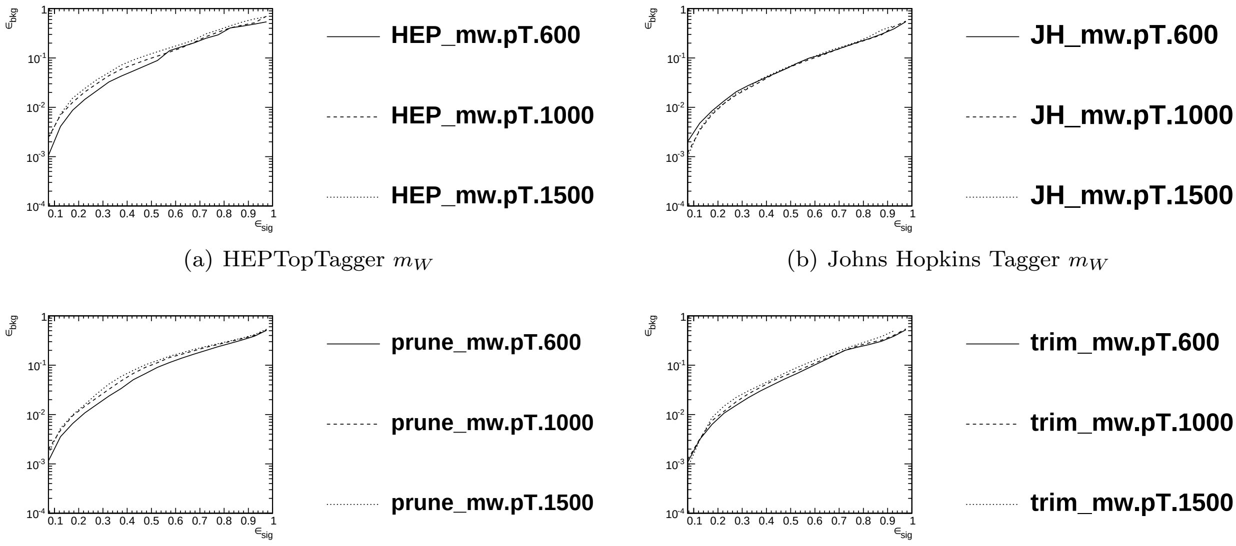


(c) Pruning m_t



(d) Trimming m_t

Fig. 28 Comparison of top mass performance of different taggers at different p_T using the anti-k_T R=0.8 algorithm.



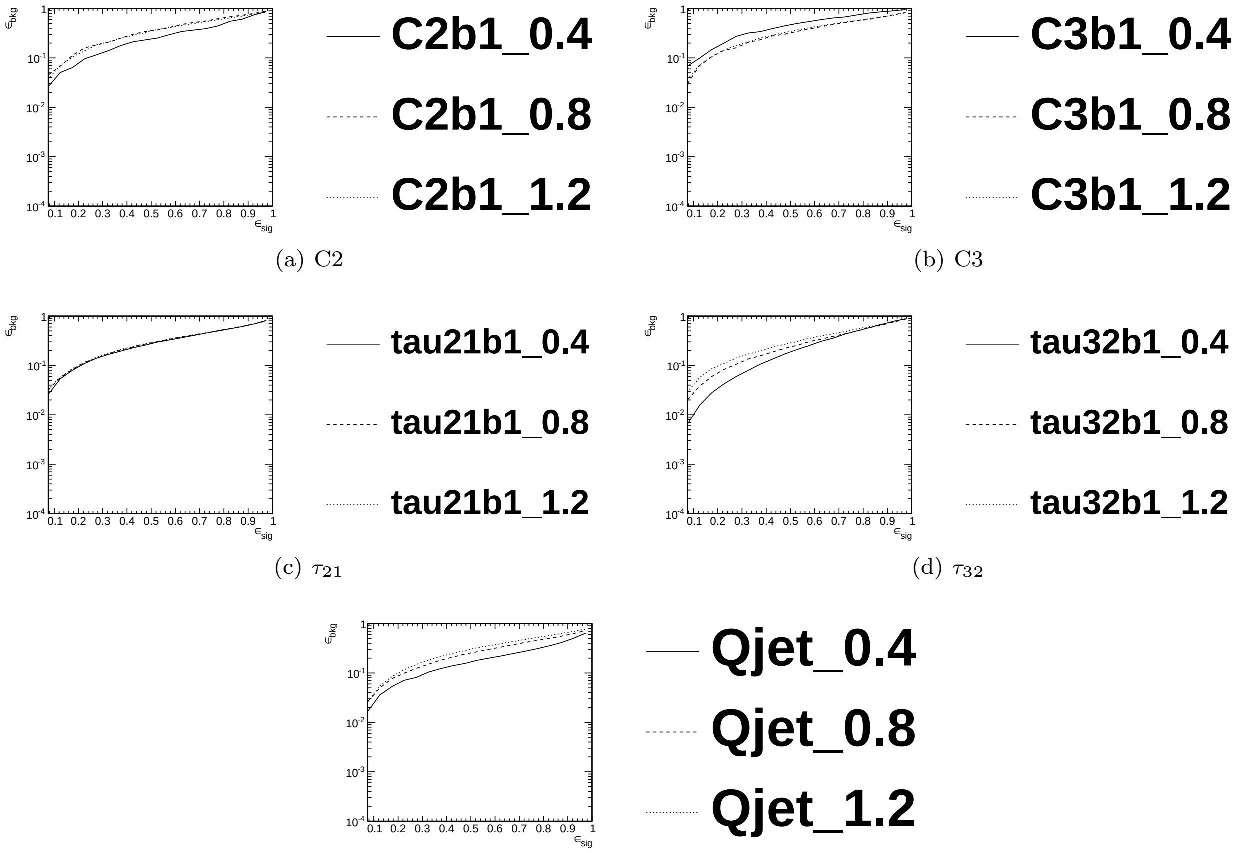
(a) HEPTopTagger m_W

(b) Johns Hopkins Tagger m_W

(c) Pruning m_W

(d) Trimming m_W

Fig. 29 Comparison of W mass performance of different taggers at different p_T using the anti-k_{T} R=0.8 algorithm.



(a) C2

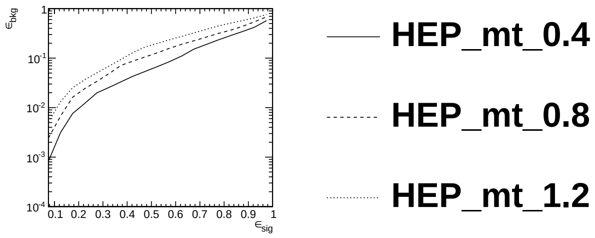
(b) C3

(c) τ_{21}

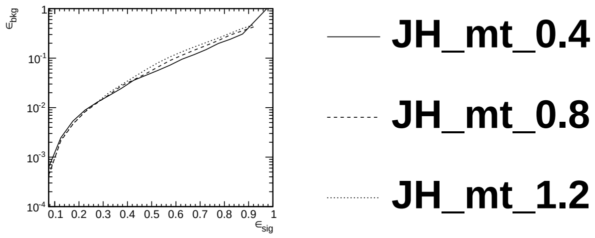
(d) τ_{32}

(e) Mass volatility

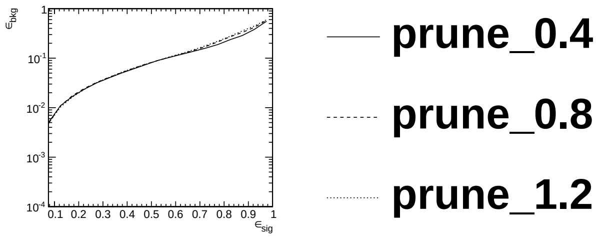
Fig. 30 Comparison of individual jet shape performance at different R in the $p_T = 1500 - 1600$ GeV bin.



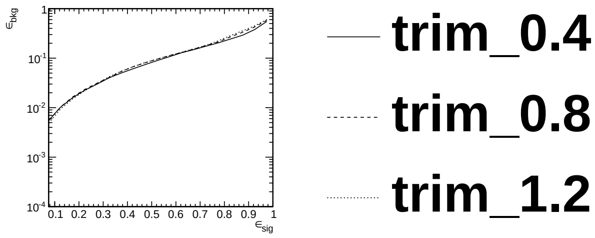
(a) HEPTopTagger m_t



(b) Johns Hopkins Tagger m_t

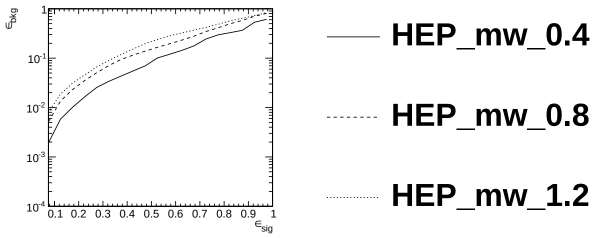


(c) Pruning m_t

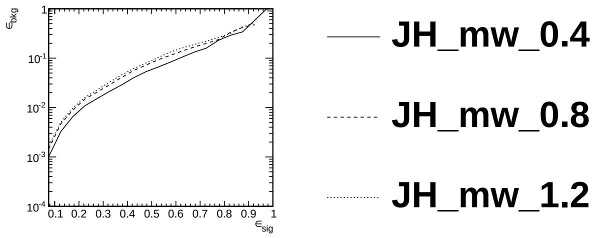


(d) Trimming m_t

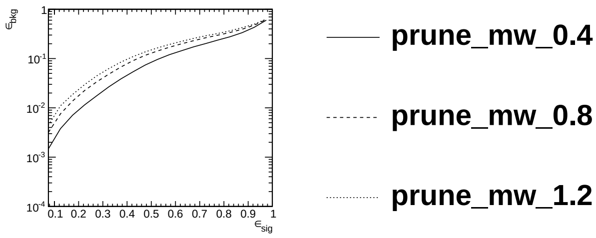
Fig. 31 Comparison of top mass performance of different taggers at different R in the $p_T = 1500 - 1600$ GeV bin.



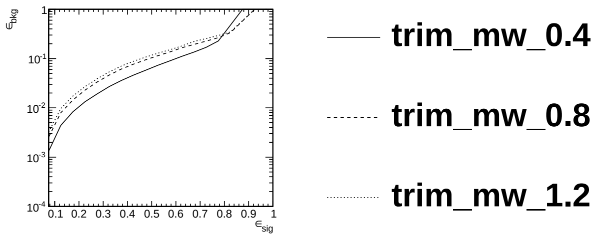
(a) HEPTopTagger m_W



(b) Johns Hopkins Tagger m_W

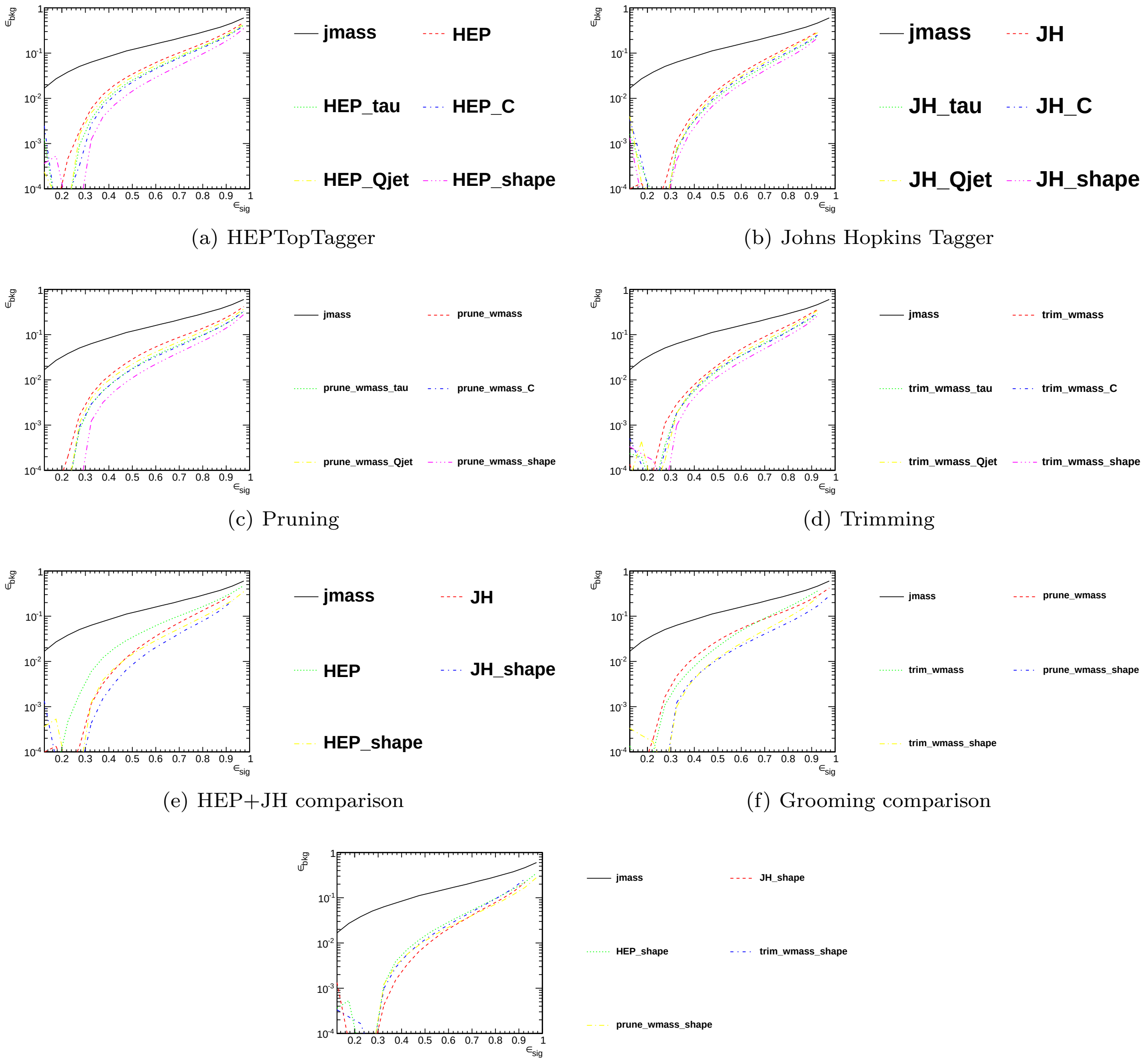


(c) Pruning m_W



(d) Trimming m_W

Fig. 32 Comparison of W mass performance of different taggers at different R in the $p_T = 1500 - 1600$ GeV bin.



(a) HEPTopTagger

(b) Johns Hopkins Tagger

(c) Pruning

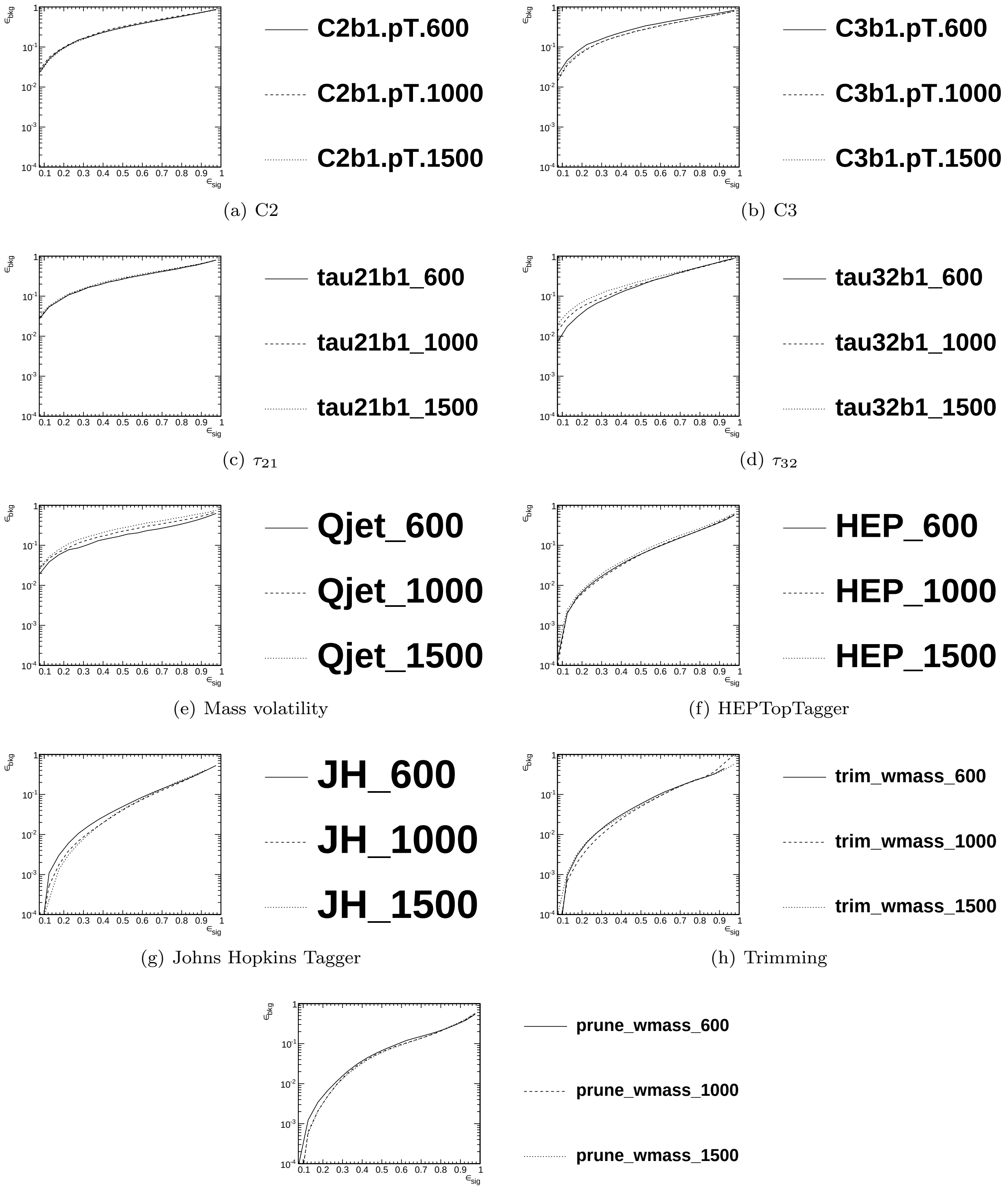
(d) Trimming

(e) HEP+JH comparison

(f) Grooming comparison

(g) Comparison of Tagger+Shape

Fig. 33 The BDT combinations in the p_T 1000-1100 GeV bin using the anti-k_{T} R=0.8 algorithm.



(a) C2 (b) C3

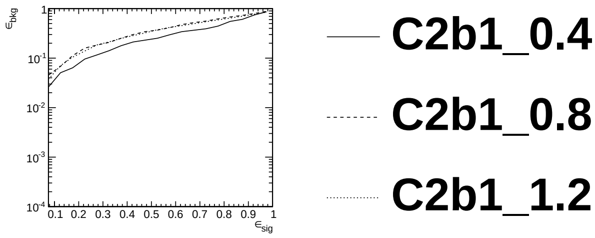
(c) τ_{21} (d) τ_{32}

(e) Mass volatility (f) HEPTopTagger

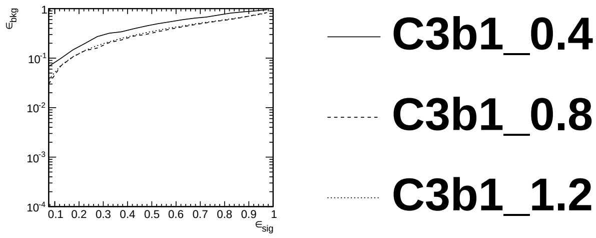
(g) Johns Hopkins Tagger (h) Trimming

(i) Pruning

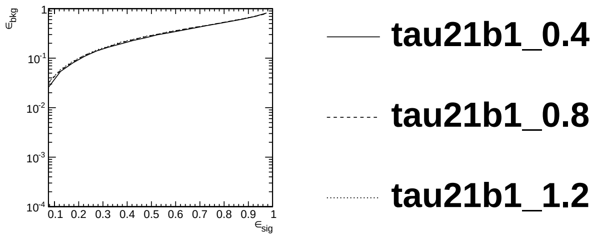
Fig. 34 Comparison of tagger and jet shape performance at different p_T using the anti-k_T R=0.8 algorithm.



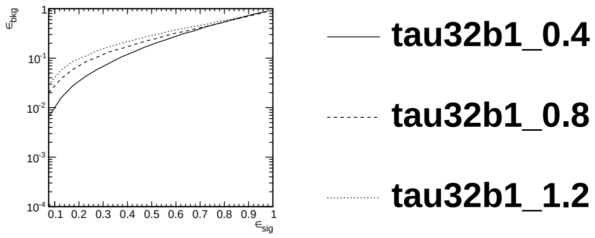
(a) C2



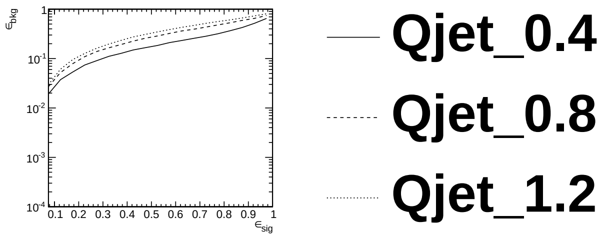
(b) C3



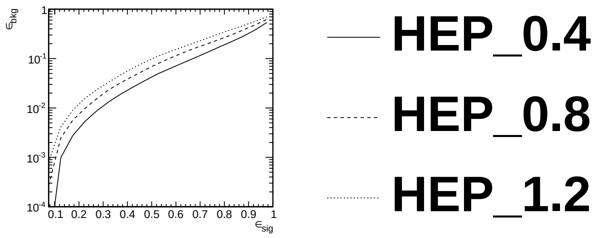
(c) τ_{21}



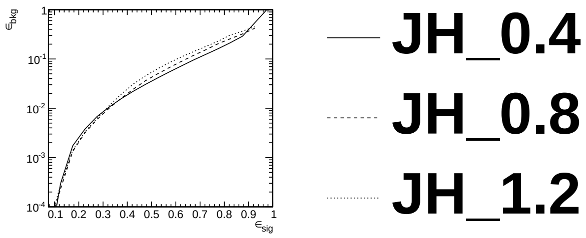
(d) τ_{32}



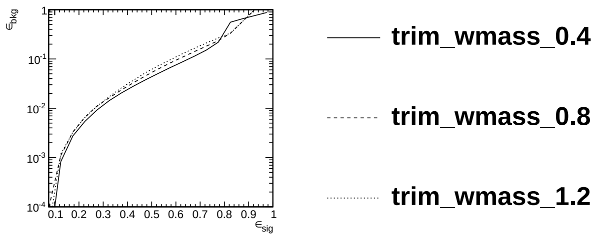
(e) Mass volatility



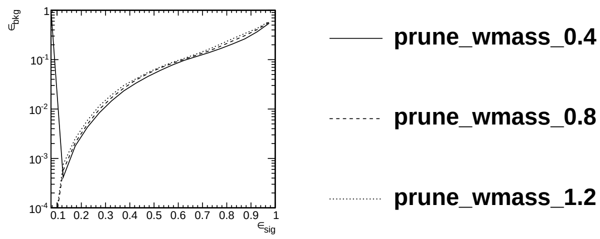
(f) HEPTopTagger



(g) Johns Hopkins Tagger



(h) Trimming



(i) Pruning

Fig. 35 Comparison of tagger and jet shape performance at different radius at p_T = 1.5-1.6 TeV.

8 Summary & Conclusions

This report discussed the correlations between observables and looked forward to jet substructure at Run II of the LHC at 14 TeV center-of-mass collisions eneergies.

Acknowledgements

We thank the Department of Physics at the University of Arizona and for hosting the conference at the Little America Hotel. We also thank Harvard University for hosting the event samples used in this report. We also thank Hallie Bolonkin for the BOOST2013 poster design and Jackson Boelts' ART465 class (fall 2012) at the University of Arizona School of Arts VisCom program. (NEED TO ASK PETER LOCH FOR MORE ACKNOWLEDGEMENTS)

References

1. A. Abdesselam, E. B. Kuutmann, U. Bitenc, G. Brooijmans, J. Butterworth, et al., *Boosted objects: A Probe of beyond the Standard Model physics*, *Eur.Phys.J.* **C71** (2011) 1661, [arXiv:1012.5412].
2. A. Altheimer, S. Arora, L. Asquith, G. Brooijmans, J. Butterworth, et al., *Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks*, *J.Phys.* **G39** (2012) 063001, [arXiv:1201.0008].
3. A. Altheimer, A. Arce, L. Asquith, J. Backus Mayes, E. Bergeaas Kuutmann, et al., *Boosted objects and jet substructure at the LHC*, arXiv:1311.2708.
4. C. Anders, C. Bernaciak, G. Kasieczka, T. Plehn, and T. Schell, *Benchmarking an Even Better HEPTopTagger*, *Phys.Rev.* **D89** (2014) 074047, [arXiv:1312.1504].