CLICdp-Note-2015-001
20 January 2015

Study of the effect of additional background channels on the top Yukawa coupling measurement at a 1.4 TeV CLIC

S. Redford[*], P. Roloff[*], M. Vogel[†]

[*] *CERN, Switzerland,* [†] *Pontificia Universidad Catolica de Chile, Chile*

Abstract

The contribution of additional samples to the background expectation in a direct measurement of the top Yukawa coupling is investigated. In a previous study, the physics potential of a direct measurement of the top Yukawa coupling was investigated, using the process $e^+e^- \to t\bar{t}H$ at a $\sqrt{s} = 1.4\,\text{TeV}$ Compact Linear Collider (CLIC) and reconstructed using the CLIC_SiD detector. The expected precision on the top Yukawa coupling was determined to be 4.27%, without beam polarisation. The inclusion of additional non-$t\bar{t}(+X)$ backgrounds has slightly increased the background expectation and an updated precision on the top Yukuwa coupling is determined to be 4.43% with no further optimisation of the selection.

Contents

1. Introduction

The top quark is the heaviest known fundamental particle, therefore its coupling to the Higgs boson, the top Yukawa coupling, g_{ttH}, represents the strongest of the Higgs couplings. The top Yukawa coupling can be accessed through the process $\mathrm{e^+e^-} \to \mathrm{t\bar{t}H}$, which is sensitive to the strength of the coupling at the ttH vertex squared.

The Compact Linear Collider (CLIC) [1] is a proposed future $\mathrm{e^+e^-}$ collider with possible staged centre-of-mass energies ($\sqrt{s}$) of 350 GeV, 1.4 TeV and 3 TeV. One of the main goals of CLIC will be precision measurements of the properties of the Higgs boson.

The present note is an update on the estimation of the statistical uncertainty of the top Yukawa coupling measured at $\sqrt{s} = 1.4\,\mathrm{TeV}$ assuming $1.5\,\mathrm{ab}^{-1}$ of integrated luminosity [2]. The motivation for this study was the need to determine whether additional sources of background, not considered in the original analysis, would contribute to the total background estimation and have an effect on the measured top Yukawa coupling uncertainty. The final states of these backgrounds and their kinematical distributions vary greatly from those of the $\mathrm{t\bar{t}H}$ signal. However, their high cross-sections warrant the study of their contribution to the total expected backgrounds.

2. Analysis strategy

The additional backgrounds were processed in the same manner as the samples in the original analysis. Two mutually exclusive event samples were produced by first identifying leptons and sorting events into two streams: one with exactly one identified lepton (semi-leptonic channel) and one with no identified leptons (hadronic channel), and then forcing jet reconstruction in each stream into 6 and 8 jets respectively. Events were selected using a boosted decision tree (BDT) as implemented in the Toolkit for Multivariate Data Analysis, TMVA [3], which is a multivariate classifier integrated into the ROOT analysis framework. Gradient boosting was used. The training for two BDTs, one for each channel, was done in the original analysis and it was applied to all the events in the additional samples. The additional samples were not used to train the BDT.

3. Simulation samples

WHIZARD 1.95 [4, 5] was used to generate all events in the additional samples. PYTHIA 6.422 [6] was used for hadronization. The full response of the CLIC_SiD detector [7, 8] was simulated for each event in GEANT4 [9, 10]. All samples are simulated with unpolarised beams.

Table 1 gives details of the simulated samples used in this study. The first two rows in the table correspond to the signal samples in the original analysis, included here for comparison. The first additional sample ($qqqq$) is mostly WW production, while the second ($qqqq\nu\nu$) is WW fusion, where "ν" can be ν or $\bar{\nu}$. The third additional sample ($qqqql\nu$) has a large fraction of $\mathrm{t\bar{t}}$, while the fourth ($qqqqll$) is mostly ZZ fusion. Here "q" can be q or $\bar{q}$ and $l = e, \mu, \tau$. The fourth column of the table shows the sample weight, which scales the simulated samples to an integrated luminosity of $1.5\,\mathrm{ab}^{-1}$.

There is a considerable fraction of $\mathrm{t\bar{t}}$ events present in $qqqql\nu$ (see Figure 1), which have already been accounted for in a dedicated sample and they need to be removed. A veto on $\mathrm{t\bar{t}}$ events is implemented by reconstructing the two top masses at generator level and excluding all events in $qqqql\nu$ where both of these reconstructed masses satisfy $150\,\mathrm{GeV/c^2} < M_t < 200\,\mathrm{GeV/c^2}$. The reconstruction of $\mathrm{t\bar{t}}$ is done by selecting events with a $\mathrm{b\bar{b}}$ pair. The lepton and neutrino are used to reconstruct a W, while the four quarks: $\mathrm{b\bar{b}}qq'$, where q' can be a bottom, are used to reconstruct (along with the first W) the second W and the two tops. The assignment of quarks that minimizes a χ^2 function parametrised in the top and W masses is selected to reconstruct the $\mathrm{t\bar{t}}$ event. This χ^2 function is defined by

Table 1: The four additional backgrounds and the two signal channels for comparison. Column 1 shows the simulated process. Column 2 shows the assumed cross-section. Column 3 shows the expected number of events in 1.5 ab^{-1}. Column 4 shows the sample weight.

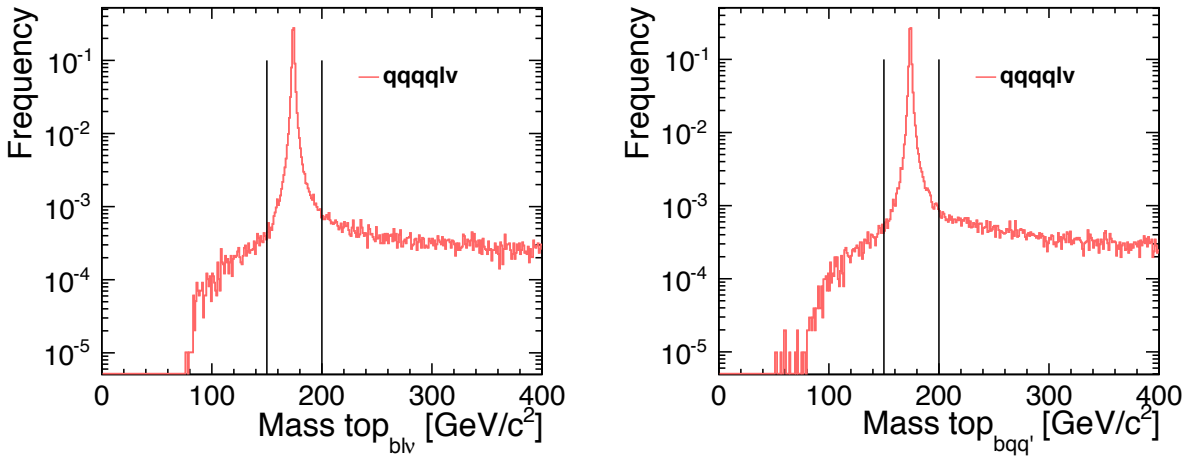
Process	Cross section (fb)	Events in 1.5 ab^{-1}	Sample weight
$t\bar{t}H$, 6 jet, $H \to b\bar{b}$	0.431	647	0.03
$t\bar{t}H$, 4 jet, $H \to b\bar{b}$	0.415	623	0.03
$qqqq$	1326	1.989×10^6	8.11
$qqqq\nu\nu$	24.7	37050	0.17
$qqqql\nu$	115.3[a]	68338	1.22
$qqqqll$	71.7	107550	0.65

[a] Before $t\bar{t}$ removal.

$$\chi^2 = \frac{(M_{bl\nu} - M_t)^2}{\sigma_t^2} + \frac{(M_{l\nu} - M_{W^\pm})^2}{\sigma_{W^\pm}^2} + \frac{(M_{bq\bar{q}'} - M_t)^2}{\sigma_t^2} + \frac{(M_{q\bar{q}'} - M_{W^\pm})^2}{\sigma_{W^\pm}^2},$$

where M_t =174.0 GeV/c^2, σ_t =1.1 GeV/c^2, $M_{W^\pm}$ =80.4 GeV/c^2, $\sigma_{W^\pm}$ =1.4 GeV/c^2 are obtained from Gaussian fits to the distributions of generator level top and W masses.

The fraction of events identified as $t\bar{t}$ and removed using this method is 40%. Unless expressly stated otherwise, all figures and table entries in this note do not include events in $qqqql\nu$ which are identified as $t\bar{t}$ in this manner. Interference effects with the removed process are small and neglected.



(a) (b)

Figure 1: $t\bar{t}$ events present in $qqqql\nu$. (a): Generator level top mass reconstructed from $bl\nu$. (b): Generator level top mass reconstructed from bqq'. The mass window used for the veto is shown in both graphs (150 GeV/c^2 < M_t <200 GeV/c^2). All distributions are normalized to unity.

4. Reconstruction

Lepton identification is described in detail in the original analysis. Electrons, muons and hadronically decaying taus are identified and events with exactly one lepton are selected for the semi-leptonic channel. Events with no identified leptons are selected for the hadronic channel. The lepton finder fails to identify leptons more often in the $qqqqll$ sample than in $qqqql\nu$ (see Table 2). However, the final states in $qqqqll$ originate mostly from ZZ fusion, hence these dileptons are mostly the beam leptons after radiation of a photon or Z bosons. These high-energy electrons and positrons have polar angle distributions that are peaked in the forward direction and many are outside of the detector acceptance (see Figure 2). Leptons in $qqqqll$ at high polar angles are mostly high energy beam electrons. Leptons in $qqqql\nu$ are from leptonic decays of W. Appendices A and B show the event shapes, b-tags and kinematic distributions of the new backgrounds in the semi-leptonic and hadronic channel, respectively. The corresponding signal distributions in each channel are also shown for comparison. As can be seen in the figures in the appendices, the distributions of these backgrounds vary greatly from those of the $t\bar{t}H$ signal, and in most samples only a few percent of the events survive the analysis selection, which is described in the next section.

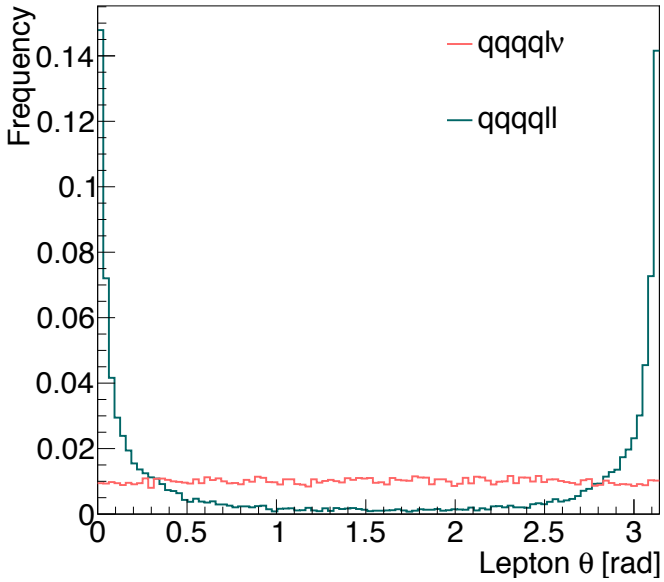


Figure 2: Generator level θ of the true lepton in $qqqqll$ (cyan) and in $qqqql\nu$ (red). All distributions are normalized to unity.

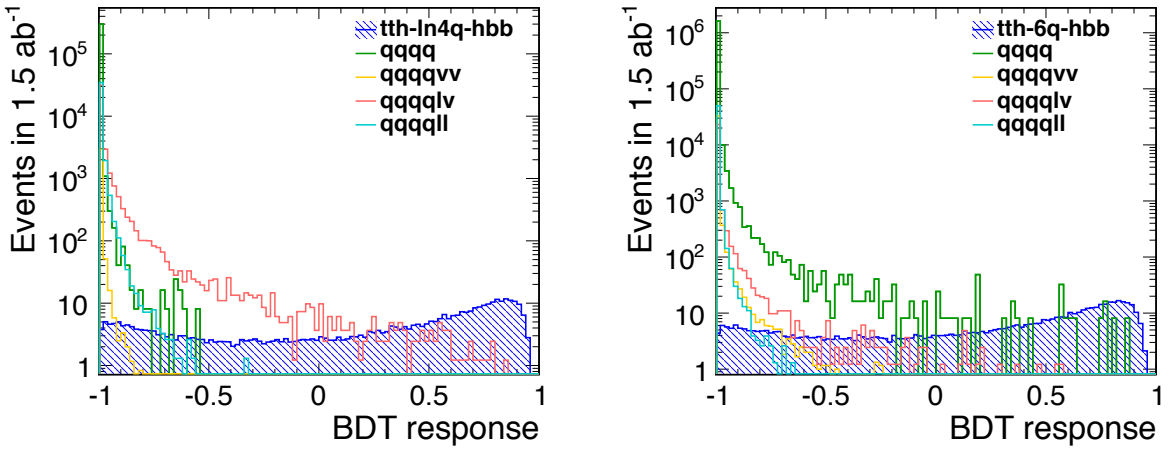
5. Selection efficiency

The cut values on the BDT response in the semi-leptonic and hadronic channles were determined in the original analysis. These correspond to the values for the optimal selection, which maximised the signal significance defined as

$$\frac{S}{\sqrt{S+B}}, \tag{1}$$

where S denotes the number of selected signal events, and B the number of selected background events. This resulted in a maximum significance of 8.36 in the semi-leptonic channel and of 9.17 in the hadronic channel, corresponding to cut values on the BDT response of 0.26 and 0.14, respectively. The BDT

response of the additional samples is shown in Figure 3. Applying the selection on the BDT response with these cuts results in the efficiencies shown in Table 2. The overall signal significance is then recalculated after adding the contribution of the additional samples to the total background expectation.



(a) (b)

Figure 3: BDT response for background samples, scaled to the number of events expected in 1.5 ab^{-1}. (a): Semi-leptonic channel. The signal in this channel (6 jets + 1 lepton) is shown (shaded) for comparison. (b): Hadronic channel. The signal in this channel (8 jets + 0 lepton) is shown (shaded) for comparison.

Table 2: Selection efficiency for each event sample. Column 1 shows the simulated process. Column 2 shows the number of events (and percent) in which 0 leptons were found. Column 3 shows the number (and percent) of these 'zero lepton' events which pass the BDT trained for the hadronic channel. Column 4 shows the number of events (and percent) in which 1 lepton was found. Column 5 shows the number (and percent) of these 'one lepton' events which pass the BDT trained for the semi-leptonic channel.

Process	Zero leptons		Pass HAD BDT		One lepton		Pass SL BDT	
$qqqq$	1.637×10^6	(82%)	195	(1%)	301343	(15%)	0	(0%)
$qqqq\nu\nu$	33760	(91%)	1	(0.3%)	3021	(8%)	0	(0%)
$qqqql\nu$	24812	(36%)	11	(4%)	38893	(57%)	70	(18%)
$qqqqll$	50865	(47%)	1	(0.1%)	37668	(35%)	0	(0%)

6. Results

After the full analysis selection, the samples studied contribute a relatively small number of new events to the total background expectation estimated in the original analysis. The additional backgrounds contribute 70 new events to the semi-leptonic channel and 208 new events to the hadronic channel. Using the updated background expectation, the significance in the semi-leptonic and in the hadronic channel is 8.12 and 8.80, respectively. These can be calculated directly from the yields obtained in the original analysis after adding the contributions of the additional backgrounds [2]. The sensitivity to the cross-section can now be calculated directly from the inverse of the signal significance. Assuming an integrated luminosity of 1.5 ab^{-1}, the cross-section can be measured with an accuracy of 12.31% in the semi-leptonic channel and 11.36% in the hadronic channel. The combined precision of the two channels is 8.35%. This is the error-weighted mean of two independent measurements.

To extract the uncertainty on the top Yukawa coupling (Δg_{ttH}) from the measured uncertainty on the cross-section ($\Delta\sigma(\mathrm{t\bar{t}H})$), the following relation is used [2]:

$$\frac{\Delta g_{\mathrm{ttH}}}{g_{\mathrm{ttH}}} = 0.53 \times \frac{\Delta\sigma(\mathrm{t\bar{t}H})}{\sigma(\mathrm{t\bar{t}H})}\,. \tag{2}$$

Therefore, the uncertainty on the measured cross-section of 8.35% translates into a precision on the top Yukawa coupling of 4.43%.

7. Possible improvements

The original selection was not altered in this analysis of additional backgrounds. The BDTs used here to reject background events were trained only with the background samples considered in the original analysis. A first step in an improved analysis of these additional samples would be the training of the BDT including these backgrounds. Additionally, the 4-jet topology of the *qqqq* sample, which is rich in events with back-to-back W bosons, could be used to further increase the rejection of this background. It would be expected that angular variables such as the $\Delta\phi$ between the two Ws on the plane perpendicular to the beam direction could be useful in separating signal from background in an MVA selection. To investigate the discriminating potential of this and other variables, the particles in *qqqq* were clustered into 4 jets in an attempt to reconstruct events with a recoiling back-to-back topology. The k_{T} algorithm including beam jets [11, 12], implemented in FastJet [13], was used in the exclusive mode to cluster the jets of each event and to reject particles originating from beam-beam backgrounds. The LCFIPlus [14] package was used to re-cluster particles not in beam jets using the Durham algorithm [15]. In both algorithms jet clustering was forced to 4 jets. The same steps were repeated with the two signal samples from the original analysis, and the distributions of $\Delta\phi$ and W mass were compared (see Figure 4). As expected, events in *qqqq* tend to aggregate to high absolute values of $\Delta\phi$, consistent with two recoiling W bosons. Unfortunately, events in the 8-jet signal also concentrate toward high values of $\Delta\phi$, and the hadronic final state is the only one affected by this background after the analysis selection. On the other hand, the shapes of the reconstructed W mass distributions exhibit significant differences between the *qqqq* background sample and the 8 jet signal sample, and this variable could be used to further reduce the contribution of this background to the hadronic channel.

8. Conclusions

The contribution of additional samples to the background expectation in a direct measurement of the top Yukawa coupling was investigated. Only two of the samples studied had non-negligible contributions to the background, with *qqqq* contributing the most events in the hadronic final state. Given that this contribution is relatively small (see Table 2) compared to the total background expectation estimated in the original analysis, no action was taken here to reduce this background. However, the first step in an improved analysis would be to reduce the 195 selected *qqqq* events in the hadronic channel by reconstructing as 4 jets.

Using the updated background estimation, the new uncertainty on the measured cross-section translates into a precision on the top Yukawa coupling of 4.43%. This is a small increase to the 4.27% estimated in the original analysis.

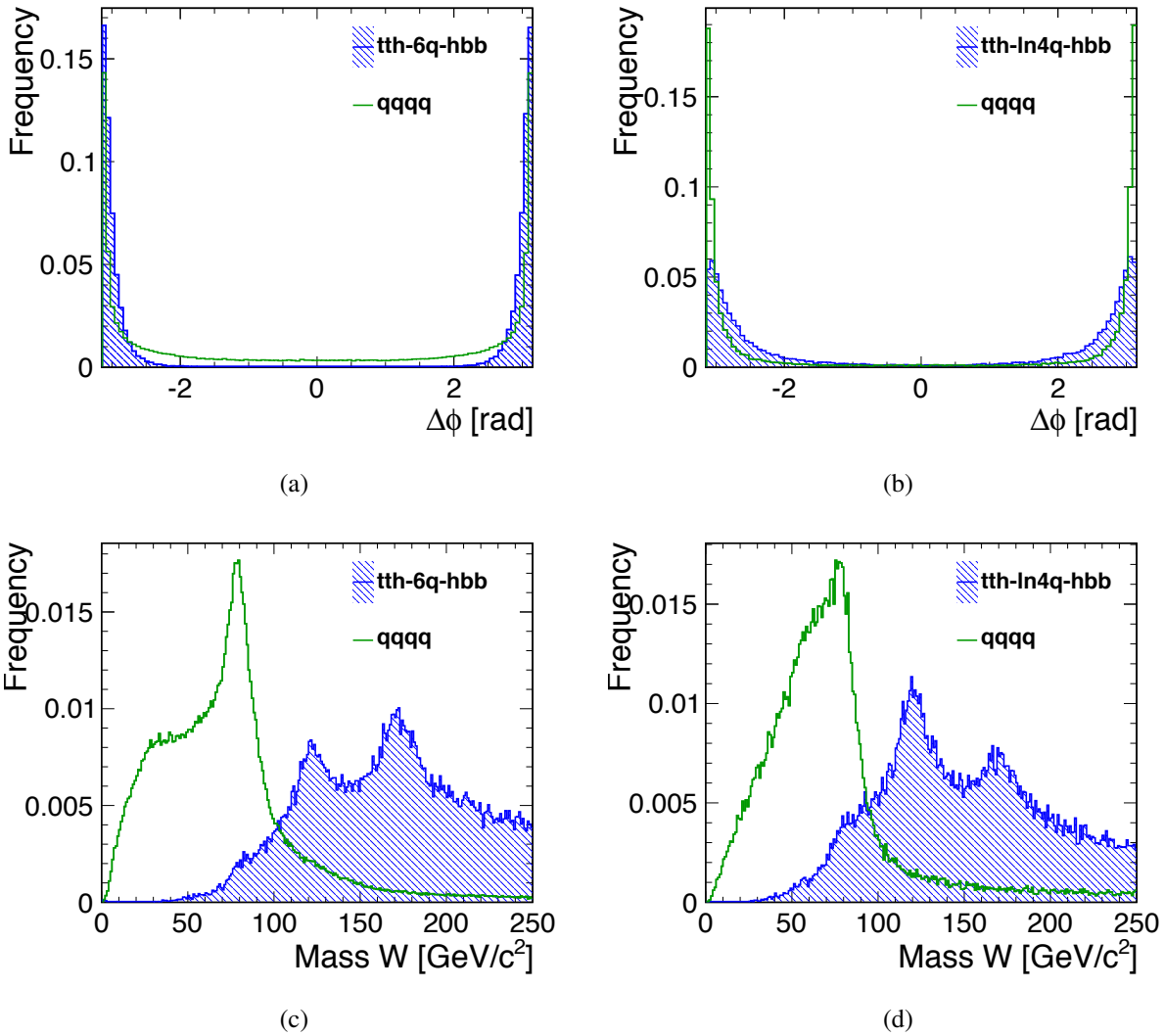
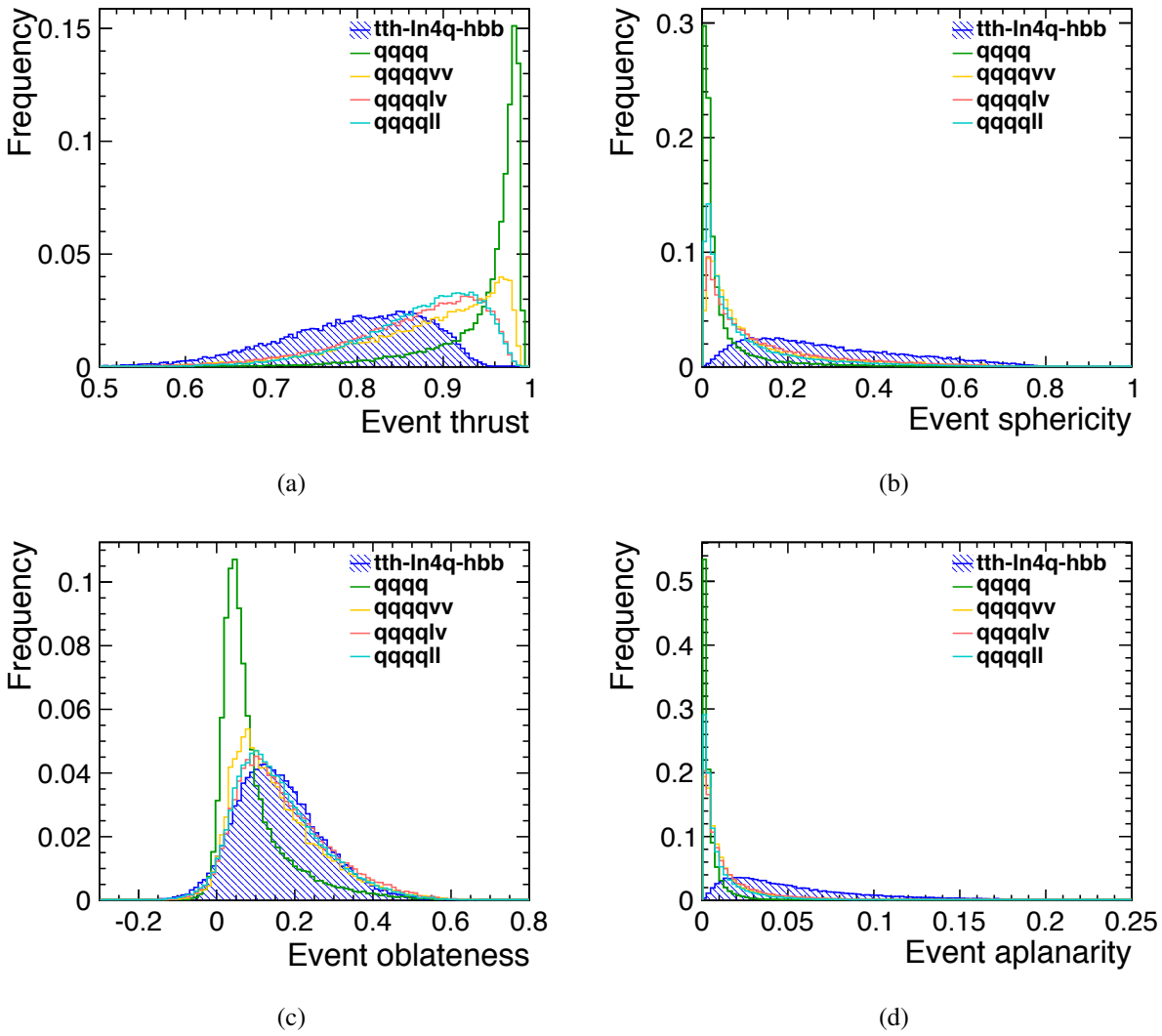


Figure 4: PFOs clustered into 4 jets. $\Delta\phi$ between reconstructed W bosons and number of leptons = 0 (a) and 1 (b). Reconstructed W mass with number of leptons = 0 (c) and 1 (d). All distributions are normalized to unity.

A. Semi-leptonic analysis variables

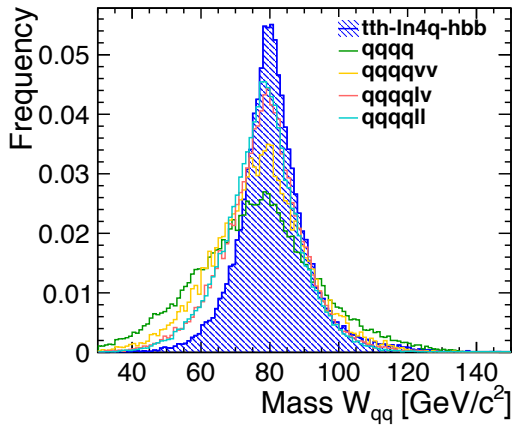
Kinematic variables, b-tags and event shapes in the semi-leptonic channel. The additional backgrounds are compared to the signal sample ($t\bar{t}H$, 4 jet, $H \to b\bar{b}$). All distributions are normalized to unity.



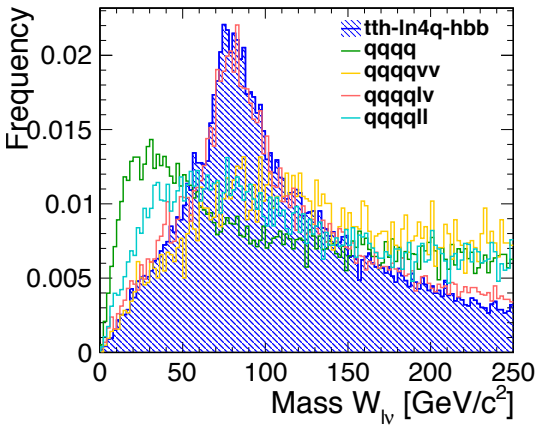
(a) (b)

(c) (d)

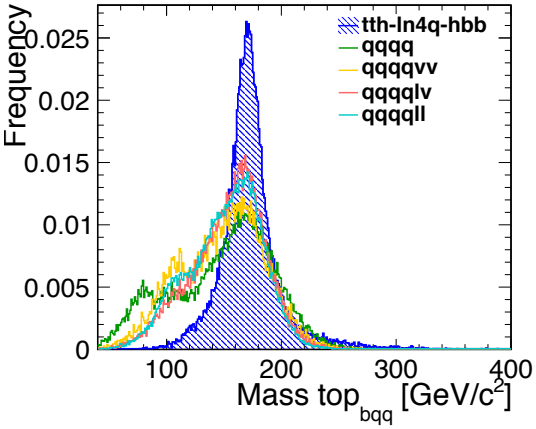
Figure 5: Event shape variable plots, calculated from all reconstructed particles in the event and normalised to unit area.



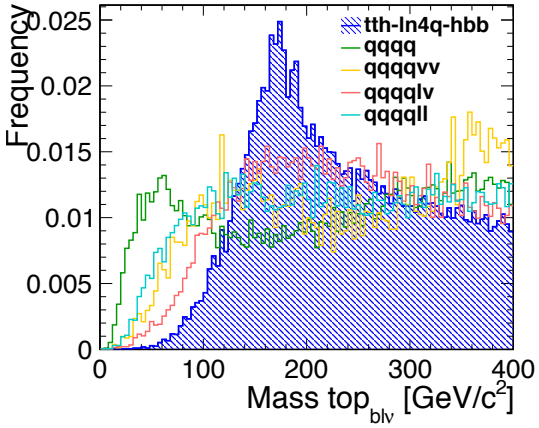
(a)



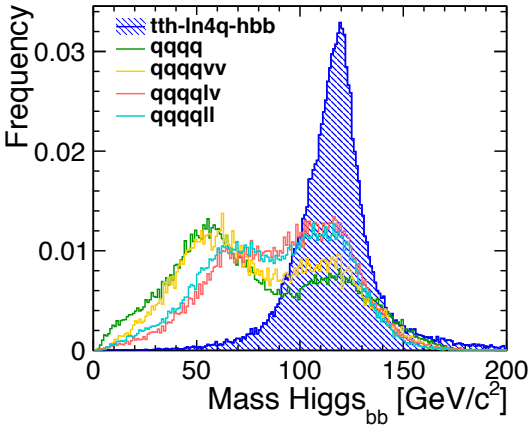
(b)



(c)

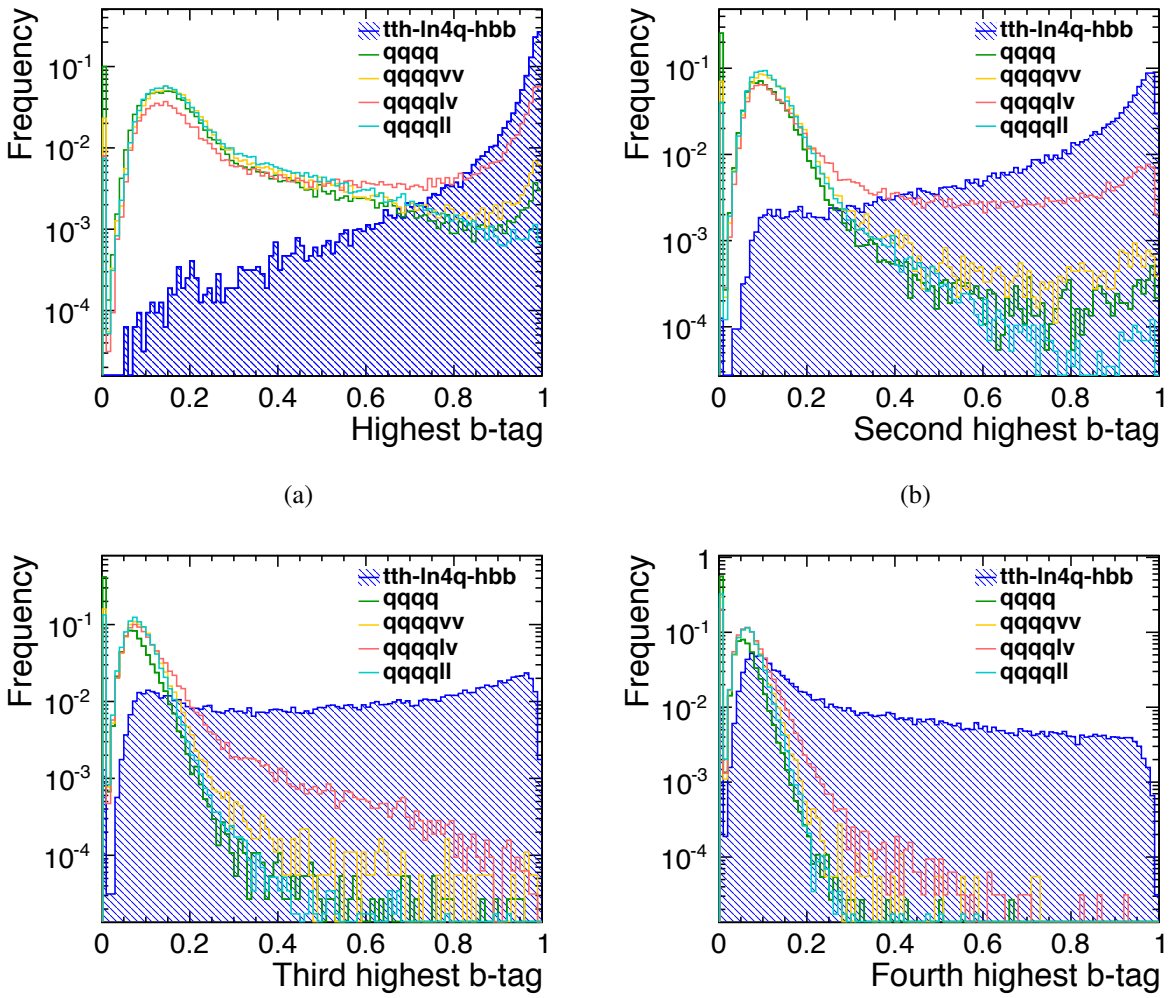


(d)



(e)

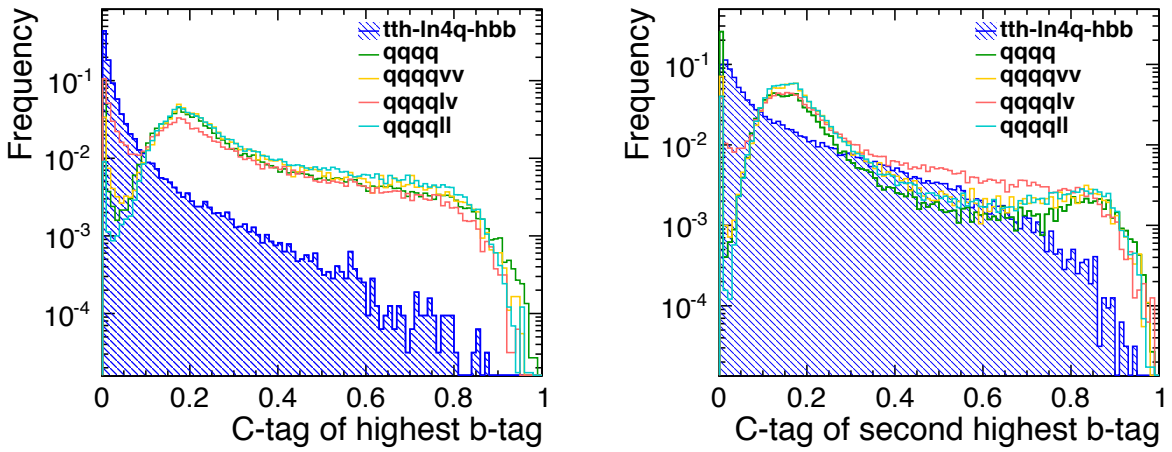
Figure 6: Reconstructed mass variable plots, normalised to unit area.



(a) (b)

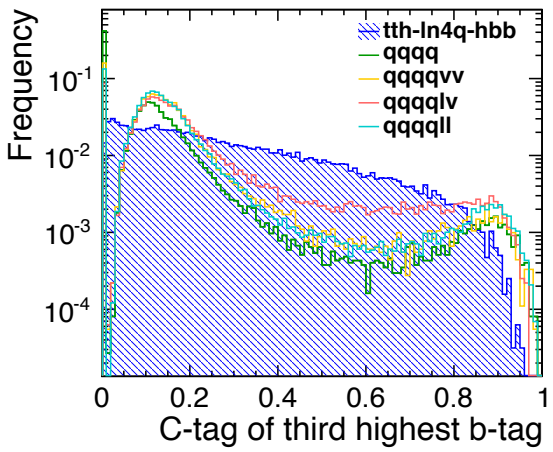
(c) (d)

Figure 7: Four highest b-tag value plots, normalised to unit area.

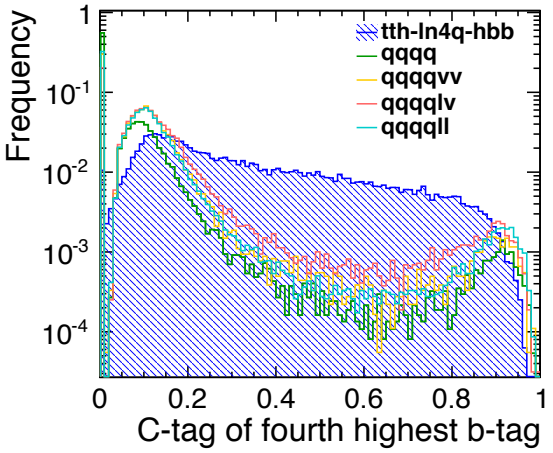


(a)

(b)

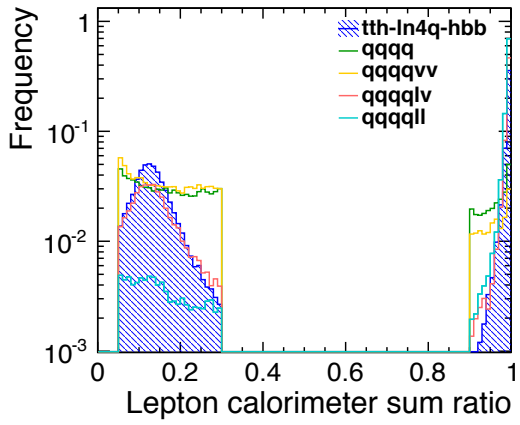


(c)

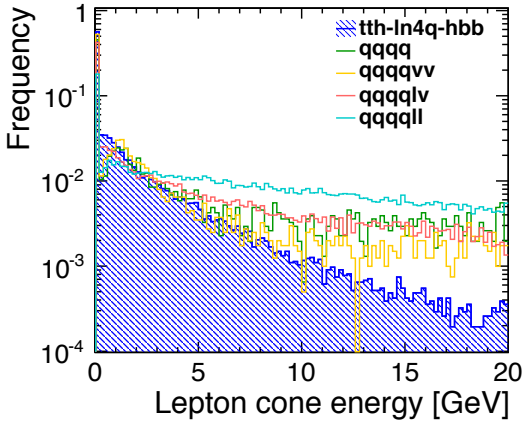


(d)

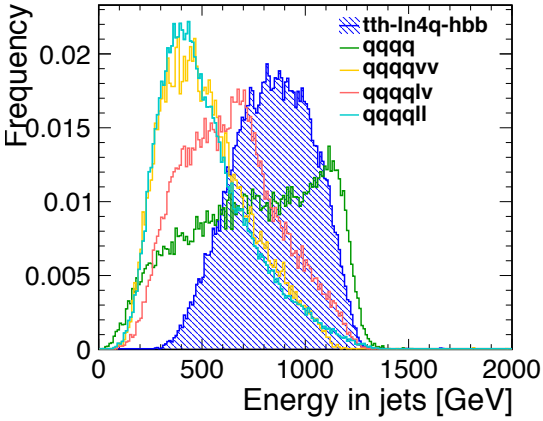
Figure 8: c-tag value plots of the four highest b-tagged jets, normalised to unit area.



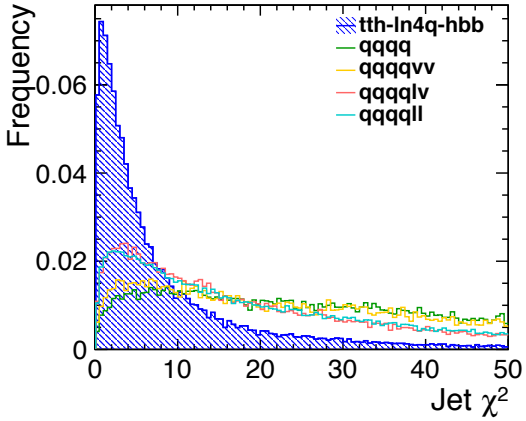
(a)



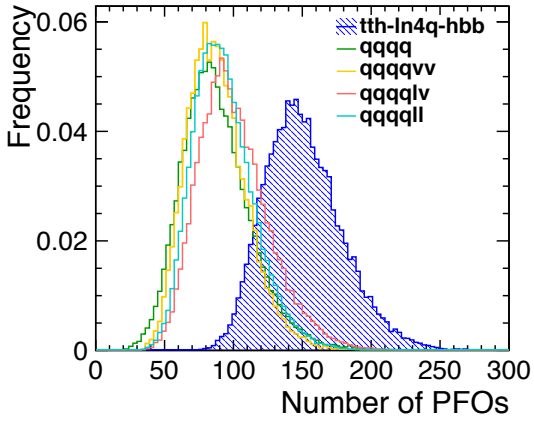
(b)



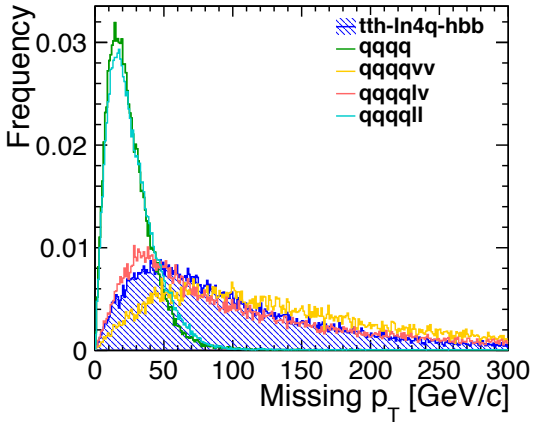
(c)



(d)

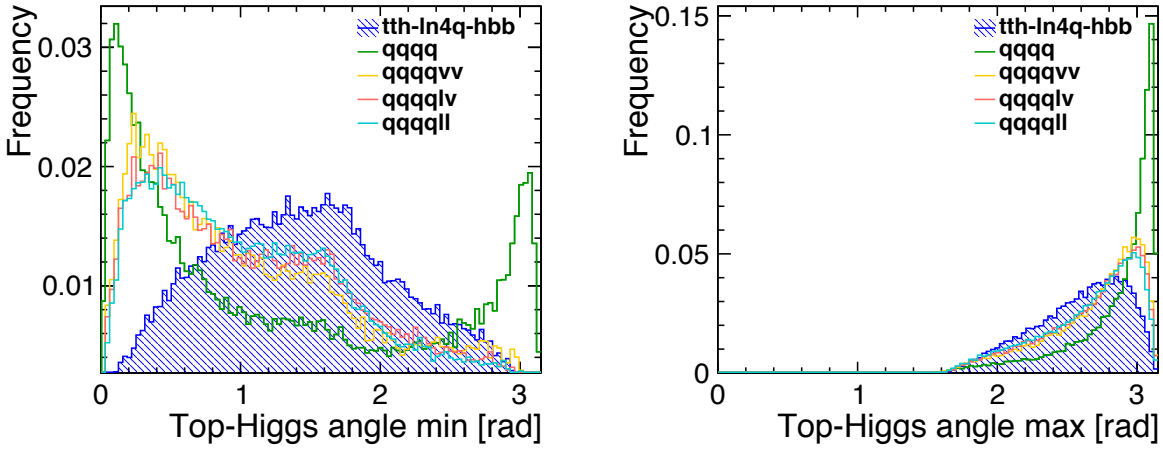


(e)



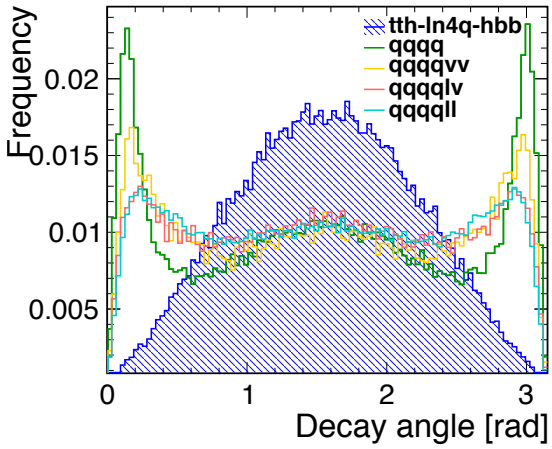
(f)

Figure 9: Lepton and event variable plots, normalised to unit area.



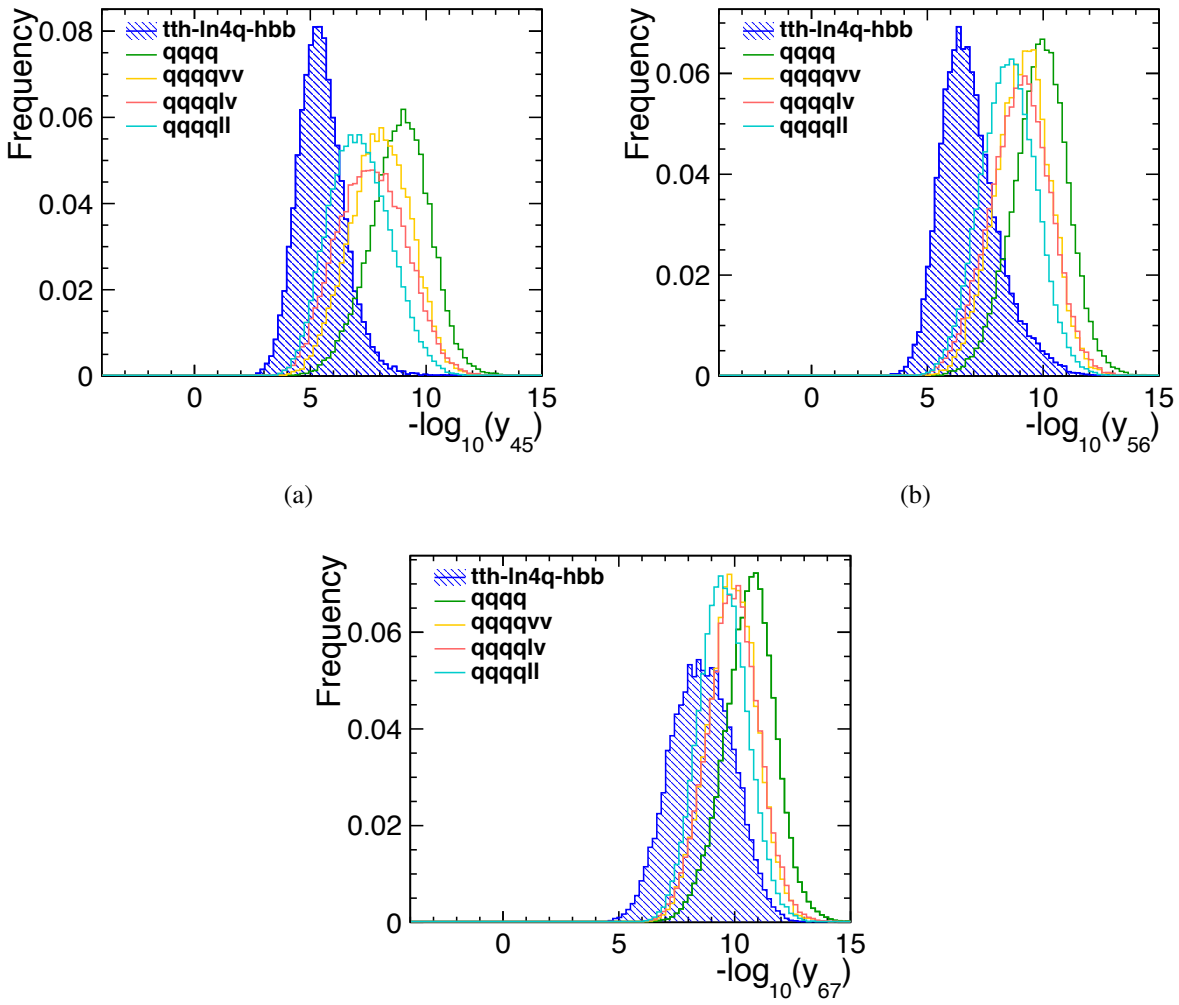
(a)

(b)



(c)

Figure 10: Angular variable plots, normalised to unit area.



(a)

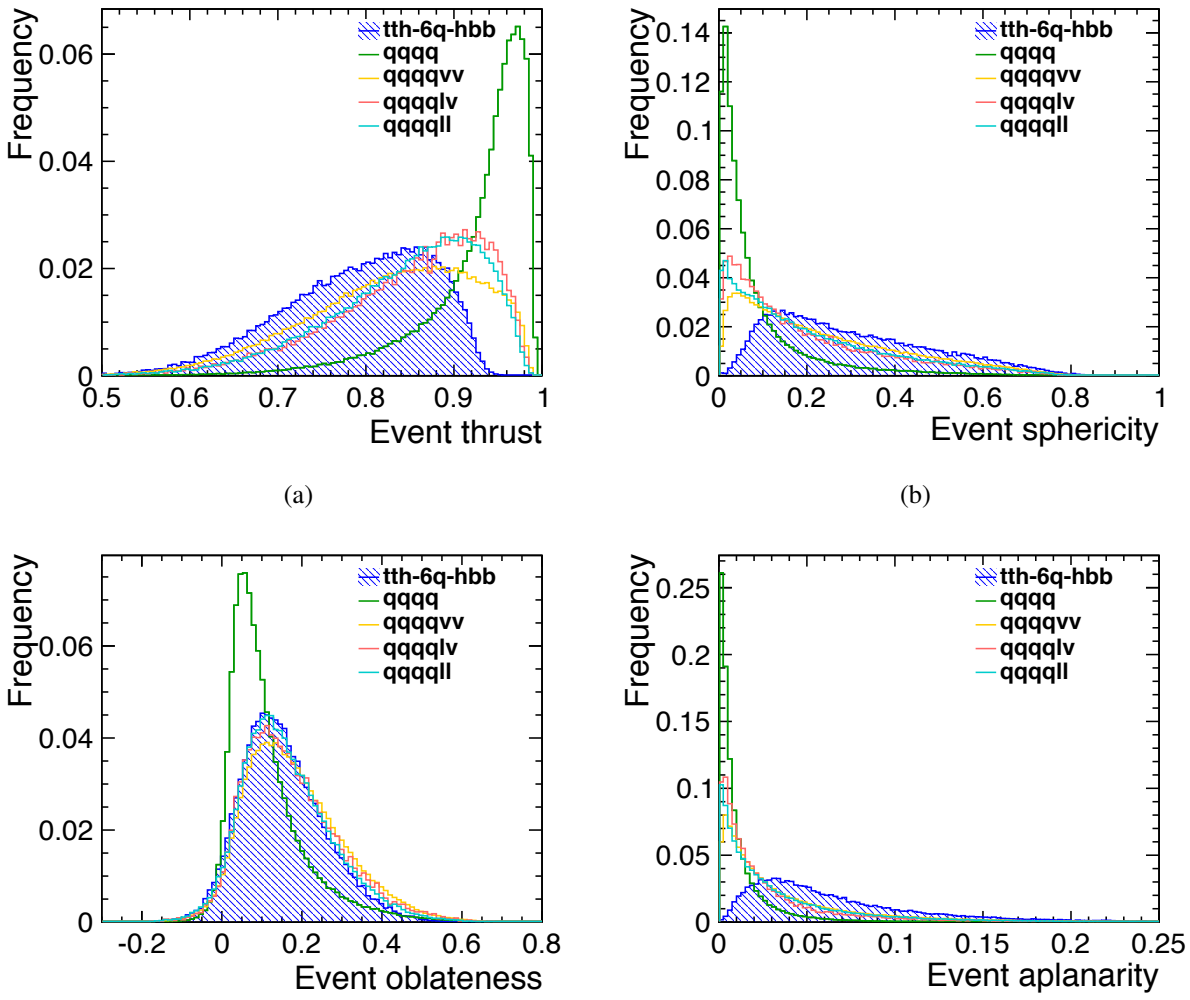
(b)

(c)

Figure 11: Jet transition variable plots, normalised to unit area.

B. Hadronic analysis variables

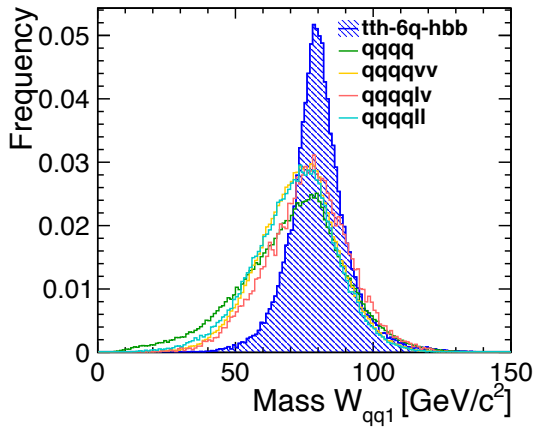
Kinematic variables, b-tags and event shapes in the hadronic channel. The additional backgrounds are compared to the signal sample ($t\bar{t}H$, 6 jet, $H \to b\bar{b}$). All distributions are normalized to unity.



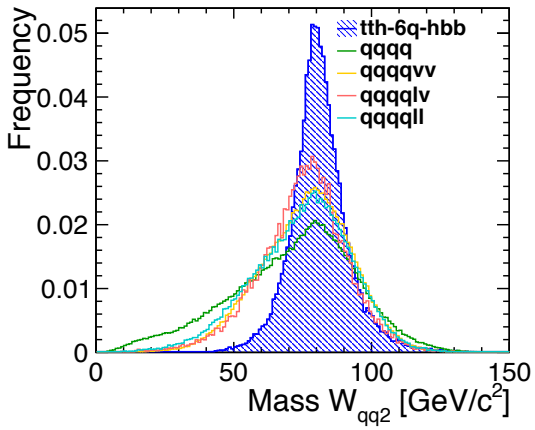
(a) (b)

(c) (d)

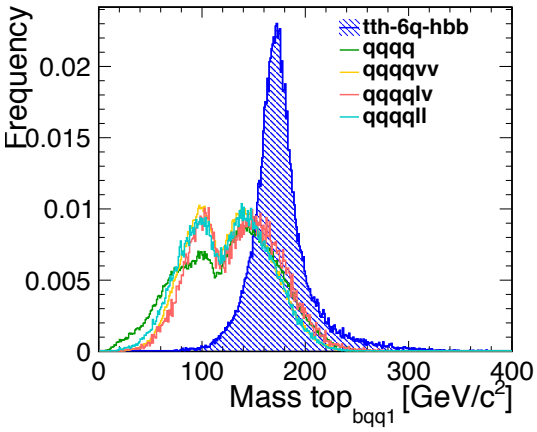
Figure 12: Event shape variable plots, normalised to unit area.



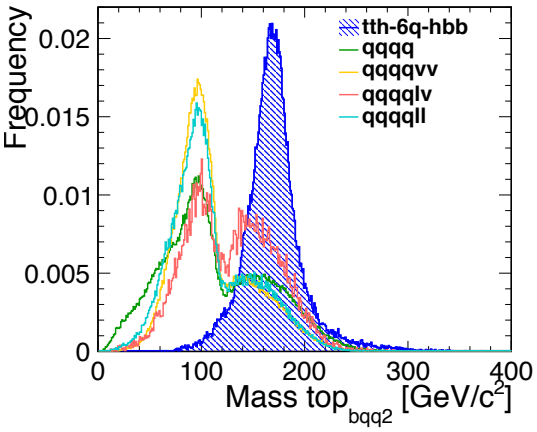
(a)



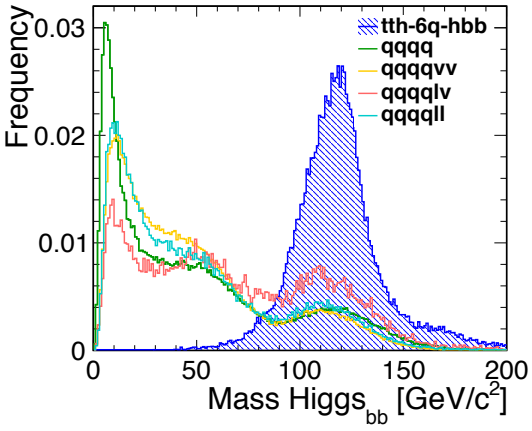
(b)



(c)

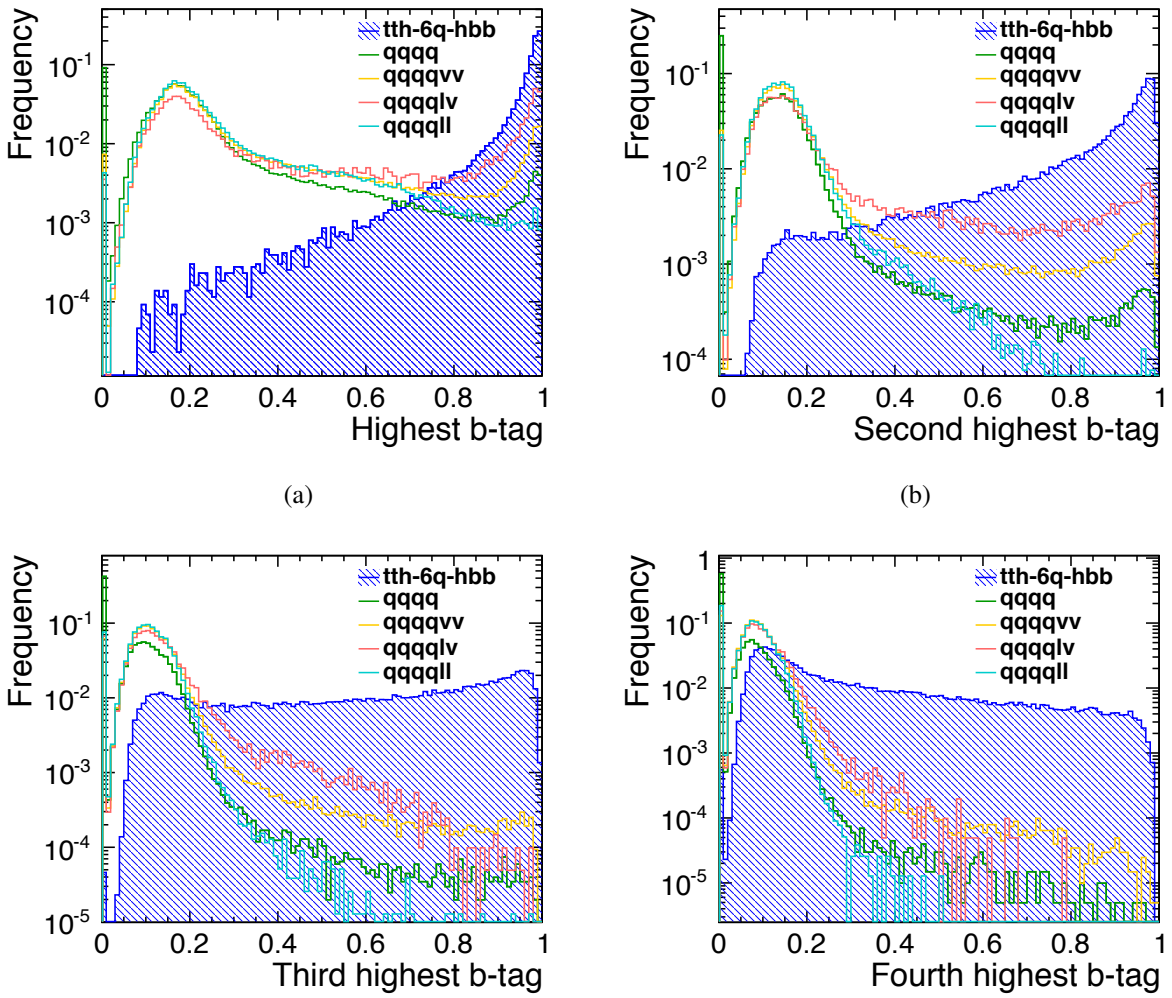


(d)



(e)

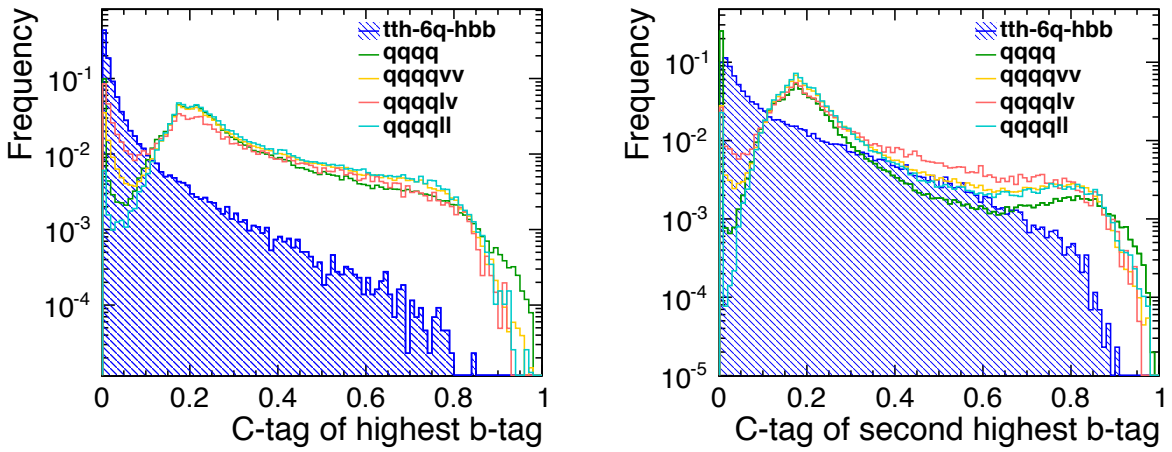
Figure 13: Reconstructed mass variable plots, normalised to unit area.



(a) (b)

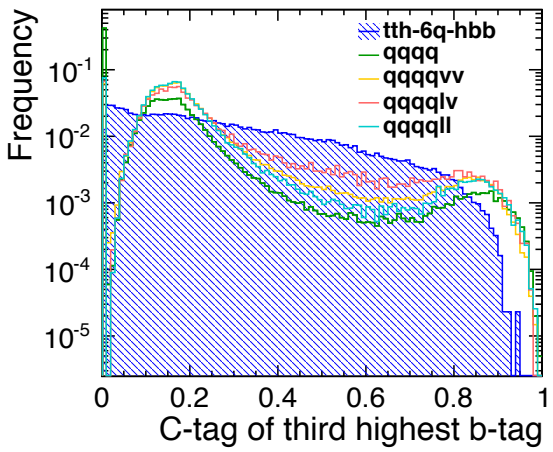
(c) (d)

Figure 14: Four highest b-tag value plots, normalised to unit area.

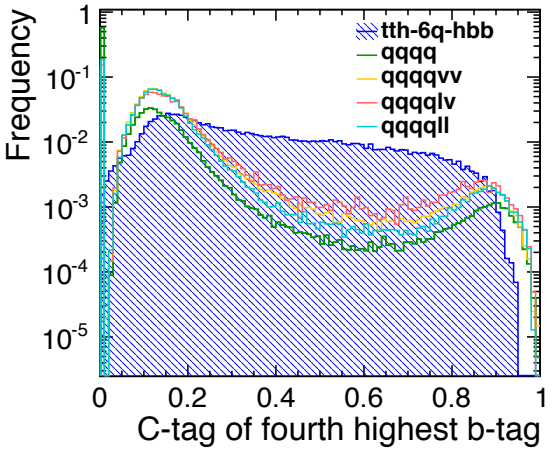


(a)

(b)

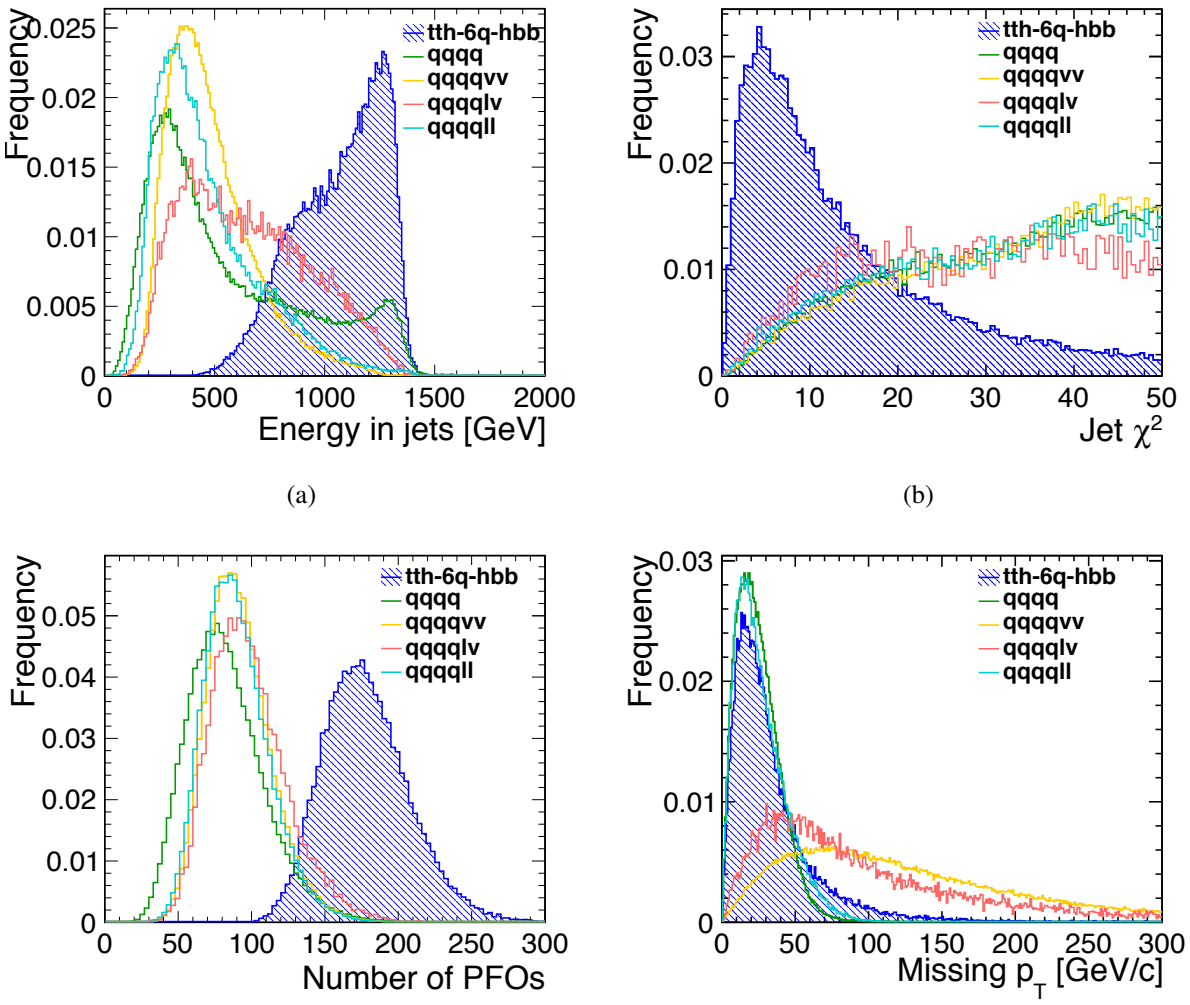


(c)



(d)

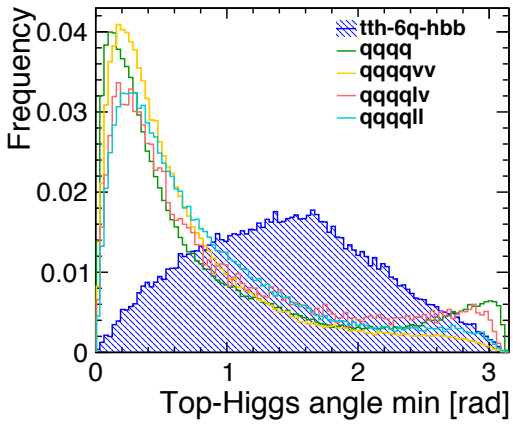
Figure 15: c-tag value plots of the four highest b-tagged jets, normalised to unit area.



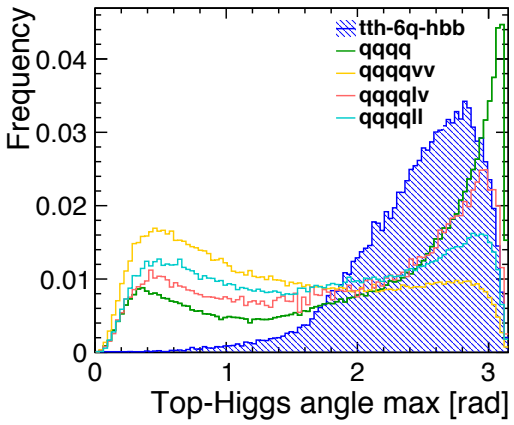
(a) (b)

(c) (d)

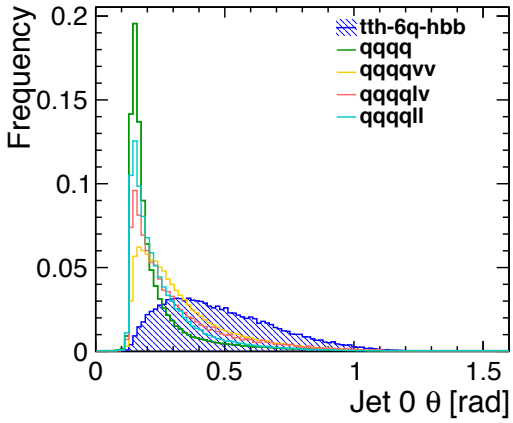
Figure 16: Event variable plots, normalised to unit area.



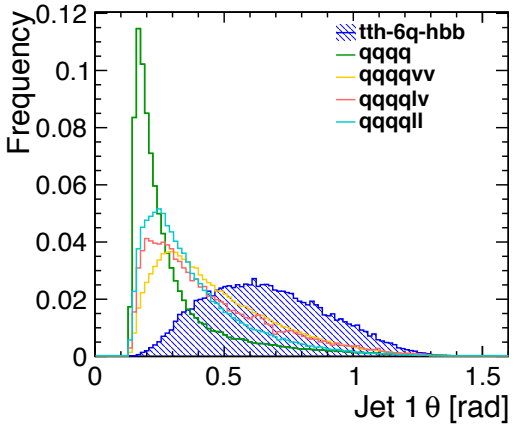
(a)



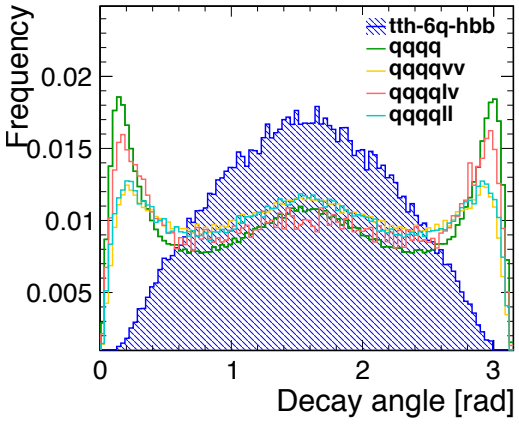
(b)



(c)

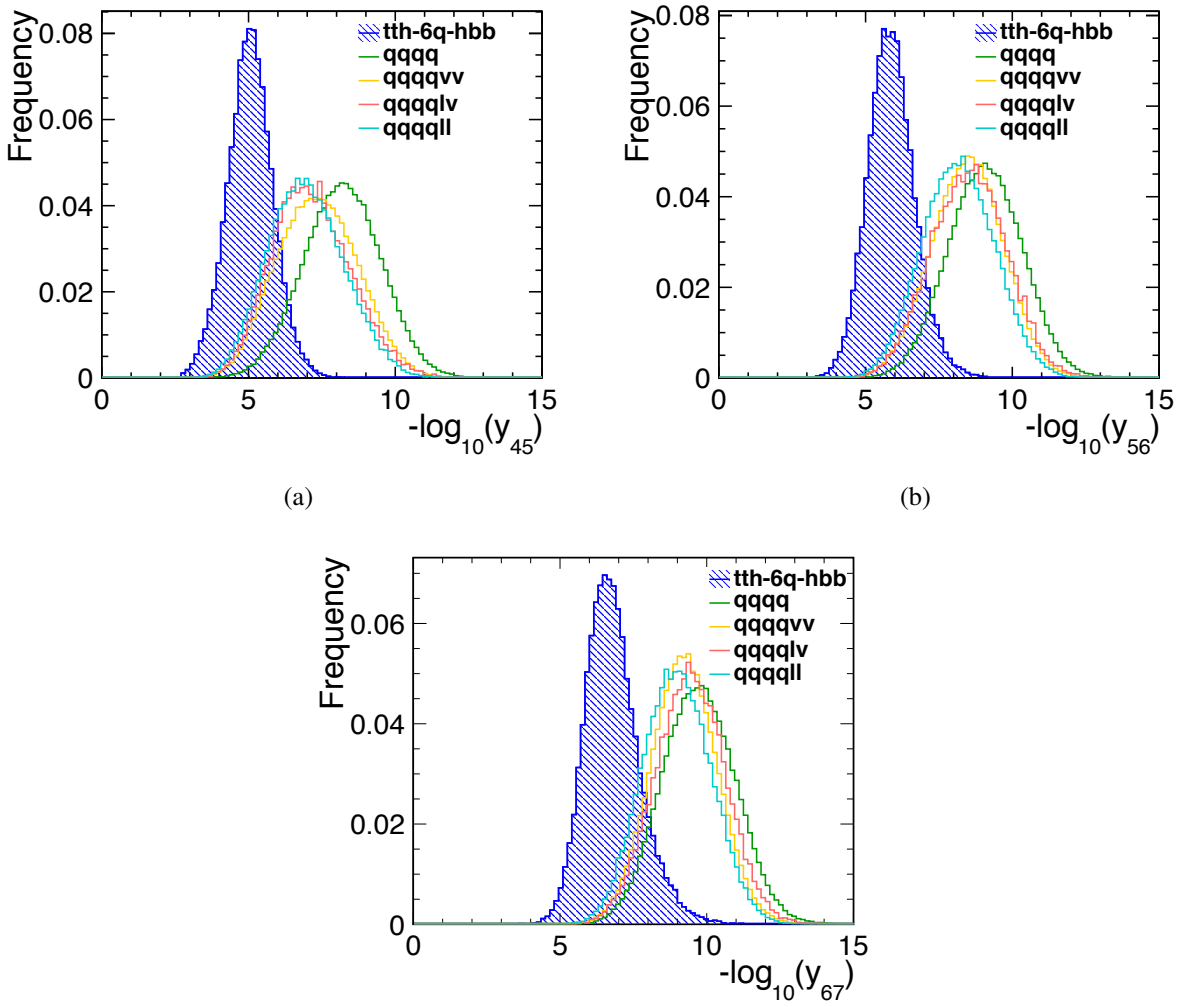


(d)



(e)

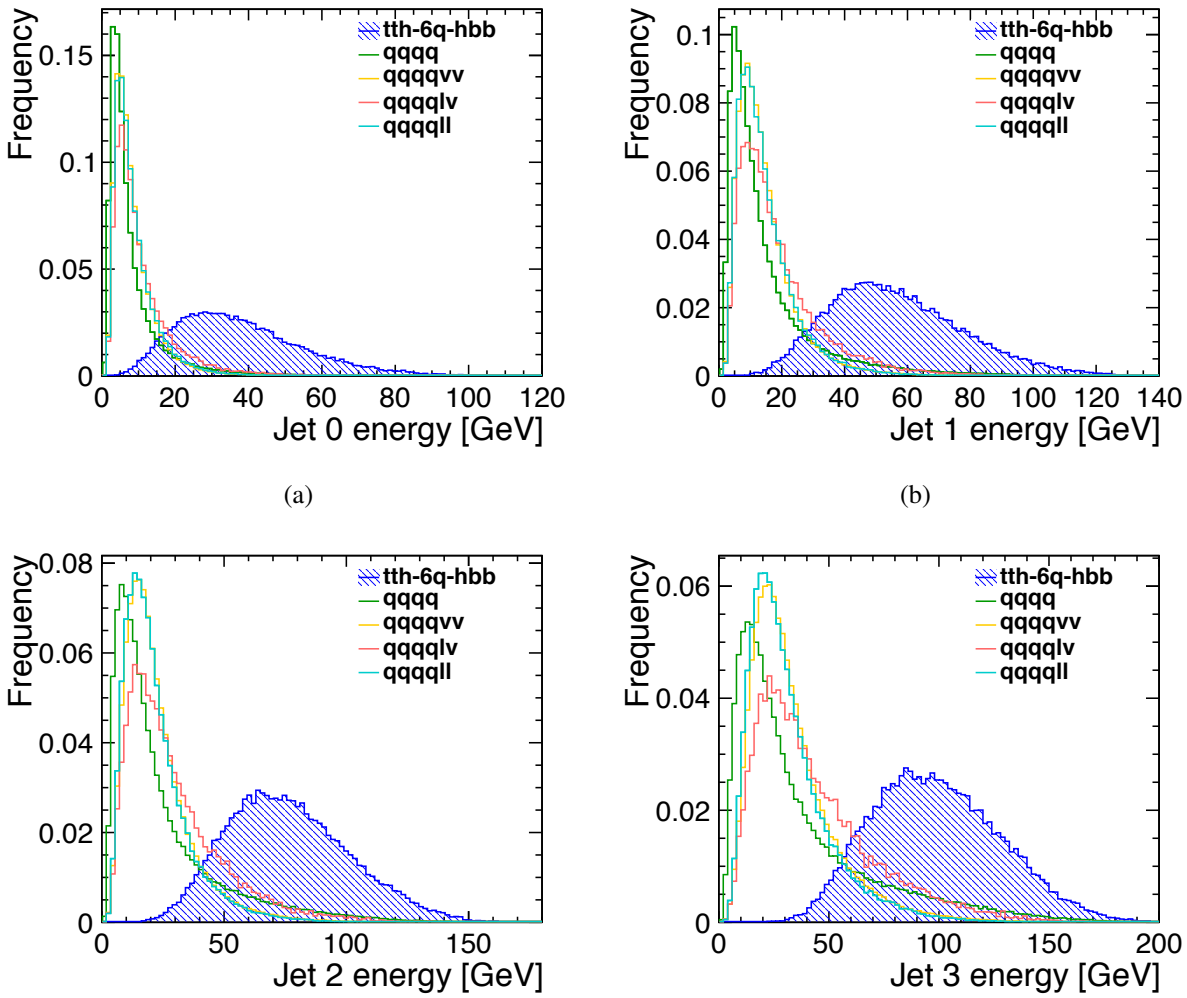
Figure 17: Angular variable plots, normalised to unit area.



(a) (b)

(c)

Figure 18: Jet transition variable plots, normalised to unit area.



(a) (b)

(c) (d)

Figure 19: Jet energy variable plots, normalised to unit area.

References

[1] M. Aicheler et al.,
A Multi-TeV Linear Collider Based on CLIC Technology: CLIC Conceptual Design Report,
tech. rep. CERN-2012-007. SLAC-R-985. KEK-Report-2012-1. PSI-12-01. JAI-2012-001,
Geneva, 2012.

[2] S. Redford, P. Roloff, M. Vogel, *Physics potential of the top Yukawa coupling measurement at a 1.4 TeV Compact Linear Collider using the CLIC SiD detector*, CLICdp-Note 2014-001,
CERN, 2014.

[3] A. Hoecker et al., *TMVA - Toolkit for Multivariate Data Analysis*, POSACAT **040** (2007).

[4] W. Kilian, T. Ohl, J. Reuter, *WHIZARD: Simulating Multi-Particle Processes at LHC and ILC*,
Eur. Phys. J. **C 71** (2011) 1742, DOI: 10.1140/epjc/s10052-011-1742-y,
arXiv: 0708.4233 [hep-ph].

[5] M. Moretti, T. Ohl, J. Reuter, *O'Mega: An Optimizing matrix element generator*,
LC-TOOL 2001-040, 2001, arXiv: hep-ph/0102195 [hep-ph].

[6] T. Sjostrand, S. Mrenna, P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **0605** (2006) 026,
DOI: 10.1088/1126-6708/2006/05/026, arXiv: hep-ph/0603175 [hep-ph].

[7] A. Miyamoto et al., *Physics and Detectors at CLIC: CLIC Conceptual Design Report*, tech. rep.
arXiv:1202.5940. CERN-2012-003. ANL-HEP-TR-12-01. DESY-12-008. KEK-Report-2011-7,
Comments: 257 p, published as CERN Yellow Report CERN-2012-003, Geneva, 2012.

[8] C. Grefe, A. Muennich,
The CLIC SiD CDR Detector Model for the CLIC CDR Monte Carlo Mass Production,
LCD-Note 2011-009, 2011.

[9] S. Agostinelli et al., *Geant4 – a simulation toolkit*,
Nucl. Instrum. Methods Phys. Res., Sect. **A 506** (2003) 250,
URL: http://www.sciencedirect.com/science/article/B6TJM-48TJFY8-5/2/23ea98096ce11c1be446850c04cfa498.

[10] J. Allison et al., *Geant4 developments and applications*, IEEE T. Nucl. Sci. **53** (2006) 270,
DOI: 10.1109/TNS.2006.869826.

[11] S. Catani et al., *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*,
Nucl. Phys. **B 406** (1993) 187, DOI: 10.1016/0550-3213(93)90166-M.

[12] S. Ellis, D. Soper, *Successive combination jet algorithm for hadron collisions*,
Phys. Rev. **D 48** (1993) 3160, DOI: 10.1103/PhysRevD.48.3160,
arXiv: hep-ph/9305266 [hep-ph].

[13] M. Cacciari, G. Salam, G. Soyez, *FastJet User Manual*, Eur. Phys. J. **C 72** (2012) 1896,
DOI: 10.1140/epjc/s10052-012-1896-2, arXiv: 1111.6097 [hep-ph].

[14] *LCFIPlus*, https://confluence.slac.stanford.edu/display/ilc/LCFIPlus.

[15] S. Catani et al., *New clustering algorithm for multijet cross sections in e^+e^- annihilation*,
Phys. Lett. **B 269** (1991) 432, ISSN: 0370-2693,
DOI: http://dx.doi.org/10.1016/0370-2693(91)90196-W, URL: http://www.sciencedirect.com/science/article/pii/037026939190196W.