DØ Note 6087-CONF

Search for the standard-model Higgs boson in the $ZH \to b\bar{b}\nu\bar{\nu}$ channel in 6.4 fb^{-1} of $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV

The DØ Collaboration
URL: http://www-d0.fnal.gov
(Dated: August 30, 2010)

A search is performed for the standard-model Higgs boson in 6.4 fb^{-1} of $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV, collected with the D0 detector at the Fermilab Tevatron Collider. The final state considered is a pair of b jets and an imbalance in transverse energy, as expected from $p\bar{p} \to ZH \to \nu\bar{\nu}b\bar{b}$ production. The search is also sensitive to the $WH \to \ell\nu b\bar{b}$ channel when the charged lepton is not identified. For a Higgs-boson mass of 115 GeV, a limit is set at the 95% C.L. on the cross section multiplied by branching fraction $[p\bar{p} \to (Z/W)H](H \to b\bar{b})$ that is a factor of 3.4 larger than the standard-model value, consistent with the expected factor of 4.2.

Preliminary Results for Summer 2010 Conferences

I. INTRODUCTION

The existence of the Higgs boson is the only fundamental element of the standard model (SM) that has yet to be confirmed. Its observation would be a key step in establishing the mechanism of electroweak symmetry breaking and mass generation. Associated ZH production in $p\bar{p}$ collisions, with $Z \to \nu\bar{\nu}$ and $H \to b\bar{b}$, is among the most sensitive processes for seeking a Higgs boson with a mass $m_H \lesssim 135$ GeV at the Fermilab Tevatron Collider [1]. The D0 Collaboration published a search for this process based on 5.2 fb^{-1} of integrated luminosity [2]. A lower limit of 114.4 GeV was set by the LEP experiments on the mass of the Higgs boson from searches for the reaction $e^+e^- \to ZH$ [3], while an indirect upper limit of 157 GeV can be inferred from precision electroweak data [4]. These limits and those given below are all defined at the 95% level of confidence (C.L.).

This note presents a new search using an integrated luminosity of 1.3 fb^{-1} beyond that used in Ref. [2]. The final-state topology consists of a pair of b jets from $H \to b\bar{b}$ decay and missing transverse energy ($\not\!\!E_T$) from $Z \to \nu\bar{\nu}$. The search is therefore also sensitive to the WH process when the charged lepton from $W \to \ell\nu$ decay is not identified. The main backgrounds arise from (W/Z)+heavy-flavor jets (jets initiated by b or c quarks), top-quark production, and multijet (MJ) events with $\not\!\!E_T$ arising from mismeasurement of jet energies.

II. DATA AND SIMULATED SAMPLES

The D0 detector is described in [5]. The data used in this analysis were recorded using triggers designed to select events with jets and $\not\!\!E_T$ [6]. The data taking period prior to March 2006 is refered to as Run IIa, while IIb denotes the period after. This division corresponds to the installation of an additional silicon vertex detector, trigger upgrades, and a significant increase in the rate of delivered luminosity. For the result described in this note the Run IIb dataset was increased in integrated luminosity and the double b tag sample was re-analyzed. The Run IIa part of the analysis and the single b tag Run IIb sample is taken from [2]. This note focusses solely on the re-analysis of the upgraded double b tag Run IIb analysis, but the limits presented incorporate also the Run IIa result and single b tag Run IIb result of [2]. After imposing data quality requirements, and combining Run IIa and Run IIb data, the total integrated luminosity [7] is 6.4 fb^{-1}for double b tag and 5.2 fb^{-1}for single b tag.

The analysis relies on (i) charged particle tracks, (ii) calorimeter jets reconstructed in a cone of radius 0.5, using the iterative midpoint cone algorithm [8], and (iii) electrons or muons identified through the association of tracks with electromagnetic calorimeter clusters or with hits in the muon detector, respectively. The $\not\!\!E_T$ is reconstructed as the opposite of the vectorial sum of transverse components of energy deposits in the calorimeter and is corrected for identified muons. Jet energies are calibrated using transverse energy balance in photon+jet events [9], and these corrections are propagated to the $\not\!\!E_T$.

Backgrounds from SM processes are determined through Monte Carlo simulation, while the instrumental MJ background is estimated from data. Events from (W/Z)+jets processes are generated with ALPGEN [10], interfaced with PYTHIA [11] for initial and final-state radiation and for hadronization. The p_T spectrum of the Z is reweighted to match the D0 measurement [12]. The p_T spectrum of the W is reweighted using the same experimental input, corrected for the differences between the Z and W p_T spectra predicted in next-to-next-to-leading order (NNLO) QCD [13]. For $t\bar{t}$ and electroweak single top quark production, the ALPGEN and COMPHEP [14] generators, respectively, are interfaced with PYTHIA, while vector boson pair production is generated with PYTHIA. The ZH and WH signal processes are generated with PYTHIA for Higgs boson masses (m_H) from 100 to 150 GeV, in 5 GeV steps. All these simulations use CTEQ6L1 parton distribution functions (PDFs) [15].

The absolute normalizations for (W/Z)+jets production are obtained from NNLO calculations of total cross sections based on [16], using the MRST2004 NNLO PDFs [17]. The heavy-flavor fractions are obtained using MCFM [18]. Cross sections for other SM backgrounds are taken from [19], or calculated with MCFM, and the cross sections for signal are taken from [20].

Signal and background samples are passed through a full GEANT3-based simulation [21] of the detector response and processed with the same reconstruction program as used for data. Events from randomly selected beam crossings are overlaid on simulated events to account for detector noise and contributions from additional $p\bar{p}$ interactions. Parametrizations of trigger efficiency are determined using events collected with independent triggers based on information from the muon detectors. Weight factors compensating for residual differences between data and simulation are applied for electron, muon and jet identification. Jet energy calibration and resolution are adjusted in simulated events to match those measured in data.

III. EVENT SELECTION

A preselection that greatly reduces the overwhelming background from multijet events is performed as follows. The primary vertex must be reconstructed within the acceptance of the silicon vertex detector, and at least three tracks must originate from that vertex. Jets with associated tracks (using only tracks that meet minimal quality criteria to ensure that the b-tagging algorithm operates efficiently) are denoted as "taggable" jets. The leading (highest p_T) jet must be taggable with no more than three taggable jets in total. These jets must have transverse momentum $p_T > 20$ GeV and pseudorapidity $|\eta| < 2.5$ [22]. The two leading jets must not be back-to-back in the plane transverse to the beam direction: $\Delta\phi(\text{jet}_1, \text{jet}_2) < 165°$. Finally, $\not\!\!E_T > 30$ GeV is required.

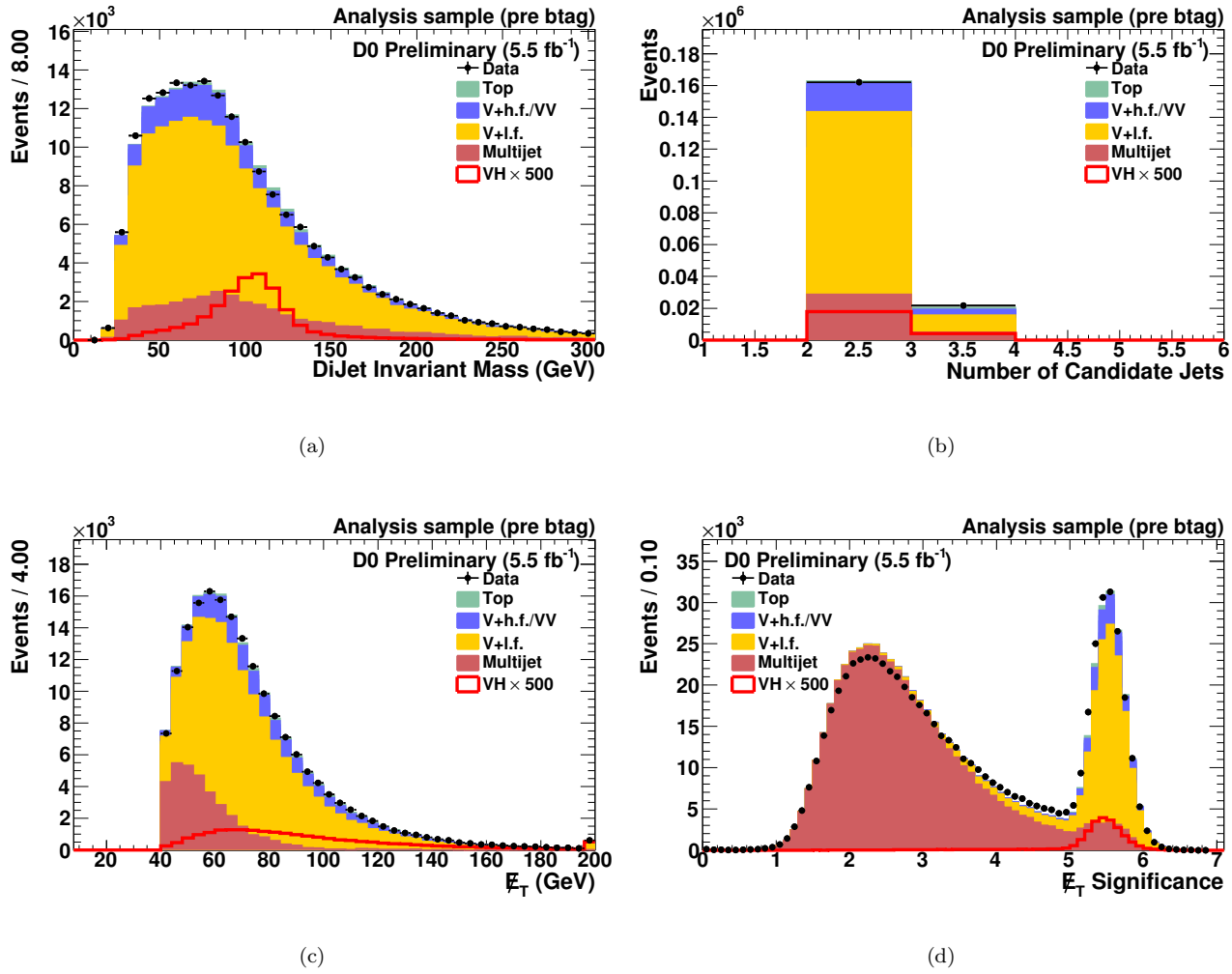


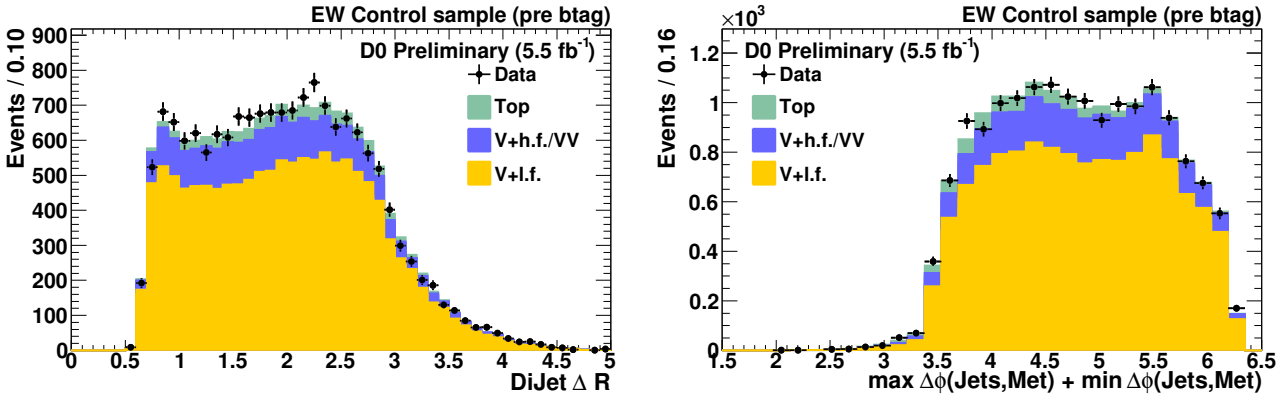
FIG. 1: Representative variable distributions in the analysis sample before the multijet veto and before b tagging: (a) Dijet invariant mass, (b) Taggable jet multiplicity, (c) Missing E_T, (d) Missing E_T significance without the requirement that it be larger than 5. The distributions for signal (VH) are multiplied by a factor of 500 and include ZH and WH production for $m_H = 115$ GeV.

Additional selection criteria define four distinct samples: (i) an analysis sample used to search for a Higgs boson signal, (ii) an electroweak (EW) control sample, enriched in $W(\to \mu\nu)$+jets events where the jet system has a topology similar to that of the analysis sample, that is used to validate the SM background simulation, (iii) a "MJ-model" sample, dominated by multijet events, used to model the MJ background in the analysis sample, and (iv) a large "MJ-enriched" sample, used to validate this modeling procedure.

The analysis sample is selected by requiring $\not\!\!E_T > 40$ GeV and a measure of the $\not\!\!E_T$ significance $\mathcal{S} > 5$ [23]. Larger values of $\mathcal{S}$ correspond to $\not\!\!E_T$ values that are less likely to be caused by fluctuations in jet energies. In signal events, the missing track p_T, $\not\!\!p_T$, defined as the opposite of the vectorial sum of the charged particle transverse momenta,

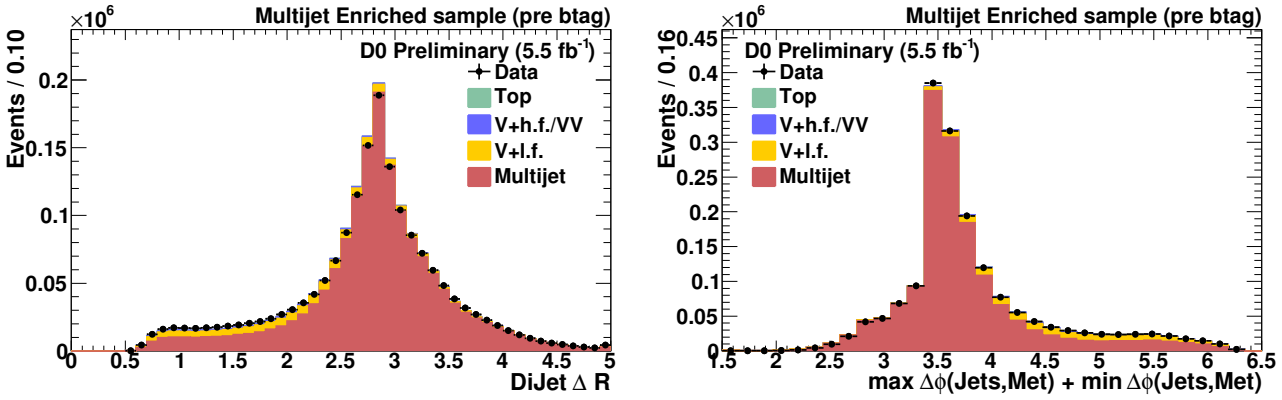
is expected to point in a direction close to that of $\not\!\!E_T$. Such a strong correlation is not expected in multijet events, where $\not\!\!E_T$ originates mainly from mismeasurement of jet energies. Advantage is taken of this feature by requiring $\mathcal{D} < \pi/2$, where $\mathcal{D} = \Delta\phi(\not\!\!E_T, \not\!\!p_T)$. Events containing an isolated electron or muon [24] with $p_T > 15$ GeV are rejected to reduce backgrounds from W+jets, top quark, and diboson production. Distributions in the analysis sample are shown in Fig. 1.

The EW-control sample is selected in a way similar to the analysis sample, except that an isolated muon with $p_T > 15$ GeV is required. The multijet content of this sample is rendered negligible by requiring the transverse mass of the muon and $\not\!\!E_T$ system to be larger than 30 GeV. To ensure similar jet topologies for the analysis and EW-control samples, $\not\!\!E_T$, not corrected for the selected muon, is required to exceed 40 GeV. Excellent agreement with the SM expectation is found for the number of selected events. The agreement for all kinematic distributions is also very good once a reweighting of the distribution of $\Delta\eta$ between the two leading taggable jets is performed, as suggested by a simulation of (W/Z)+jets using the SHERPA generator [25]. Distributions in the EW-control sample are shown in Fig. 2.



(a) (b)

FIG. 2: Representative variable distributions in the EW-contol sample before b tag requirement: (a) Dijet ΔR, (b) $\max \Delta\phi(\not\!\!E_T, \mathrm{jet_i}) + \min \Delta\phi(\not\!\!E_T, \mathrm{jet_i})$.



(a) (b)

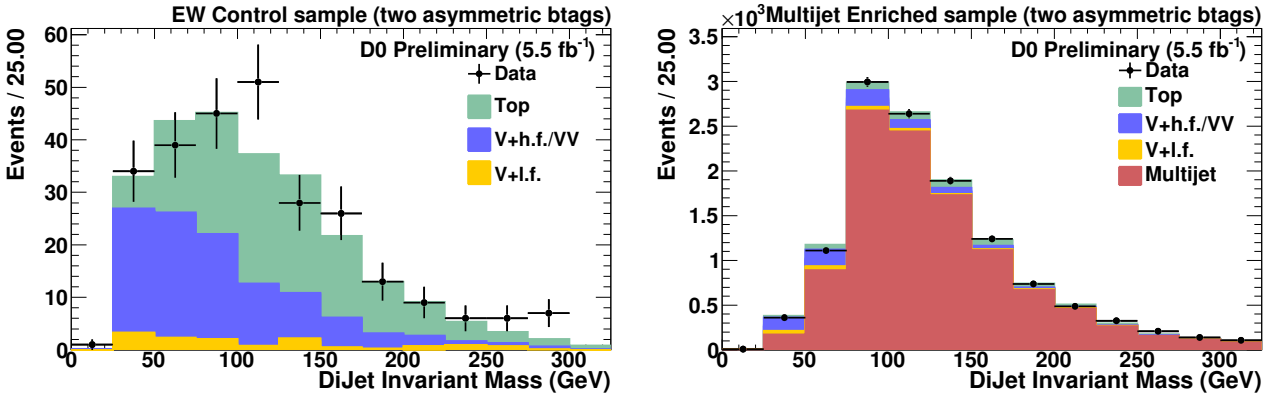
FIG. 3: Representative variable distributions in the MJ-enriched sample before b tag requirement: (a) Dijet ΔR, (b) $\max \Delta\phi(\not\!\!E_T, \mathrm{jet_i}) + \min \Delta\phi(\not\!\!E_T, \mathrm{jet_i})$.

The MJ-model sample, used to determine the MJ background, is selected as the analysis sample, except that the requirement of $\mathcal{D} < \pi/2$ is inverted. The small contribution from non-MJ SM processes in the $\mathcal{D} > \pi/2$ region

is subtracted, and the resulting sample is used to model the MJ background in the analysis sample. After adding contributions from SM backgrounds, the MJ background is normalized so that the expected number of events is identical to the number observed in the analysis sample.

The MJ-enriched sample is used to test the validity of this approach and is defined as the analysis sample, except that the $\not\!\!E_T$ threshold is reduced to 30 GeV and no requirement is imposed on $\mathcal{S}$. As a result, the MJ background dominates the entire range of $\mathcal{D}$ values, and this sample is used to verify that the events with $\mathcal{D} > \pi/2$ correctly model those with $\mathcal{D} < \pi/2$. Distributions in the MJ-enriched sample are shown in Fig. 3.

The large branching fraction for $H \to b\bar{b}$ is exploited by requiring that both of the two leading taggable jets be b tagged. The double-tag sample is selected with asymmetric requirements on the outputs of a b-tagging neural network algorithm [26], such that one jet is tagged with an efficiency of $\approx 70\%$ ("loose tag"), and the other with an efficiency of $\approx 50\%$ ("tight tag"). These values apply for taggable jets with $p_T \approx 45$ GeV and $|\eta| \approx 0.8$. The mistag rates , i.e., the probablilities to tag light (u, d, s, g) jets as b jets, are $\approx 6.5\%$ and $\approx 0.5\%$ for the loose and tight tags, respectively. The flavor-dependent b-tagging efficiencies are adjusted in simulated events to match those measured in dedicated data samples. Fig. 4 shows the dijet invariant mass for the control samples for the double tag selections.



(a) (b)

FIG. 4: (a) Dijet invariant mass with double b tag in the EW-control sample, (b) Dijet invariant mass with double b tag in the MJ-control sample.

IV. ANALYSIS USING DECISION TREES

A boosted decision tree (DT) technique [27] takes advantage of different kinematics in signal and background processes. For each m_H, a "MJ DT" (multijet-rejection DT), used to discriminate between signal and MJ-model events, is trained using 20 kinematic variables before b tagging is applied. These variables include the number of jets, jet p_T, dijet p_T, $\not\!\!E_T$, angles between jets, between second jet and $\not\!\!E_T$, and the dijet mass, where the dijet system is defined by the two leading taggable jets. New variables in this updated result include: (i) a color-flow variable proposed in Ref. [28] to estimate if the jets are from decays of color singlets, or initiated by gluons, (ii) the cosine of the angle between the direction of the Higgs candidate and the leading jet, boosted to the rest frame of the Higgs candidate; this distribution is expected to be flat for the decay of a scalar, in contrast to gluon splitting, and (iii) the p_T weighted distance in $(\eta - \phi)$ space between the second leading jet and other jets. The full list of input variables to the Decision Trees is given in Table I.

The MJ-DT output (multijet discriminant) is shown in Fig. 5 for $m_H = 115$ GeV. A value of the multijet discriminant in excess of 0 is required (multijet veto), which removes over 95% of the multijet background and 65% of the non-MJ SM backgrounds, while retaining 70% of the signal. The number of expected signal and background events, as well as the number of observed events, are given in Table II, after imposing the multijet veto. Distributions in the analysis sample after the multijet veto are shown in Fig. 6, before b-tag requirement and in Fig. 7 for double tagged events.

To discriminate signal from SM backgrounds, additional "SM DT" (SM-rejection DT) are trained, using the same kinematic variables as in the MJ DT. The output of the SM DT after the multijet veto (final discriminant) is shown in Fig. 8(a) for $m_H = 115$ GeV.

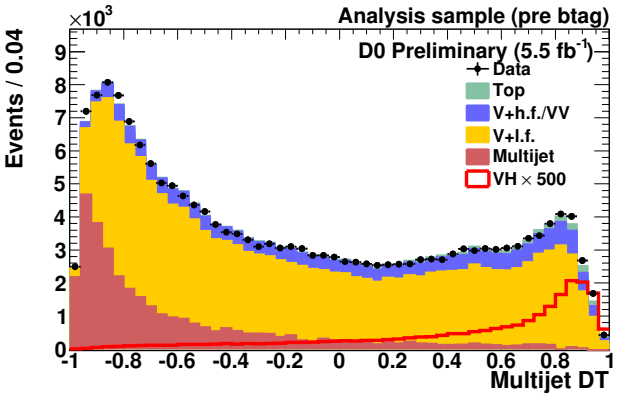


FIG. 5: MJ-Decision tree output for $m_H = 115$ GeV. The distribution for signal (VH) is multiplied by a factor of 500 and includes ZH and WH production.

TABLE I: Variables used as input to the Decision Trees. See the auxiliary material for the variable definitions.

Number of taggable jets
leading jet p_T
second jet p_T
$\Delta\eta(\text{jet}_1, \text{jet}_2)$
$\Delta\phi(\text{jet}_1, \text{jet}_2)$
$\Delta R(\text{jet}_1, \text{jet}_2)$
p_T weighted $\Delta R(\text{jet}_2, \text{jet}_i)$
$\not\!\!E_T$
$\Delta\phi(\not\!\!E_T, \text{jet}_2)$
$\max \Delta\phi(\not\!\!E_T, \text{jet}_i) + \min \Delta\phi(\not\!\!E_T, \text{jet}_i)$
$\max \Delta\phi(\not\!\!E_T, \text{jet}_i) - \min \Delta\phi(\not\!\!E_T, \text{jet}_i)$
H_T(scalar sum of jet p_T)
$\not\!\!H_T$(vectorial sum of jet p_T)
$\not\!\!H_T$/ H_T
dijet p_T
dijet mass
dijet transverse mass
Cos(Higgs decay angle)
color flow leading jet
color flow second jet

V. LIMIT SETTING PROCEDURE

Once systematic uncertainties are taken into account, agreement is found between data and expectation from SM and MJ backgrounds, both in the number of selected events (Table II) and in the distribution of final discriminant (Fig. 8). A modified frequentist approach [29] is used to set limits on the cross section for SM Higgs-boson production, where the test statistic is a joint log-likelihood ratio (LLR) for the background-only and signal+background hypotheses. The

TABLE II: The number of expected signal and background events, and the number observed after the multijet veto, prior to b tagging and for double tags. The signal corresponds to $m_H = 115$ GeV, "Top" includes pair and single top quark production, and VV is the sum of all diboson processes. The uncertainties quoted arise from the statistics of the simulation and from the sources of systematic uncertainties mentioned in the text.

Sample	ZH	WH	W+jets	Z+jets	Top	VV	Multijet	Total background	Observed
Pretag	18.56 ± 0.07	18.26 ± 0.11	48 468	14931	2075	1876	3325	70 678 ± 246	70 815
Double tag	5.30 ± 0.03	4.91 ± 0.05	286	148	336	32	47 ± 17	849 ± 21	881

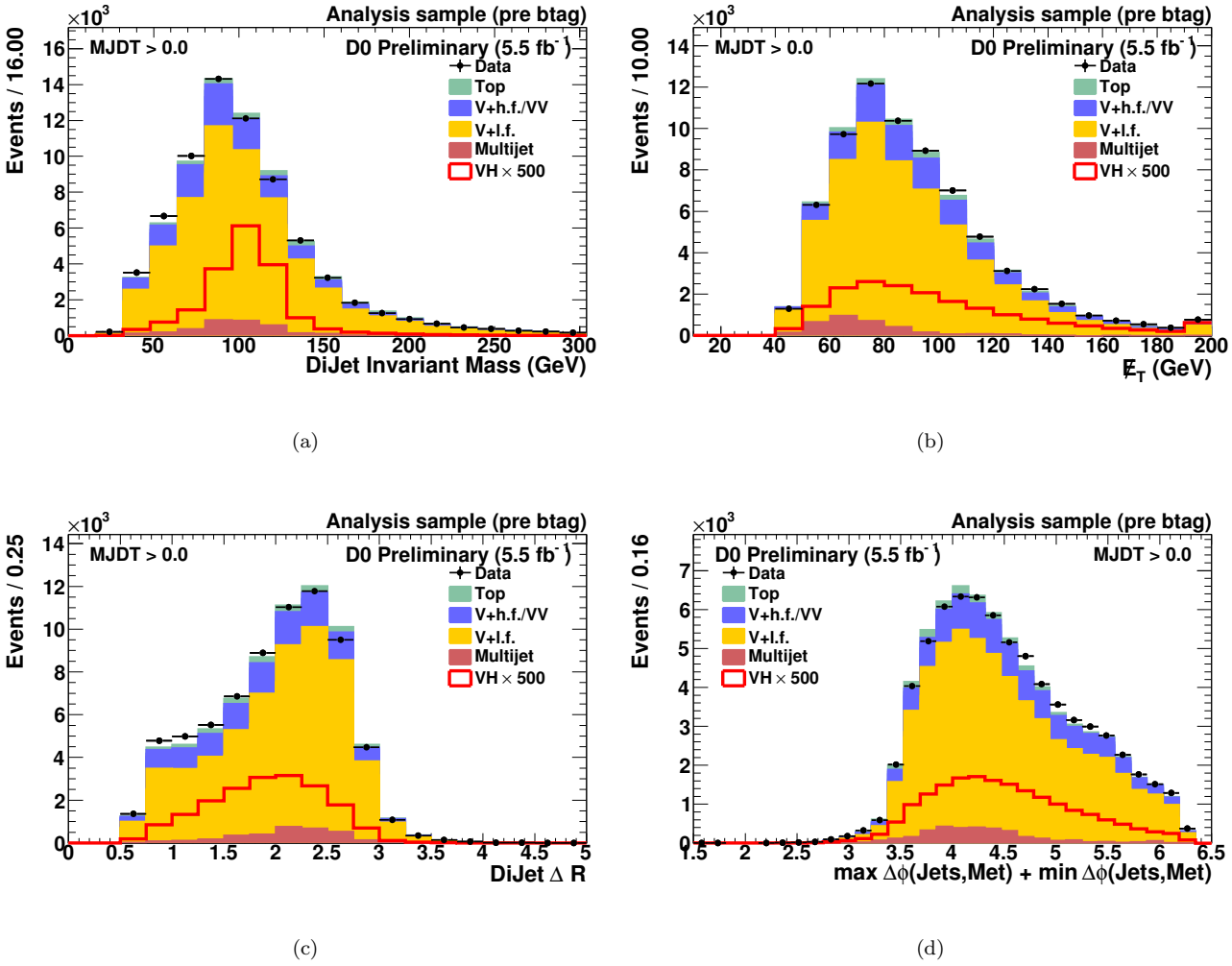


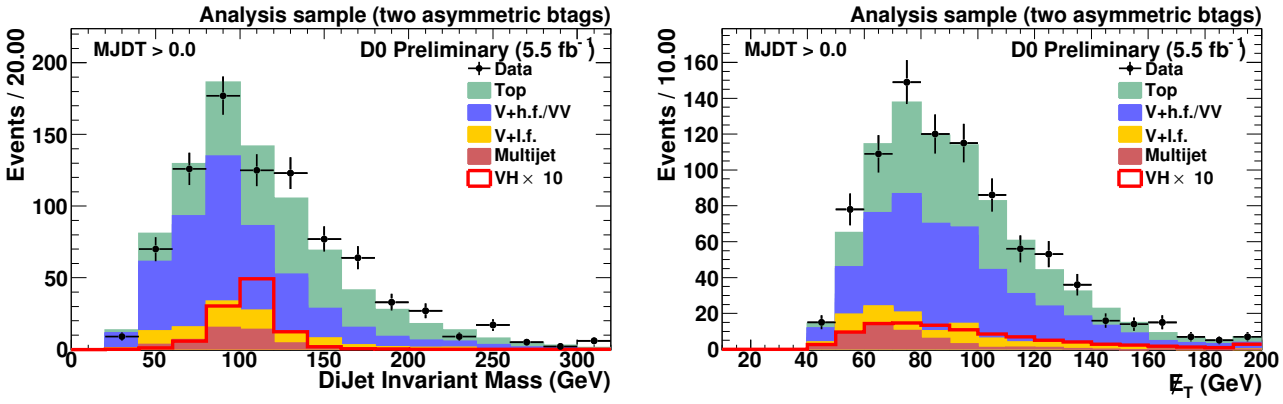
FIG. 6: Representative variable distributions in the analysis sample after the multijet veto: (a) Dijet invariant mass, (b) Missing E_T, (c) Dijet ΔR, (d) $\max \Delta\phi(\not\!\!E_T, \text{jet}_i) + \min \Delta\phi(\not\!\!E_T, \text{jet}_i)$. The distributions for signal (VH) are multiplied by a factor of 500 and include ZH and WH production for $m_H = 115$ GeV.

result is obtained by summing LLR values over the bins in the final discriminant shown in Fig. 8(a). The impact of systematic uncertainties on the sensitivity of the analysis is reduced by maximizing a "profile" likelihood function [30] in which these uncertainties are given Gaussian constraints associated with their priors and the final discriminant after the maximization is shown in Fig. 8(b). This is further illustrated in Fig. 9 where a comparison of the DT distribution after profiling between background-subtracted data and the expected signal scaled by the observed limit in the Run IIb double tag analysis for the $m_H = 115$ GeV hypothesis is shown. In this plot the background prediction and its uncertainties have been determined from the fit to data under the background-only hypothesis.

VI. SYSTEMATIC UNCERTAINTIES

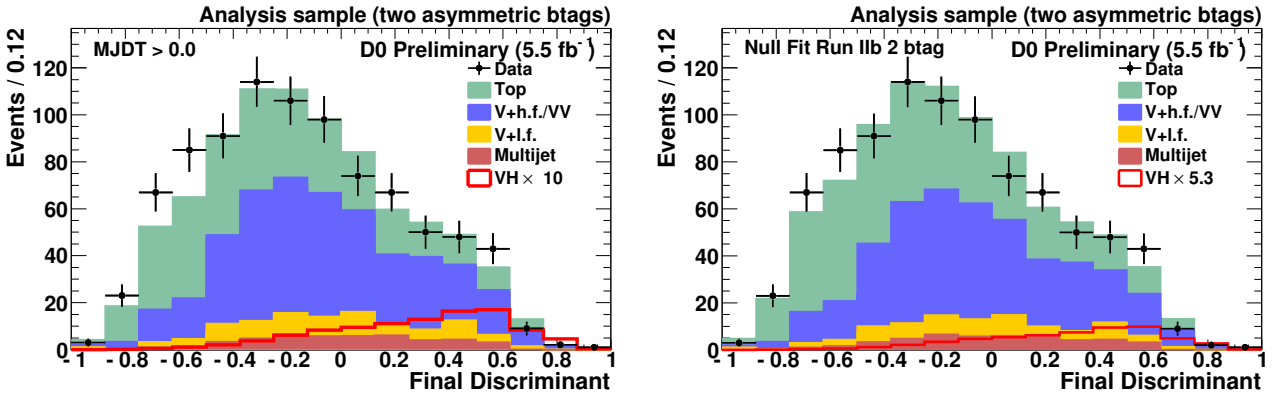
Experimental uncertainties arise from trigger simulation (3%), jet energy calibration and resolution (3% for signal and $4\% - 5\%$ for background), jet reconstruction and taggability ($2\% - 3\%$), lepton identification ($1\% - 2\%$), and b tagging (from 2% for signal in the single-tag sample to 8% for background in the double-tag sample). Their impact is assessed on overall normalizations and shapes of distributions in final discriminants. Correlations among systematic uncertainties in signal and background are taken into account in extracting the final results, including a 6.1% uncertainty on the integrated luminosity.

Theoretical uncertainties on cross sections for SM processes are estimated as follows. For (W/Z)+jets production,



(a) (b)

FIG. 7: (a) Dijet invariant mass with double tag in the analysis sample after the multijet veto, (b) Missing $E_T$with double tag in the analysis sample after the multijet veto. The distributions for signal (VH) are multiplied by a factor of 10 and include ZH and WH production for $m_H = 115$ GeV.



(a) (b)

FIG. 8: (a) Decision tree output for $m_H = 115$ GeV for the SM DT following the multijet veto for double tag. The distribution for signal (VH) is multiplied by a factor of 10 and includes ZH and WH production for $m_H = 115$ GeV. (b) Decision tree output for $m_H = 115$ GeV for the SM DT following the multijet veto for double tag after profiling. The signal histogram (VH) is scaled by 5.3, the observed limit in the Run IIb double tag analysis.

an uncertainty of 10% is assigned to the total cross sections, and an uncertainty of 20% on the heavy-flavor fractions (estimated from MCFM). For other SM backgrounds, uncertainties are taken from [19] or from MCFM, and range from 6% to 10%. The uncertainties on cross sections for signal (6% for $m_H = 115$ GeV) are taken from [20]. Uncertainties on the shapes of the final discriminants arise from (i) the modeling of (W/Z)+jets, assessed by varying the renormalization-and-factorization scale and by comparing ALPGEN interfaced with HERWIG [31] to ALPGEN interfaced with PYTHIA, and (ii) the choice of PDFs, estimated using the prescription of [15].

VII. RESULTS

The results of the analysis are given in terms of LLR values in Fig. 10(a), and as limits in Table IV and Fig. 10(b), as a function of m_H. These limits incorporate the updated double-tag analysis presented in this note, combined with the Run IIa and single-tag Run IIb result from [2]. The observed LLRs are within 1 standard deviation of expectation

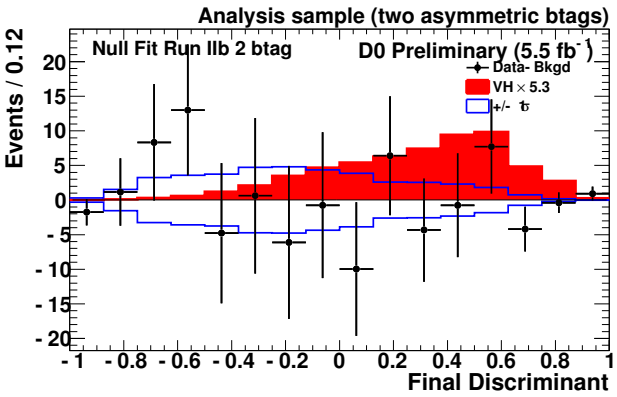


FIG. 9: Data after subtracting the fitted background (points) and scaled SM signal expectation (red histogram) as a function of the DT output for $m_H = 115$ GeV. The histogram for signal is scaled by the observed limit in the Run IIb double tag analysis. Also shown is the ± 1 standard deviation band on the total background after fitting.

TABLE III: Systematic uncertainties in percent of the overall signal and background yields. "Jet EC" and "Jet ER" stand for jet energy calibration and resolution, respectively. "Jet R&T" stands for jet reconstruction and taggability. "Signal" includes ZH and WH production for $m_H = 115$ GeV.

Systematic Uncertainty	Signal	Background
double tag		
Jet EC – Jet ER	1.1	3.3
Jet R&T	4.0	6.3
b tagging	4.7	8.7
Trigger	1.7	3.3
Lepton identification	0.6	1.3
Heavy-flavor fractions	–	8.4
Cross sections	6.0	7.8
Luminosity	6.1	6.1
Multijet normalization	–	1.4
Total	12.4	16.5

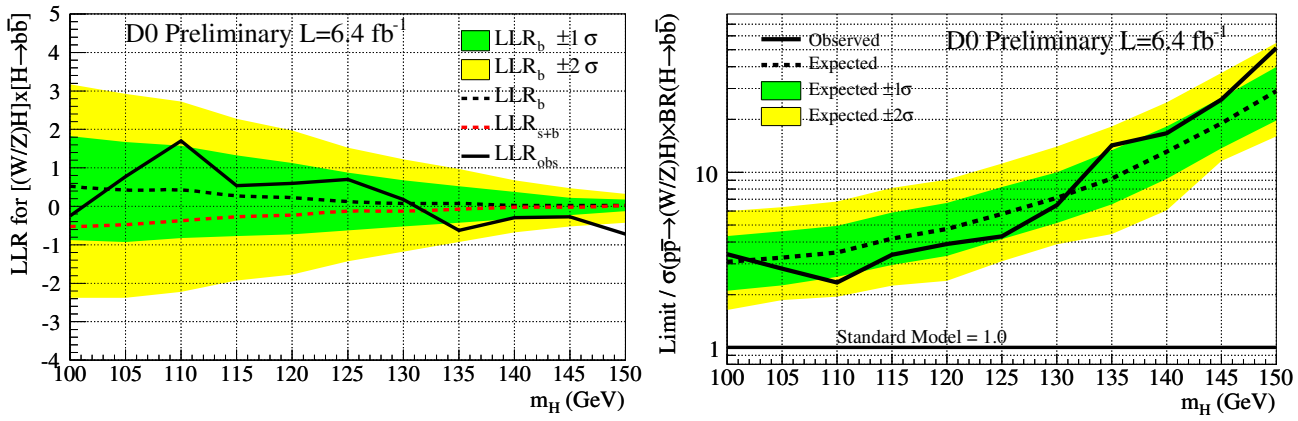
(the median of the LLR for the background-only hypothesis). For $m_H = 115$ GeV, the observed and expected limits on the combined cross section of ZH and WH production, multiplied by the branching fraction for $H \to b\bar{b}$, are factors of 3.4 and 4.2 larger than the SM value, respectively.

Acknowledgments

We thank the staffs at Fermilab and collaborating institutions, and acknowledge support from the DOE and NSF (USA); CEA and CNRS/IN2P3 (France); FASI, Rosatom and RFBR (Russia); CNPq, FAPERJ, FAPESP and FUNDUNESP (Brazil); DAE and DST (India); Colciencias (Colombia); CONACyT (Mexico); KRF and KOSEF (Korea); CONICET and UBACyT (Argentina); FOM (The Netherlands); STFC and the Royal Society (United Kingdom); MSMT and GACR (Czech Republic); CRC Program, CFI, NSERC and WestGrid Project (Canada);

TABLE IV: As a function of m_H, observed and expected upper limits on the $(W/Z)H$ production cross section multiplied by branching fraction for $H \to b\bar{b}$, relative to the SM expectation.

m_H	100	105	110	115	120	125	130	135	140	145	150
Expected	3.08	3.26	3.47	4.18	4.74	5.77	7.13	9.19	13.11	19.03	29.06
Observed	3.40	2.82	2.35	3.38	3.90	4.30	6.47	14.23	16.67	26.00	51.25



(a) (b)

FIG. 10: (a) The observed LLR is shown as a solid black line, the expected LLRs for the background-only and signal+background hypotheses are shown as black dots and red dashes, respectively, and the heavy green and light yellow shaded areas correspond to 1 and 2 standard deviations (s.d.) around the expected LLR for the background-only hypothesis. (b) Ratio of the observed (solid black) and expected (dotted red) exclusion limits to the SM production cross section multiplied by branching fraction for $H \to b\bar{b}$, as a function of m_H. These limits incorporate the updated double tag analysis presented in this note combined with the Run IIa and single b tag Run IIb result from [2].

BMBF and DFG (Germany); SFI (Ireland); The Swedish Research Council (Sweden); and CAS and CNSF (China).

[1] M. Carena *et al.*, `arXiv:hep-ph/0010338`; CDF and D0 Collaborations, Report No. FERMILAB-PUB-03/320-E, 2003.
[2] V.M. Abazov *et al.* (D0 Collaboration), Phys. Rev. Lett. **104**, 071801 (2010).
[3] R. Barate *et al.* (LEP Working Group for Higgs boson searches), Phys. Lett. B **565**, 61 (2003).
[4] LEP, Tevatron and SLD Electroweak Working Groups, `arXiv:0911.2604`.
[5] V.M. Abazov *et al.* (D0 Collaboration), Nucl. Instrum. Methods Phys. Res., Sect. A **565**, 463 (2006); M. Abolins *et al.*, Nucl. Instrum. Methods Phys. Res., Sect. A **584**, 75 (2008); R. Angstadt *et al.*, `arXiv:0911.2522`.
[6] C. Ochando, Report No. FERMILAB-THESIS-2008-78.
[7] T. Andeen *et al.*, Report No. FERMILAB-TM-2365, 2007.
[8] G.C. Blazey *et al.*, `arXiv:hep-ex/0005012`.
[9] M. Voutilainen, Report No. FERMILAB-THESIS-2008-19.
[10] M.L. Mangano *et al.*, J. High Energy Phys. 07 (2003) 001; version 2.11 was used.
[11] T. Sjöstrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026; version 6.409 was used.
[12] V.M. Abazov *et al.*, (D0 Collaboration), Phys. Rev. Lett. **100**, 102002 (2008).
[13] K. Melnikov and F. Petriello, Phys. Rev. D **74**, 114017 (2006).
[14] E. Boos *et al.* (CompHEP Collaboration), Nucl. Instrum. Methods Phys. Res., Sect. A **534**, 250 (2004).
[15] J. Pumplin *et al.*, J. High Energy Phys. 07 (2002) 012; D. Stump *et al.*, J. High Energy Phys. 10 (2003) 046.
[16] R. Hamberg, W.L. van Neerven, and W.B. Kilgore, Nucl. Phys. **B359**, 343 (1991); **B644**, 403 (2002).
[17] A.D. Martin, R.G. Roberts, W.J. Stirling, and R.S. Thorne, Phys. Lett. B **604**, 61 (2004).
[18] J.M. Campbell and R.K. Ellis, Phys. Rev. D **60**, 113006 (1999).
[19] N. Kidonakis and R. Vogt, Phys. Rev. D **78**, 074005 (2008); N. Kidonakis, Phys. Rev. D **74**, 114012 (2006).
[20] T. Hahn *et al.*, `arXiv:hep-ph/0607308`.
[21] R. Brun and F. Carminati, CERN Program Library Long Writeup W5013, 1993 (unpublished).
[22] The pseudorapidity is defined as $\eta = -\ln[\tan(\theta/2)]$, where θ is the polar angle with respect to the proton beam direction.
[23] A. Schwartzman, Report No. FERMILAB-THESIS-2004-21.
[24] V.M. Abazov *et al.* (D0 Collaboration), Phys. Rev. D **76**, 092007 (2007).
[25] T. Gleisberg *et al.*, J. High Energy Phys. 02 (2004) 056; J. Alwall *et al.*, Eur. Phys. J. C **53**, 473 (2008).
[26] V. M. Abazov *et al.* [The D0 Collaboration], Nucl. Instrum. Methods in Phys. Res. Sect. **A 620**, 400 (2010).
[27] L. Breiman *et al.*, *Classification and Regression Trees* (Wadsworth, Belmont, CA, 1984); Y. Freund and R.E. Schapire, in *Machine Learning: Proceedings of the Thirteenth International Conference, Bari, Italy, 1996,* edited by L. Saitta (Morgan Kaufmann, San Francisco, 1996), p148.

[28] J. Gallicchio and M. D. Schwartz, Phys. Rev. Lett. **105**, 022001 (2010).
[29] T. Junk, Nucl. Instrum. Methods Phys. Res., Sect. A **434**, 435 (1999); A. Read, J. Phys. G **28**, 2693 (2002).
[30] W. Fisher, Report No. FERMILAB-TM-2386-E, 2006.
[31] G. Corcella *et al.*, J. High Energy Phys. 01 (2001) 010.
[32] V. M. Abazov *et al.*, *"Observation of ZZ production in $p\bar{p}$ collisions at $\sqrt{s}$ = 1.96 TeV"*, Phys. Rev. Lett. **101**, 171803 (2008)

Auxiliary material

In the following, the decision tree input variables are displayed in the analysis sample without any b tag requirements before the multijet veto (Figs. 11-12), and after (Figs. 13-14), and in the double b tag sample after the multijet veto (Figs. 15-16). For each sample the variables are shown in the following order:

1. Missing transverse energy ($\not\!\!E_T$)
2. Dijet invariant mass constructed from the leading two jets
3. The difference in ϕ of the leading two jets ($\Delta\phi(\mathrm{jet}_1, \mathrm{jet}_2)$)
4. The number of candidate jets
5. The transverse momentum of the leading jet
6. The transverse momentum of the next-to-leading jet
7. Scalar sum of all jets (H_T)
8. Vector sum of all jest ($\not\!\!H_T$)
9. Ratio of vector sum and scalar sum of all jets ($\not\!\!H_T$/ H_T)
10. Difference in ϕ between the $\not\!\!E_T$ and the next-to-leading jet ($\Delta\phi(\not\!\!E_T, \mathrm{jet}_2)$)
11. Mismeasurements of jet energies, which plague the multijet background, can cause di-jet events to have a jet alligned with the $\not\!\!E_T$ while jets back-to-back with the $\not\!\!E_T$ is expected from Higgs decays. To optimize this topological difference the sum of the maximum and minimum difference in ϕ of the $\not\!\!E_T$ with any jet ($\max\Delta\phi(\not\!\!E_T, \mathrm{jet_i}) + \min\Delta\phi(\not\!\!E_T, \mathrm{jet_i})$).
12. With similar motivation to the previous variable we use the difference of the maximum and minimum difference in ϕ of the $\not\!\!E_T$ with any jet ($\max\Delta\phi(\not\!\!E_T, \mathrm{jet_i}) - \min\Delta\phi(\not\!\!E_T, \mathrm{jet_i})$)
13. The color-flow variable for the leading jet is used to help discriminate a pair of b jets originating from a color singlet, the Higgs, and a color octet, a gluon. The angle in ϕ-η space between the jet pull, $\vec{\rho} = \Sigma_i \frac{E_T^i |r_i|}{E_T^{\mathrm{jet}_1}} \vec{r}_i$ where the summation is over the cells that make up the jet and $\vec{r}_i$ is in the direction of the jet center to the cell, and the direction from the center of jet_1 to jet_2 is defined as the leading jet color-flow.
14. The color-flow variable for the next-to-leading jet where the jet pull is defined as previous variable with jet_1 and jet_2 interchanged.
15. The cosine of the angle between the direction of the Higgs candidate and the leading jet, in the rest frame of the Higgs candidate. This variable attempts to capitalize on the angular difference between gluon splitting and a scalar particle decay.
16. The modified "recoil-subtracted" dijet p_T component along the unit vector $\hat{a}_T$ perpendicular to the dijet thrust axis. The modification seeks to find the minimum sensible dijet p_T by subtracting twice the larger of the $\not\!\!E_T$ or $-\Sigma_k \vec{p}_T \dot{} \hat{a}_T$ where the sum is over the non-leading jets. [recoil subtracted dijet$p_\perp = p_{T,t}^{jj} + 2 \times min(-\not\!\!E_{T\,uncorr} \cdot \hat{a}_t, \sum_{jets} p_T^{jet} \cdot \hat{a}_t, 0)$]. This variable is less sensitive to detector effects so one expects lower valuesfor instrumental backgrounds and is closely related to that defined in Ref [32].
17. Distance in η-ϕ space between the two leading jets $\Delta R(\mathrm{jet}_1, \mathrm{jet}_2)$
18. Difference in η of the two leading jets $\Delta\eta(\mathrm{jet}_1, \mathrm{jet}_2)$
19. The p_T weighted distance in $(\eta - \phi)$ space between the second leading jet and other jets ($\Sigma(\mathrm{jet}_2) = \frac{\Sigma_i \Delta R(\mathrm{jet}_2, jet_i) \times p_T^{jet_i}}{\Sigma_{jets} p_T^{jet}}$)
20. Transverse mass of the two leading jets

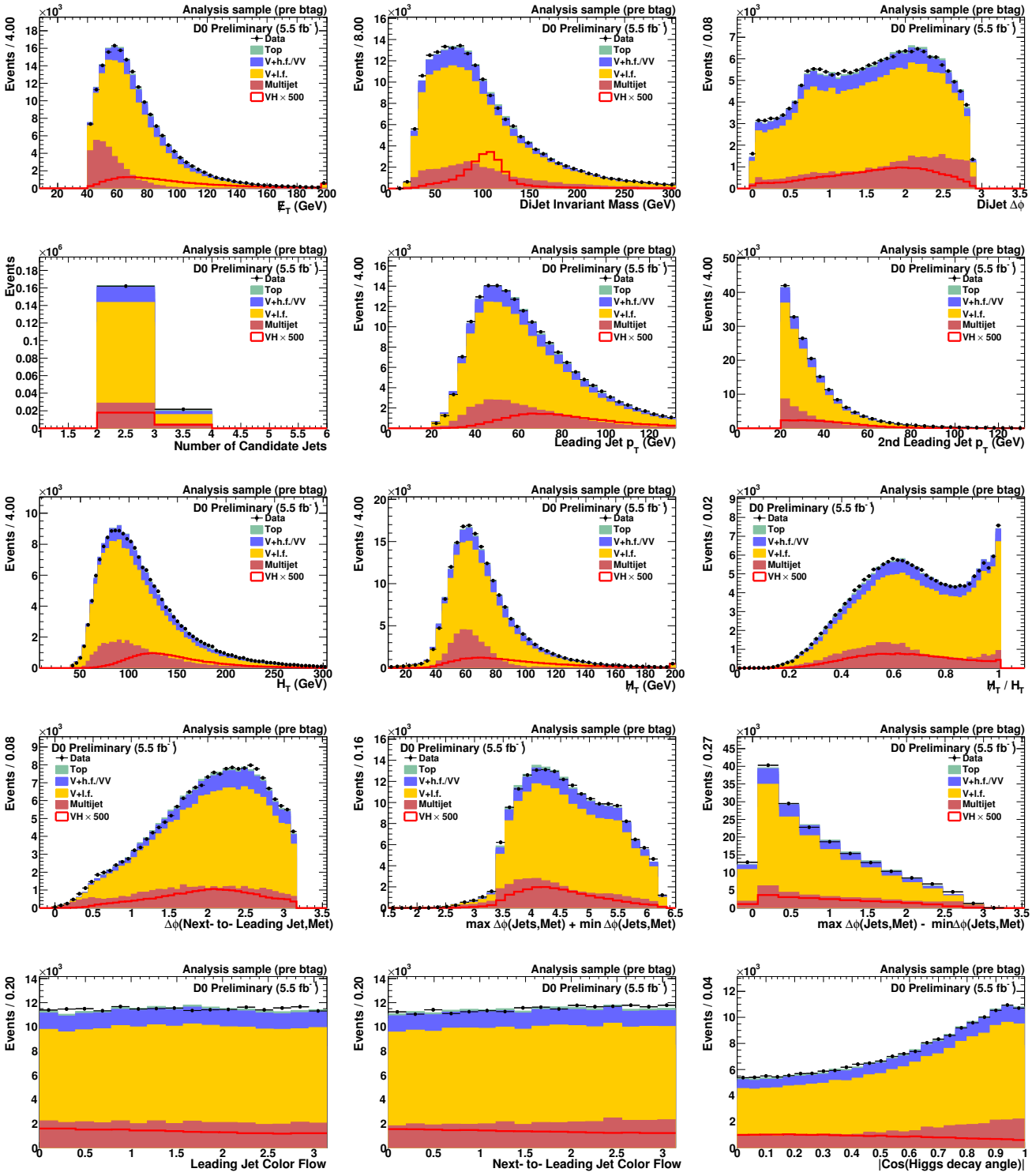


FIG. 11: Decision Tree input variables at the pretag level.

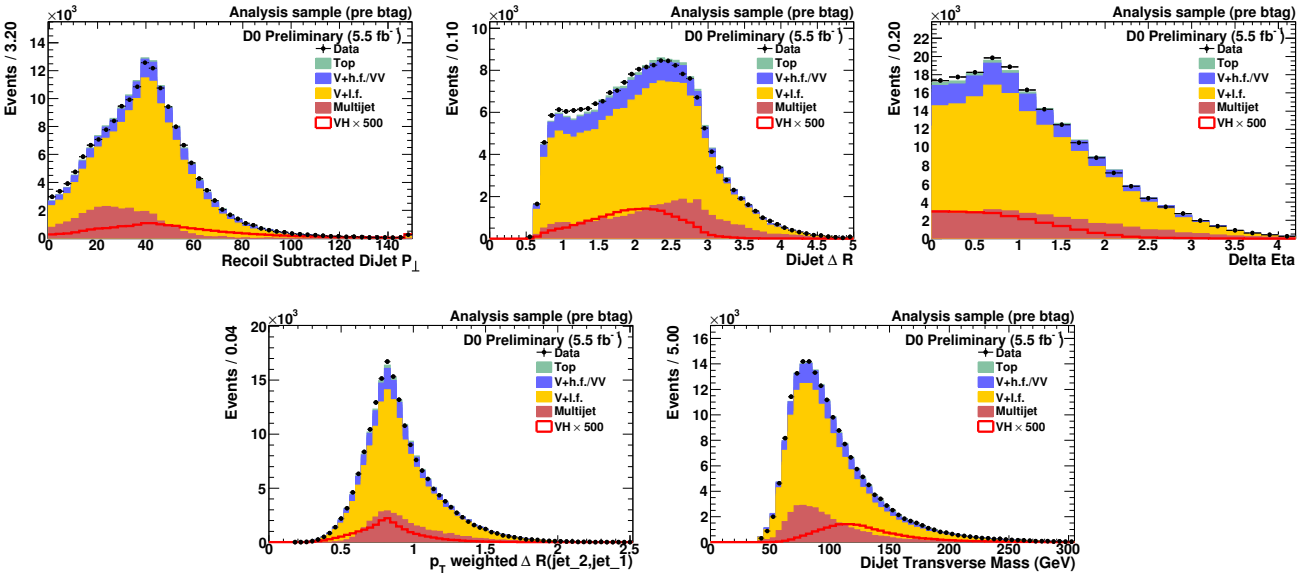


FIG. 12: Decision Tree input variables at the pretag level.

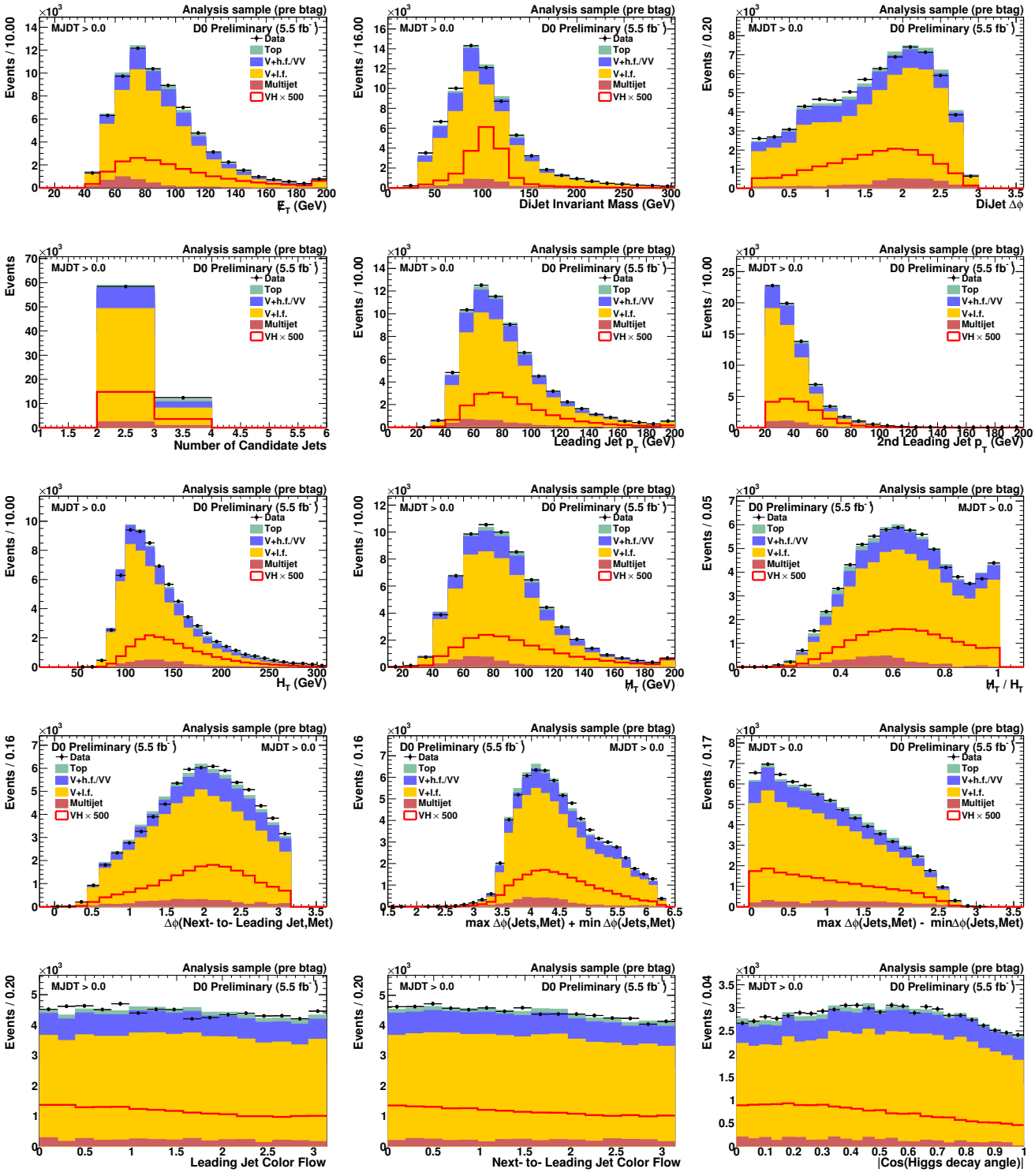


FIG. 13: Decision Tree input variables at the pretag level after the multijet veto.

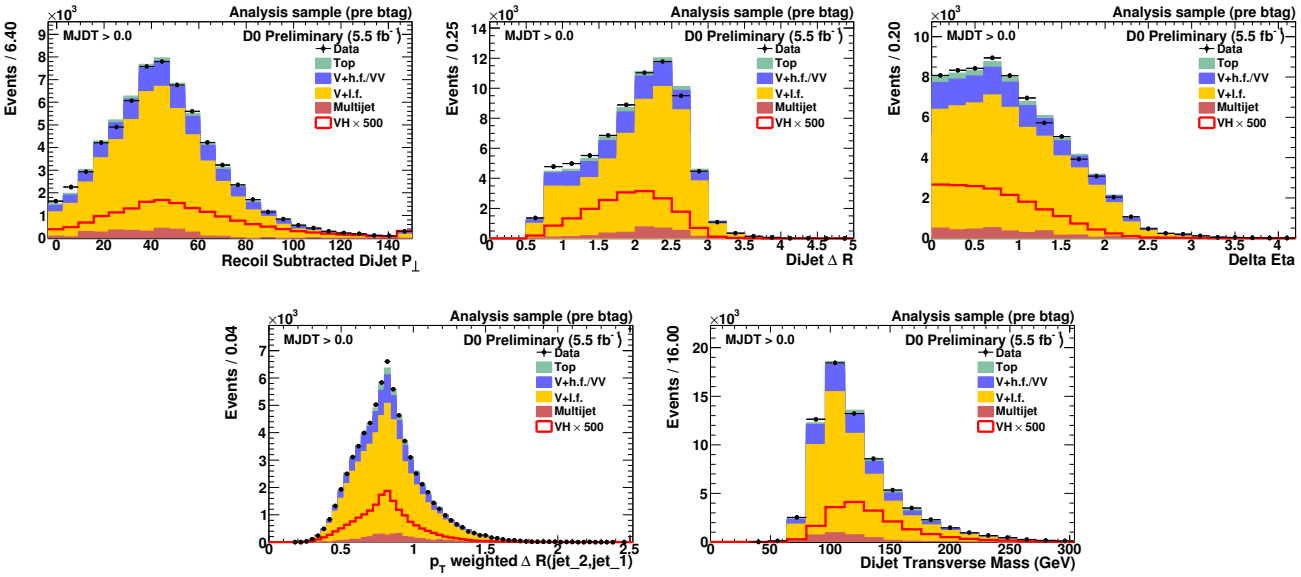


FIG. 14: Decision Tree input variables at the pretag level after the multijet veto.

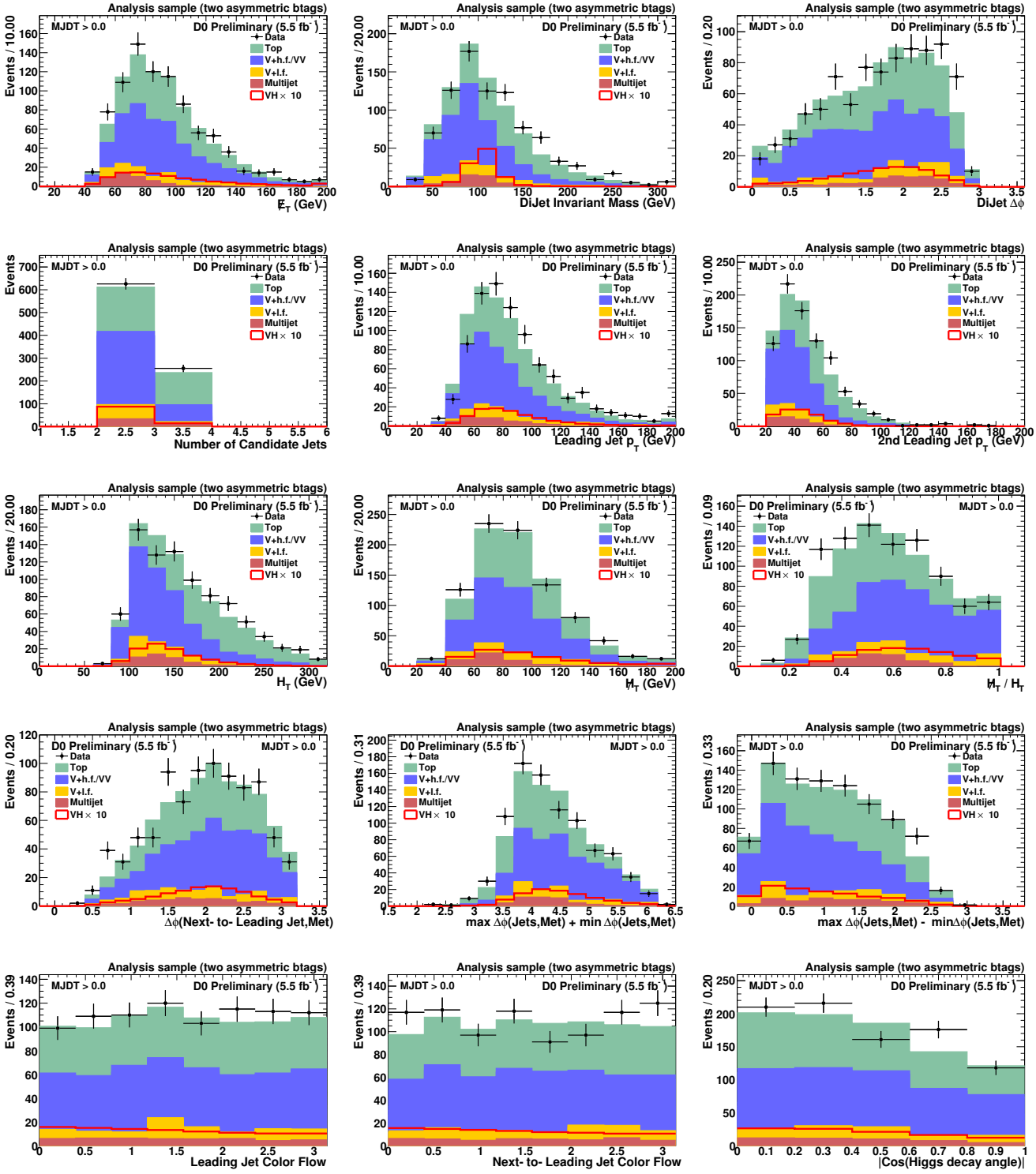


FIG. 15: Decision Tree input variables after the multijet veto in the double b tag sample.

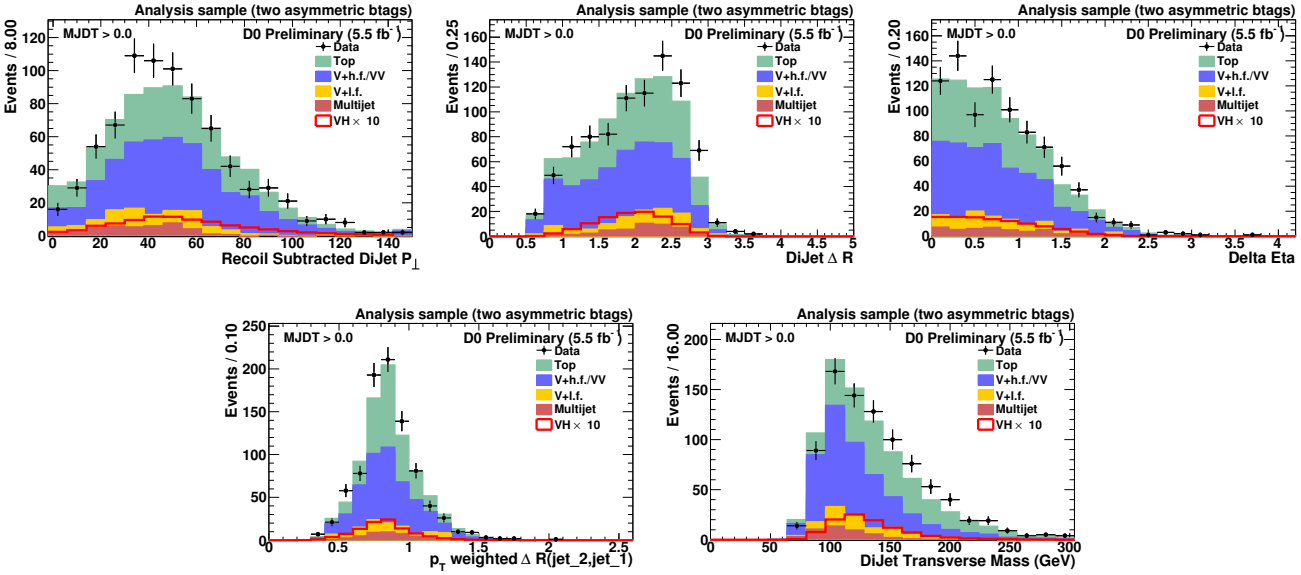


FIG. 16: Decision Tree input variables after the multijet veto in the double b tag sample.