

A Tale of two Higgs: Search for pair  
production of Higgs bosons in the  $b\bar{b}b\bar{b}$   
final state using proton–proton collisions at  
 $\sqrt{s} = 13$  TeV with the ATLAS detector

A DISSERTATION PRESENTED

BY

BAOJIA TONG

TO

THE DEPARTMENT OF PHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

PHYSICS

HARVARD UNIVERSITY  
CAMBRIDGE, MASSACHUSETTS

MAY 2018

©2017-2018 – BAOJIA TONG  
ALL RIGHTS RESERVED.

# A Tale of two Higgs: Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector

## ABSTRACT

This thesis presents a search for Higgs boson pair production, with the  $b\bar{b}b\bar{b}$  final state. This search uses the full 2015 and 2016 data collected by the ATLAS Collaboration at  $\sqrt{s} = 13$  TeV, corresponding to  $3.2 \pm 0.2 \text{ fb}^{-1}$  of 2015 and  $32.9 \pm 1.1 \text{ fb}^{-1}$  of 2016  $pp$  collision data. Improvements with respect to the previous analysis come from the increased dataset, detailed background estimation and additional signal regions. Search sensitivity is specially enhanced for the resonance signals between 2500 GeV and 3000 GeV. The data is found to be compatible with the Standard Model predictions, and no signs of new physics have been observed. The results are interpreted in the context of the bulk Randall-Sundrum warped extra dimension model with a Kaluza-Klein graviton  $k/\bar{M}_{\text{Pl}} = 1.0$  or  $2.0$  decaying to  $hh$ , and the Type 2 two-Higgs doublet model (2HDM) where the neutral heavy CP-even  $H$  scalar decays to  $hh$ .

# Contents

o INTRODUCTION	I
1 THEORY AND MOTIVATION	5
1.1 The Standard Model and the Higgs Boson . . . . .	5
1.2 Standard Model di-Higgs production . . . . .	7
1.3 Beyond the Standard Model physics di-Higgs production . . . . .	9
1.4 Di-Higgs decay and LHC previous search results . . . . .	13
2 LHC AND ATLAS	16
2.1 The Large Hadron Collider . . . . .	17
2.2 A Toroidal LHC ApparatuS . . . . .	19
3 RECONSTRUCTION AND OBJECTS	25
3.1 ID Tracks and Vertices . . . . .	26
3.2 Jets . . . . .	27
3.3 Flavor Tagging . . . . .	30
3.4 Leptons . . . . .	33
3.5 Resolved and Boosted . . . . .	35
4 DATA AND SIMULATION	37
4.1 Data . . . . .	37
4.2 MC . . . . .	39
4.3 Signal . . . . .	40
5 EVENT SELECTION	42
5.1 Data Cleaning . . . . .	42
5.2 Trigger . . . . .	43
5.3 Object Selection . . . . .	44
5.4 Resolved Veto . . . . .	49
5.5 2D Higgs Mass Cut . . . . .	51

5.6	Number of $b$ -tagging requirement . . . . .	53
5.7	Signal efficiency and cutflow . . . . .	54
<b>6</b>	<b>BACKGROUND ESTIMATION</b>	<b>58</b>
6.1	Background Estimation . . . . .	58
<b>7</b>	<b>BACKGROUND ESTIMATION</b>	<b>144</b>
7.1	Background Estimation . . . . .	144
<b>8</b>	<b>SYSTEMATICS</b>	<b>227</b>
<b>9</b>	<b>RESULT</b>	<b>257</b>
<b>10</b>	<b>INTERPRETATION</b>	<b>284</b>
<b>II</b>	<b>CONCLUSION</b>	<b>286</b>
	<b>REFERENCES</b>	<b>288</b>

# Listing of figures

<p>1.1 Fermions and bosons of the Standard Model and their properties<sup>1</sup>, where all the values are measured experimentally. . . . .</p> <p>1.2 Leading order Feynman diagrams contributing to di-Higgs production via gluon-gluon fusion, through the Higgs-fermion Yukawa interactions 1.2a and the Higgs boson self-coupling 1.2b. Only Figure 1.2b probes <math>\lambda_{hhh}</math>. . . . .</p> <p>1.3 Total cross sections (y-axis) at the NLO in QCD for the six largest di-Higgs production channels at p-p colliders at different energy (x-axis). The thickness of the lines corresponds to the scale and PDF uncertainties added linearly. <math>H</math> refers to the SM Higgs. . . . .</p> <p>1.4 BSM Higgs boson pair production: non-resonant production proceeds through changes in the SM Higgs couplings in 1.4a and 1.4b, resonant production proceeds through 1.4c an intermediate resonance, <math>X</math>. <math>H</math> and <math>b</math> both refers to the SM Higgs. . . . .</p> <p>1.5 Total cross sections (y-axis) at the LO and NLO in QCD for di-Higgs production channels, at the <math>\sqrt{s} = 14</math> TeV LHC as a function of the self-interaction coupling <math>\lambda</math> (x-axis). The dashed (solid) lines and light- (dark-) color bands correspond to the LO (NLO) results and to the scale and PDF uncertainties added linearly. The SM values of the cross sections are obtained at <math>\frac{\lambda}{\lambda_{SM}} = 1</math>. <math>H</math> refers to the SM Higgs. . . . .</p> <p>1.6 Parton luminosity ratios as a function of resonance mass <math>M_X</math> for <math>13/8</math> TeV<sup>2</sup>. For a <math>2</math> TeV <math>X</math>, the luminosity ratio is almost 10. . . . .</p> <p>1.7 Summary of di-Higgs final states and their ratios. Top left, <math>b\bar{b}b\bar{b}</math>, has the largest branching ratio. . . . .</p> <p>1.8 The observed and expected 95% CL upper limits of <math>\sigma(gg \rightarrow H) \times BR(H \rightarrow bb)</math> at <math>\sqrt{s} = 8</math> TeV as functions of the heavy Higgs boson mass <math>m_H</math>, combining resonant searches in Higgs boson pair to <math>b\bar{b}\tau^+\tau^-</math>, <math>W^+W^-\gamma\gamma</math>, <math>b\bar{b}\gamma\gamma</math>, and <math>b\bar{b}b\bar{b}</math> final states. The expected limits from individual searches are also shown. The green and yellow bands represent <math>\pm 1\sigma</math> and <math>\pm 2\sigma</math> uncertainty ranges of the expected combined limits. The improvement above <math>m_H = 500</math> GeV reflects the sensitivity of the <math>b\bar{b}b\bar{b}</math> analysis. The results beyond 1 TeV are only from the <math>b\bar{b}b\bar{b}</math> final state alone. . . . .</p>	<p>6</p> <p>7</p> <p>8</p> <p>9</p> <p>10</p> <p>12</p> <p>13</p> <p>14</p>
---	---

2.1	A schematic view of the LHC ring <sup>3</sup> . LINAC <sub>2</sub> , Booster, PS, SPS, and LHC accelerate the protons in order. Four main experiments are located at interaction points along the ring. ATLAS and CMS are general purpose experiments, while ALICE focuses on heavy ion collisions and LHC <sub>b</sub> is dedicated to $B$ physics. . . . .	17
2.2	The luminosity-weighted distribution of the mean number of interactions per crossing for the 2015 and 2016 $p\bar{p}$ collision data at 13 TeV centre-of-mass energy. <sup>4</sup> . . . . .	19
2.3	A detailed computer-generated image of the ATLAS detector and its systems. . . . .	20
2.4	Gemoetry of IBL, PIXEL and SCT detectors in Run2. . . . .	22
2.5	The overall layout of the ATLAS MuonSpectrometer. . . . .	23
2.6	Cumulative luminosity vs. time delivered to (green) and recorded by ATLAS (yellow) during stable beams for $p\bar{p}$ collisions at 13 TeV centre-of-mass energy. . . . .	24
3.1	Illustration of particle interactions in ATLAS. . . . .	26
3.2	Uncalibrated (dashed line) and calibrated (solid line) reconstructed jet mass distribution 3.2a, and the jet mass resolution vs jet $p_T$ 3.2b for calorimeter-based jet mass, $m_{calo}$ (red), track-assisted jet mass $m_{TA}$ (black) and the invariant mass of four-vector sum of tracks associated to the large-radius calorimeter jet $m_{track}$ (blue) for W/Z-jets <sup>5</sup> . . . . .	29
3.3	Secondary vertex reconstruction rate and $MV_{2c10}$ output for b-jets (solid blue), c-jets (dashed green) and light-jets (dotted red) evaluated with simulated $t\bar{t}$ events. . . . .	31
3.4	$b$ -jet efficiency for the fixed cut working point with a $b$ -jet efficiency of 77% as a function of the jet $p_T$ for the comparison between the $MV_{2c10}$ $b$ -tagging algorithm employed for the 2016 analyses (2016 config) and the previous version of the tagger, $MV_{2c20}$ (2015 config), which has 15% $c$ -fraction in the training . . . . .	33
3.5	Light-flavour jet and $c$ -jet rejection as a function of jet $p_T$ for the previous (2015 config) $MV_{2c20}$ and the current $MV_{2c10}$ configuration (2016 config). A fixed cut at 77% $b$ -jet efficiency operating point is used <sup>6</sup> . . . . .	34
3.6	Event display of the same event using 3.6a resolved and 3.6b boosted topologies. ID is in grey, ECAL is in green, HCAL is in red and MS is in blue. Jets are gray cones, and ID tracks are colored lines in the ID. The resolved reconstruction ends up with a $m_{4J}$ of 873 GeV, and the boosted reconstruction gives $m_{2J}$ of 852 GeV. . . . .	36
5.1	Different trigger efficiencies as a function of the signal resonance mass with respect to all events with no selection (left) and with respect to events passing the two large- $R$ jets $p_T > 400$ GeV and leading/subleading jet $p_T > 250$ GeV (right). For 1.4 TeV signal, the trigger efficiency is about 98%. . . . .	43

5.2	Large- $R$ jet trigger efficiencies, defined as the fraction of events fired trigger with a given highest large- $R$ jet $p_T$ , measured in 2015 Data ( <code>HLT_j360_a10_lcw</code> , left) and 2016 Data ( <code>HLT_j420_a10_lcw</code> , right) and MC. . . . .	44
5.3	Percentages of truth Higgs to large- $R$ jet $\Delta R < 1.0$ matching as a function of $G_{KK}^*$ mass. Both Higgs almost never match to the same large- $R$ jet. . . . .	45
5.4	Normalized $\Delta R$ between the truth Higgs (leading on left, subleading on right) and the truth children $b$ -quarks for $G_{KK}^*$ MCs. Lines are drawn at $\Delta R = 0.4$ ( $R$ of small- $R$ jets) and $\Delta R = 1.0$ ( $R$ of large- $R$ jets). . . . .	46
5.5	Percentage of $\Delta R < 0.2$ matching truth $b$ 's to track jets (leading Higgs on the left, subleading Higgs on the right) for different $G_{KK}^*$ mass. The cases listed in the legend are orthogonal to each other. The cases not listed on the legend (including when a truth $b$ is not contained in the large- $R$ jet) happen in total at most 1.6% of the time for a given $G_{KK}^*$ mass. . . . .	47
5.6	Kinematics of 1 TeV $G_{KK}^*$ before and after muon-in-jet corrections. The reconstructed Higgs masses (top row, left for leading large- $R$ jet, right for subleading large- $R$ jet) are closer to 125 GeV after the correction, which improves the signal efficiency for the signal region selection by $\sim 10\%$ (bottom row, left for $m_{JJ}$ , right for event distribution differences on the leading-subleading large- $R$ jet mass plane.). . . . .	48
5.7	After large- $R$ jet requirements, normalized $\Delta\eta_{JJ}$ distribution in 1.5 TeV $G_{KK}^*$ and data, where the data consists of mostly multijet events ( $> 90\%$ ). The background multijet event is flatter in $\Delta\eta_{JJ}$ distribution. . . . .	49
5.8	For RSG $c = 1.0$ samples, number of events as a function of leading Higgs candidate mass and subleading Higgs candidate mass, for 1.2 TeV (left) signal and 2 TeV (right) signal samples. The red dotted line in the center correspond to the signal region, passing $X_{hh} < 1.6$ . . . . .	52
5.9	Signal fraction in different $b$ -tag categories (left) and detailed fraction in different number of track jet and $b$ -tag categories (right) as a function of signal resonance mass hypothesis for selection cuts. The efficiencies are relative to the total number of events passing the 2D mass cut. . . . .	54

5.10	Absolute (left) and relative (right) signal efficiency as a function of RSG $c=1.0$ signal resonance mass hypothesis for selection cuts. The relative efficiency is defined from the previous cut, where the order of cuts is given by the legend. PassTrig means the event passes the trigger selection; PassDijetPt means the event passes the leading and sub-leading jet $p_T$ cuts; PassDijetEta means the event passes the leading and sub-leading jet $\eta$ cuts; PassDeltaH means the events passes the $ \Delta\eta  < 1.7$ cut; PassBJetSkim means the event contains at least two $b$ -tagged track jets, inclusive of $2b$ , $2bs$ , $3b$ and $4b$ configurations; PassSignal means the event passes the signal region cut $X_{hh} < 1.6$ . . . . .	55
6.1	Values of the $X_{hh}$ and $R_{hh}$ variables, which are shapes in the two-dimensional plane of the large- $R$ jet masses used to defined signal, control, and sideband regions. For both variables, a smaller value indicates the jets are closer to the Higgs mass. . . . .	59
6.2	Detailed signal efficiency in different signal/control/sideband regions as in $2bs$ (top left, $3b$ (top right), $4b$ (bottom left) and inclusive $b$ -tagged regions, which include $2b$ , $1b$ and $ob$ as well, (bottom right) as a function of signal resonance mass hypothesis for selection cuts. The efficiencies are relative to the total number of events in the preselection. . . . .	61
6.3	$m_j^{lead}$ vs. $m_j^{subl}$ in data in the $1b$ -tag (top) and $2b$ -tag (bottom) selection, the plots show the boundary between the Sideband (left) and Control (right) regions. . . . .	64
6.4	comparison between the $2b$ , $3b$ , and $4b$ shapes for the di-large- $R$ -jet mass distributions (the final discriminant) in the SR. . . . .	68
6.5	Simultaneous fit of $\mu_{multijet}$ and $\alpha_{t\bar{t}}$ in $4b$ (top) and $3b$ (middle) and $2b$ (bottom) sideband region using leading $large - R$ calorimeter jet mass spectrum. . . . .	71
6.6	Comparison of different trackjet $p_T$ distributions. Top row is for leading $p_T$ Higgs candidate, and bottom row is for subleading $p_T$ Higgs candidate. Left column is for the leading $p_T$ trackjet of the Higgs candidate, and right column is for the subleading $p_T$ trackjet of the Higgs candidate. Shown in the plot are just data distributions, inclusive of SB, CR, and SR regions for $ob$ and $1b$ , while for $2bs$ only the SB region is shown. $1b$ sample is further split into four subcategories, depending on which trackjet gets $b$ tagged. OneTag lead on lead means the $b$ tagged trackjet is the leading trackjet of the leading Higgs candidate, OneTag lead on subl means the $b$ tagged trackjet is the subleading trackjet of the leading Higgs candidate, OneTag subl on lead means the $b$ tagged trackjet is the leading trackjet of the subleading Higgs candidate, and OneTag subl on subl means the $b$ tagged trackjet is the subleading trackjet of the subleading Higgs candidate. At the bottom ratio plot, all the ratio are taken with respect to the $ob$ tagged distribution. . . . .	73



- 6.10 For  $3b$  background estimate: the fits to the ratio of the data in the  $2b$  category, of the leading Higgs candidate  $2b$ -tagged events's leading Higgs candidate distributions(black point), over the subleading Higgs candidate  $1b$ -tagged events's leading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$  jet  $p_T$  (left), the large- $R$  jet's leading trackjet  $p_T$  (middle), and large- $R$  jet's subleading trackjet  $p_T$  (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottow row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . . 80
- 6.11 For  $4b$  background estimate: the fits to the ratio of the data in the  $2b$  category, of the sub-leading Higgs candidate  $2b$ -tagged events's subleading Higgs candidate distributions(black point), over the leading Higgs candidate  $2b$ -tagged events's subleading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$  jet  $p_T$  (left), the large- $R$  jet's leading trackjet  $p_T$  (middle), and large- $R$  jet's subleading trackjet  $p_T$  (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottow row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . . 81
- 6.12 For  $4b$  background estimate: the fits to the ratio of the data in the  $2b$  category, of the leading Higgs candidate  $2b$ -tagged events's leading Higgs candidate distributions(black point), over the subleading Higgs candidate  $2b$ -tagged events's leading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$  jet  $p_T$  (left), the large- $R$  jet's leading trackjet  $p_T$  (middle), and large- $R$  jet's subleading trackjet  $p_T$  (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottow row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . . 82
- 6.13 Reweighted  $2bs$  Sideband region predictions comaprison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . . 84

6.14	Reweighted $3b$ Sideband region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$ jet $p_T$ , third row is the leading large- $R$ jet's leading trackjet $p_T$ and the last row subleading large- $R$ jet's leading trackjet $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . . .	85
6.15	Reweighted $4b$ Sideband region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$ jet $p_T$ , third row is the leading large- $R$ jet's leading trackjet $p_T$ and the last row subleading large- $R$ jet's leading trackjet $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . . .	86
6.16	Reweighted $2bs$ Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$ jet $p_T$ , third row is the leading large- $R$ jet's leading trackjet $p_T$ and the last row subleading large- $R$ jet's leading trackjet $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . . .	87
6.17	Reweighted $3b$ Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$ jet $p_T$ , third row is the leading large- $R$ jet's leading trackjet $p_T$ and the last row subleading large- $R$ jet's leading trackjet $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . . .	88
6.18	Reweighted $4b$ Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$ jet $p_T$ , third row is the leading large- $R$ jet's leading trackjet $p_T$ and the last row subleading large- $R$ jet's leading trackjet $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . . .	89
6.19	Kinematics of the lead large- $R$ jet in data and prediction in the sideband region after requiring $4 b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	91
6.20	Kinematics of the sub-lead large- $R$ jet in data and prediction in the sideband region after requiring $4 b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	92
6.21	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the sideband region after requiring $4 b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	93

6.22	Kinematics of the large- $R$ jet system in data and prediction in the sideband region after requiring 4 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	94
6.23	Kinematics of the lead large- $R$ jet in data and prediction in the sideband region after requiring 3 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	95
6.24	Kinematics of the sub-lead large- $R$ jet in data and prediction in the sideband region after requiring 3 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	96
6.25	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the sideband region after requiring 3 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	97
6.26	Kinematics of the large- $R$ jet system in data and prediction in the sideband region after requiring 3 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	98
6.27	Kinematics of the lead large- $R$ jet in data and prediction in the sideband region after requiring 2 $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	99
6.28	Kinematics of the sub-lead large- $R$ jet in data and prediction in the sideband region after requiring 2 $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	100
6.29	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the sideband region after requiring 2 $b$ -tags split. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	101
6.30	Kinematics of the large- $R$ jet system in data and prediction in the sideband region after requiring 2 $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	102

6.31	Kinematics of the lead large- $R$ jet in data and prediction in the control region after requiring 4 $b$ -tags. . . . .	104
6.32	Kinematics of the sub-lead large- $R$ jet in data and prediction in the control region after requiring 4 $b$ -tags. . . . .	105
6.33	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the control region after requiring 4 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	106
6.34	Kinematics of the large- $R$ jet system in data and prediction in the control region after requiring 4 $b$ -tags. . . . .	107
6.35	Kinematics of the lead large- $R$ jet in data and prediction in the control region after requiring 3 $b$ -tags. . . . .	108
6.36	Kinematics of the sub-lead large- $R$ jet in data and prediction in the control region after requiring 3 $b$ -tags. . . . .	109
6.37	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the control region after requiring 3 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	110
6.38	Kinematics of the large- $R$ jet system in data and prediction in the control region after requiring 3 $b$ -tags. . . . .	III
6.39	Kinematics of the lead large- $R$ jet in data and prediction in the control region after requiring 2 $b$ -tags split. . . . .	II2
6.40	Kinematics of the sub-lead large- $R$ jet in data and prediction in the control region after requiring 2 $b$ -tags split. . . . .	II3
6.41	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the control region after requiring 2 $b$ -tags split. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	II4
6.42	Kinematics of the large- $R$ jet system in data and prediction in the control region after requiring 2 $b$ -tags split. . . . .	II5

6.43	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requiring 4 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	117
6.44	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 4 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	118
6.45	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 4 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. Data is blinded, and will be added after unblinding. . . . .	119
6.46	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 4 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	120
6.47	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	121
6.48	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	122
6.49	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. Data is blinded, and will be added after unblinding. . . . .	123
6.50	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 3 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	124
6.51	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. Data is blinded, and will be added after unblinding. . . . .	125
6.52	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. Data is blinded, and will be added after unblinding. . . . .	126
6.53	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. Data is blinded, and will be added after unblinding. . . . .	127
6.54	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 2 $b$ -tags split. Data is blinded, and will be added after unblinding. . . . .	128

6.55 Comparison of the $4b$ , $3b$ and $2bs$ signal region $t\bar{t}$ dijet mass shape. On the left is the linear scale, and on the right is the log scale. Both distributions are normalized to 1 for comparison. . . . .	130
6.56 Fits for background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $4b$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	131
6.57 Fits for background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $3b$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	132
6.58 Fits for background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $2bs$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	133
6.59 Smoothed background estimations the $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal regions. Only smoothing statistical uncertainties are shown here. . . . .	134
6.60 Normalized Scaled dijet mass distributions for the $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal regions. For comparison, the unscaled distributions are shown on the same plot. . . . .	136
6.61 Fits for scaled background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $4b$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	137

6.62	Fits for scaled background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $3b$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	138
6.63	Fits for scaled background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $2bs$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	139
6.64	Smoothed MJJ (left) and scaled MJJ (right) background estimations the $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal regions. Smoothing statistical and systematic uncertainties from smoothing parameter variations are shown here. . . . .	140
6.65	Background prediction for $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal region using scaled di-jet mass before smoothing. The uncertainty band includes only statistical uncertainties. . . . .	141
6.66	Background prediction for $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal region using scaled di-jet mass after smoothing. The uncertainty band includes only statistical uncertainties. . . . .	142
7.1	$m_j^{lead}$ vs. $m_j^{subl}$ in data in the $1b$ -tag (top) and $2b$ -tag (bottom) selection, the plots show the boundary between the Sideband (left) and Control (right) regions. . . . .	148
7.2	comparison between the $2b$ , $3b$ , and $4b$ shapes for the di-large- $R$ -jet mass distributions (the final discriminant) in the SR. . . . .	151
7.3	Simultaneous fit of $\mu_{\text{multijet}}$ and $\alpha_{t\bar{t}}$ in $4b$ (top) and $3b$ (middle) and $2b$ (bottom) sideband region using leading $large - R$ calorimeter jet mass spectrum. . . . .	154

7.4	Comparison of different trackjet $p_T$ distributions. Top row is for leading $p_T$ Higgs candidate, and bottom row is for subleading $p_T$ Higgs candidate. Left column is for the leading $p_T$ trackjet of the Higgs candidate, and right column is for the subleading $p_T$ trackjet of the Higgs candidate. Shown in the plot are just data distributions, inclusive of SB, CR, and SR regions for $ob$ and $1b$ , while for $2bs$ only the SB region is shown. $1b$ sample is further split into four subcategories, depending on which trackjet gets $b$ tagged. OneTag lead on lead means the $b$ tagged trackjet is the leading trackjet of the leading Higgs candidate, OneTag lead on subl means the $b$ tagged trackjet is the subleading trackjet of the leading Higgs candidate, OneTag subl on lead means the $b$ tagged trackjet is the leading trackjet of the subleading Higgs candidate, and OneTag subl on subl means the $b$ tagged trackjet is the subleading trackjet of the subleading Higgs candidate. At the bottom ratio plot, all the ratio are taken with respect to the $ob$ tagged distribution. . . . .	156
7.5	For $2bs$ background estimate: the fits to the ratio of the data in the $1b$ category, of the subleading Higgs candidate $1b$ -tagged events's subleading Higgs candidate distributions(black point), over the leading Higgs candidate $1b$ -tagged events's subleading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$ jet $p_T$ (left), the large- $R$ jet's leading trackjet $p_T$ (middle), and large- $R$ jet's subleading trackjet $p_T$ (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottow row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . .	160
7.6	For $2bs$ background estimate: the fits to the ratio of the data in the $1b$ category, of the leading Higgs candidate $1b$ -tagged events's leading Higgs candidate distributions(black point), over the subleading Higgs candidate $1b$ -tagged events's leading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$ jet $p_T$ (left), the large- $R$ jet's leading trackjet $p_T$ (middle), and large- $R$ jet's subleading trackjet $p_T$ (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottow row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . .	161

- 7.7 For  $3b$  background estimate: the fits to the ratio of the data in the  $2b$  category, of the subleading Higgs candidate  $2b$ -tagged events's subleading Higgs candidate distributions(black point), over the leading Higgs candidate  $1b$ -tagged events's subleading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$  jet  $p_T$  (left), the large- $R$  jet's leading trackjet  $p_T$  (middle), and large- $R$  jet's subleading trackjet  $p_T$  (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottom row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . . 162
- 7.8 For  $3b$  background estimate: the fits to the ratio of the data in the  $2b$  category, of the leading Higgs candidate  $2b$ -tagged events's leading Higgs candidate distributions(black point), over the subleading Higgs candidate  $1b$ -tagged events's leading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$  jet  $p_T$  (left), the large- $R$  jet's leading trackjet  $p_T$  (middle), and large- $R$  jet's subleading trackjet  $p_T$  (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottom row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . . 163
- 7.9 For  $4b$  background estimate: the fits to the ratio of the data in the  $2b$  category, of the subleading Higgs candidate  $2b$ -tagged events's subleading Higgs candidate distributions(black point), over the leading Higgs candidate  $2b$ -tagged events's subleading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$  jet  $p_T$  (left), the large- $R$  jet's leading trackjet  $p_T$  (middle), and large- $R$  jet's subleading trackjet  $p_T$  (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottom row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . . 164

- 7.10 For  $4b$  background estimate: the fits to the ratio of the data in the  $2b$  category, of the leading Higgs candidate  $2b$ -tagged events's leading Higgs candidate distributions(black point), over the subleading Higgs candidate  $2b$ -tagged events's leading Higgs candidate distributions(yellow). Distributions and fits to the estimated QCD background for large- $R$  jet  $p_T$  (left), the large- $R$  jet's leading trackjet  $p_T$  (middle), and large- $R$  jet's subleading trackjet  $p_T$  (right) are shown. Figure are shown before reweighting (top row), after the first iteration(second row), after the fourth iteration(third row), and after the last iteration (bottom row). The green line is the spline extrapolation; and the red line is a polynomial fit. . . . . 165
- 7.11 Reweighted  $2bs$  Sideband region predictions comaprison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . 167
- 7.12 Reweighted  $3b$  Sideband region predictions comaprison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . 168
- 7.13 Reweighted  $4b$  Sideband region predictions comaprison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . 169
- 7.14 Reweighted  $2bs$  Control region predictions comaprison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . 170
- 7.15 Reweighted  $3b$  Control region predictions comaprison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . 171
- 7.16 Reweighted  $4b$  Control region predictions comaprison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting. . . 172

7.17 Kinematics of the lead large- $R$ jet in data and prediction in the sideband region after requiring 4 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	174
7.18 Kinematics of the sub-lead large- $R$ jet in data and prediction in the sideband region after requiring 4 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	175
7.19 First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the sideband region after requiring 4 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	176
7.20 Kinematics of the large- $R$ jet system in data and prediction in the sideband region after requiring 4 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	177
7.21 Kinematics of the lead large- $R$ jet in data and prediction in the sideband region after requiring 3 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	178
7.22 Kinematics of the sub-lead large- $R$ jet in data and prediction in the sideband region after requiring 3 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	179
7.23 First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the sideband region after requiring 3 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	180
7.24 Kinematics of the large- $R$ jet system in data and prediction in the sideband region after requiring 3 $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	181
7.25 Kinematics of the lead large- $R$ jet in data and prediction in the sideband region after requiring 2 $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	182

7.26 Kinematics of the sub-lead large- $R$ jet in data and prediction in the sideband region after requiring 2 $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	183
7.27 First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the sideband region after requiring 2 $b$ -tags split. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	184
7.28 Kinematics of the large- $R$ jet system in data and prediction in the sideband region after requiring 2 $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction. . . . .	185
7.29 Kinematics of the lead large- $R$ jet in data and prediction in the control region after requiring 4 $b$ -tags. . . . .	187
7.30 Kinematics of the sub-lead large- $R$ jet in data and prediction in the control region after requiring 4 $b$ -tags. . . . .	188
7.31 First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the control region after requiring 4 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	189
7.32 Kinematics of the large- $R$ jet system in data and prediction in the control region after requiring 4 $b$ -tags. . . . .	190
7.33 Kinematics of the lead large- $R$ jet in data and prediction in the control region after requiring 3 $b$ -tags. . . . .	191
7.34 Kinematics of the sub-lead large- $R$ jet in data and prediction in the control region after requiring 3 $b$ -tags. . . . .	192
7.35 First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the control region after requiring 3 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	193
7.36 Kinematics of the large- $R$ jet system in data and prediction in the control region after requiring 3 $b$ -tags. . . . .	194

7.37	Kinematics of the lead large- $R$ jet in data and prediction in the control region after requiring 2 $b$ -tags split. . . . .	195
7.38	Kinematics of the sub-lead large- $R$ jet in data and prediction in the control region after requiring 2 $b$ -tags split. . . . .	196
7.39	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the control region after requiring 2 $b$ -tags split. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	197
7.40	Kinematics of the large- $R$ jet system in data and prediction in the control region after requiring 2 $b$ -tags split. . . . .	198
7.41	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requiring 4 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	200
7.42	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 4 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	201
7.43	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 4 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. Data is blinded, and will be added after unblinding. . . . .	202
7.44	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 4 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	203
7.45	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	204
7.46	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	205
7.47	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. Data is blinded, and will be added after unblinding. . . . .	206
7.48	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 3 $b$ -tags. Data is blinded, and will be added after unblinding. . . . .	207

7.49	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. Data is blinded, and will be added after unblinding. . . . .	208
7.50	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. Data is blinded, and will be added after unblinding. . . . .	209
7.51	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. Data is blinded, and will be added after unblinding. . . . .	210
7.52	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 2 $b$ -tags split. Data is blinded, and will be added after unblinding. . . . .	211
7.53	Comparison of the 4 $b$ , 3 $b$ and 2 $bs$ signal region $t\bar{t}$ dijet mass shape. On the left is the linear scale, and on the right is the log scale. Both distributions are normalized to 1 for comparison. . . . .	213
7.54	Fits for background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the 4 $b$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	214
7.55	Fits for background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the 3 $b$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	215
7.56	Fits for background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the 2 $bs$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	216
7.57	Smoothed background estimations the 4 $b$ (top), 3 $b$ (middle), and 2 $bs$ (bottom) signal regions. Only smoothing statistical uncertainties are shown here. . . . .	217

7.58 Normalized Scaled dijet mass distributions for the $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal regions. For comparison, the unscaled distributions are shown on the same plot. . . . .	219
7.59 Fits for scaled background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $4b$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	220
7.60 Fits for scaled background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $3b$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	221
7.61 Fits for scaled background smoothing are shown for QCD (top row) and $t\bar{t}$ (bottom row) in the $2bs$ signal region. The left figures show the distributions with linear $y$ -axis scale along with the fit central value and variations. The right figures show the distributions with log $y$ -axis scale along with the fit central value and the fit variations as determined by the varying the fit parameters within uncertainties whilst taking into account parameter correlations. . . . .	222
7.62 Smoothed MJJ (left) and scaled MJJ (right) background estimations the $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal regions. Smoothing statistical and systematic uncertainties from smoothing parameter variations are shown here. . . . .	223
7.63 Background prediction for $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal region using scaled di-jet mass before smoothing. The uncertainty band includes only statistical uncertainties. . . . .	224
7.64 Background prediction for $4b$ (top), $3b$ (middle), and $2bs$ (bottom) signal region using scaled di-jet mass after smoothing. The uncertainty band includes only statistical uncertainties. . . . .	225
8.1 Total background estimation ( $qcd + t\bar{t}$ ) with different $t\bar{t}MC$ variations. The different variations agree with the default within the statistical uncertainties. . . . .	230

8.2	Illustration of ZZ (left) and TT (right) signal region as shown in the orange shaded region. Control region shown in green, and Sideband region in blue. The white circle in the midde is the real Signal region, and it is blinded. . . . .	237
8.3	ZZ signal region distribution of di-jet mass (left column) and leading large-R jet mass (right column) in low mass signal, for $4b$ (top row), $3b$ (middle row) and $2b$ split (bottom row). The plots are with only statistical uncertainty. . . . .	240
8.4	TT signal region distribution of di-jet mass (left column) and leading large-R jet mass (right column) in low mass signal, for $4b$ (top row), $3b$ (middle row) and $2b$ split (bottom row). The plots are with only statistical uncertainty. . . . .	242
8.5	(left) Shape of the $t\bar{t}$ di-large- $R$ -jet mass in the sideband region, comparing the $3b$ shape with that of the $2b$ , in order to asses the systematic effect of additional $b$ -tags changing the dijet mass distribution. The $m_{JJ}$ distributions is shown on the left, and the ratio of $3b$ to $2b$ distributions on the right. . . . .	244
8.6	Dijet mass distribution in the CR along with the prediction (left) and the ratio of the prediction to the CR distribution (right) for the $2bs$ (top) $3b$ (middle) and $4b$ (bottom) samples. Ratios are from the smoothed distributions, the data uncertainty band contains the smoothing parameter variations, and the prediction uncertainty band also contains smoothing parameter variations. . . . .	246
8.7	Dijet mass distribution SR prediction fit with several fit ranges (left) and the ratio of nominal to fits with different fit ranges (right) for the $2b$ (top) $3b$ (middle) and $4b$ (bottom) samples. . . . .	248
8.8	Dijet mass distribution SR prediction fit with several fit functions (left) and the ratio of nominal to fits with different fit functions (right) for the $2b$ (top) $3b$ (middle) and $4b$ (bottom) samples. The additional fit functions are from Table 8.12. . . . .	250
8.9	Impact of each systematic on the signal prediction as a function of the signal mass, in the $4b$ (left) and $3b$ (middle) and $2bs$ signal regions. . . . .	253
8.10	The total background estimation in $4b$ signal region, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down. . . . .	254
8.11	The total background estimation in $3b$ signal region, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down. . . . .	254

8.12	The total background estimation in $2b$ s signal region, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down. . . . .	255
8.13	The total background estimation in $4b$ signal region, scaled mJJ, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down. . . . .	255
8.14	The total background estimation in $3b$ signal region, scaled mJJ, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down. . . . .	256
8.15	The total background estimation in $2b$ s signal region, scaled mJJ, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down. . . . .	256
9.1	Unscaled dijet mass distribution in the $4b$ Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucn- certainties are shown on the plot. . . . .	260
9.2	Unscaled dijet mass distribution in the $3b$ Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucn- certainties are shown on the plot. . . . .	260
9.3	Unscaled dijet mass distribution in the $2b$ s Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucn- certainties are shown on the plot. . . . .	261
9.4	Scaled dijet mass distribution in the $4b$ Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucner- tainties are shown on the plot. . . . .	261
9.5	Scaled dijet mass distribution in the $3b$ Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucner- tainties are shown on the plot. . . . .	261
9.6	Scaled dijet mass distribution in the $2b$ s Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucn- certainties are shown on the plot. . . . .	262
9.7	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requir- ing $4 b$ -tags. . . . .	263

9.8	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 4 $b$ -tags. . . . .	264
9.9	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 4 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	265
9.10	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 4 $b$ -tags. . . . .	266
9.11	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. . . . .	267
9.12	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. . . . .	268
9.13	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 3 $b$ -tags. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	269
9.14	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 3 $b$ -tags. . . . .	270
9.15	Kinematics of the lead large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. . . . .	271
9.16	Kinematics of the sub-lead large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. . . . .	272
9.17	First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$ track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$ jet in data and prediction in the signal region after requiring 2 $b$ -tags split. Third row shows the $\Delta R$ between two leading small- $R$ track-jets associated to the leading (left) and sub-leading (right) large- $R$ jet. . . . .	273
9.18	Kinematics of the large- $R$ jet system in data and prediction in the signal region after requiring 2 $b$ -tags split. . . . .	274
9.19	Local $p_o$ of the (a) scalar, (b) $c=1$ Graviton and (c) $c=2$ Graviton. . . . .	275

9.20	The expected and observed 95% C.L. upper exclusion limits for the boosted $4b$ analysis calculated including all systematic uncertainties for the narrow scalar model. The dot-dashed line shows the expected limit when only statistical uncertainties are included. The limits are derived within the asymptotic approximation. . . . .	276
9.21	The expected and observed 95% C.L. upper exclusion limits for the boosted $4b$ analysis calculated including all systematic uncertainties for the $c=1.0$ Graviton. The dot-dashed line shows the expected limit when only statistical uncertainties are included. The limits are derived within the asymptotic approximation. . . . .	277
9.22	The expected and observed 95% C.L. upper exclusion limits for the boosted $4b$ analysis calculated including all systematic uncertainties for the $c=2.0$ Graviton. The dot-dashed line shows the expected limit when only statistical uncertainties are included. The limits are derived within the asymptotic approximation. . . . .	278
9.23	Nuisance parameters associated with the background modelling, after the conditional likelihood fit for a bulk RS graviton signal with $m_{G_{KK}^*} = 2$ TeV and $k/\bar{M}_{\text{Pl}} = 1.0$ . The tight constraints of $2b\_QCD\_CRShape$ and $3b\_QCD\_CRShape$ are a result of the nuisance parameter prior being unconstrained due to a lack of control region data at high mass. . . . .	280
9.24	Postfit distributions after fitting the data with the 2000 GeV signal hypothesis. The signal strength is slightly positive. . . . .	281
9.25	Postfit distributions after fitting the data with the 2500 GeV signal hypothesis. The signal strength is zero. . . . .	282
9.26	The impact of nuisance parameters on the fitted cross section, ranked by their postfit impact. The signal mass used in this fits is 2000 GeV, and the signal model is (a) narrow scalar, (b) $c=1$ Graviton and (c) $c=2$ Graviton. . . . .	283

# Listing of tables

2.1	LHC nominal and operational parameters . . . . .	20
5.1	Physics objects and their technical names in the boosted analysis. . . . .	44
5.2	The selection efficiency for $G_{KK}^* \rightarrow hh \rightarrow b\bar{b}b\bar{b}$ events ( $c = 1.0$ ) at each stage of the event selection. Uncertainties are the MC stat uncertainty only. . . . .	56
5.3	The selection efficiency for $G_{KK}^* \rightarrow hh \rightarrow b\bar{b}b\bar{b}$ events ( $c = 2.0$ ) at each stage of the event selection. Uncertainties are the MC stat uncertainty only. . . . .	56
5.4	The selection efficiency for $H \rightarrow hh \rightarrow b\bar{b}b\bar{b}$ events at each stage of the event selection. . . . .	57
6.1	Definitions of the Signal (SR), Sideband (SB) and Control (CR) regions. . . . .	63
6.2	Background scaling parameters ( $\mu_{\text{multijet}}$ and $\alpha_{t\bar{t}}$ ) estimated from fits to the leading jet mass distributions in $4b/3b/2bs$ sideband regions. $\rho(\mu_{qcd}, \alpha_{t\bar{t}}) = \frac{\text{Cov}(qcd, t\bar{t})}{\sqrt{qcd} \sqrt{t\bar{t}}}$ . . .	69
6.3	Smoothing parameters in $4b$ and $3b$ and $2bs$ signal regions, the correlation between parameters is almost always 0.99. . . . .	130
6.4	Smoothing parameters in $4b$ and $3b$ and $2bs$ signal regions for scaled mass distributions, the correlation between parameters is almost always 0.99. . . . .	143
7.1	Definitions of the Signal (SR), Sideband (SB) and Control (CR) regions. . . . .	147
7.2	Background scaling parameters ( $\mu_{\text{multijet}}$ and $\alpha_{t\bar{t}}$ ) estimated from fits to the leading jet mass distributions in $4b/3b/2bs$ sideband regions. $\rho(\mu_{qcd}, \alpha_{t\bar{t}}) = \frac{\text{Cov}(qcd, t\bar{t})}{\sqrt{qcd} \sqrt{t\bar{t}}}$ . . .	152
7.3	Smoothing parameters in $4b$ and $3b$ and $2bs$ signal regions, the correlation between parameters is almost always 0.99. . . . .	213
7.4	Smoothing parameters in $4b$ and $3b$ and $2bs$ signal regions for scaled mass distributions, the correlation between parameters is almost always 0.99. . . . .	226
8.1	Agreement between data and prediction in $4b$ tag CR. Showing stat uncertainty only. . . . .	235
8.2	Agreement between data and prediction in $3b$ tag CR. Showing stat uncertainty only. . . . .	235
8.3	Agreement between data and prediction in $2bs$ tag CR. Showing stat uncertainty only. . . . .	235

8.4	Background prediction in SR/CR/SB for ZZ SR in $4b$ -tag region. Uncertainties are stat only. . . . .	238
8.5	Background prediction in SR/CR/SB for ZZ SR in $3b$ -tag region. Uncertainties are stat only. . . . .	238
8.6	Background prediction in SR/CR/SB for ZZ SR in $2bs$ -tag region. Uncertainties are stat only. . . . .	238
8.7	Agreement between data and prediction in ZZ SR in $4b$ , $3b$ and $2bs$ regions. . . . .	239
8.8	Background prediction in SR/CR/SB for TT SR in $4b$ -tag region. Uncertainties are stat only. . . . .	239
8.9	Background prediction in SR/CR/SB for TT SR in $3b$ -tag region. Uncertainties are stat only. . . . .	239
8.10	Background prediction in SR/CR/SB for TT SR in $2bs$ -tag region. Uncertainties are stat only. . . . .	241
8.11	Agreement between data and prediction in TT SR in $4b$ , $3b$ and $2bs$ regions. . . . .	241
8.12	Functions used to fit the QCD dijet mass distributions, where $x = m_{jj}/\sqrt{s}$ . . . . .	249
8.13	Percent impact of the dominant systematics on the background acceptance and on the signal acceptance of RS $c = 1.0$ graviton predictions in the $4b$ signal region. . . . .	252
8.14	Percent impact of the dominant systematics on the background acceptance and on the signal acceptance of RS $c = 1.0$ graviton predictions in the $3b$ signal region. . . . .	252
8.15	Percent impact of the dominant systematics on the background acceptance and on the signal acceptance of RS $c = 1.0$ graviton predictions in the $2bs$ signal region. . . . .	253
9.1	Unblinded Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. . . . .	258
9.2	$4b$ unblinded Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals. . . . .	258
9.3	$3b$ unblinded Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals. . . . .	258
9.4	$2bs$ unblinded Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals. . . . .	258

9.5	$4b$ unblinded Scaled dijet mass Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals. . . . .	258
9.6	$3b$ unblinded Scaled dijet mass Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals. . .	259
9.7	$2bs$ unblinded Scaled dijet mass Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals. . .	259

VERITAS SHALL MAKE YOU FREE.

# Acknowledgments

THANKS TO EVERYONE WORKING AT CERN. CERN is a unique and special place. I love how the streets are named by physicists. Without the support from the IT department, I could not log into lxplus, check the twikis and finish my work. CERN user's office also made traveling to Europe a much nicer and easier experience for me.

THANKS TO EVERYONE WORKING ON THE LARGE HADRON COLLIDER. Without a fully functional accelerator, there would have been no data for me to study. The LHC performed outstandingly since 2015, and all the valuable data in this thesis is produced from it.

THANKS TO EVERYONE THE ATLAS COLLABORATION, who has supported this remarkable program and has contributed to every bit of the result in my thesis. I am standing on the ATLAS(member)'s shoulders—eight floors high and don't shrug. Without the excellent work on detector design, commissioning, operational works, reconstruction, data processing, performance studies and recommendations, software support, computing support, and analysis discussions and guidance, I could not have completed this thesis. I sincerely thank all ATLAS members for their work.

*It was the best of times, it was the worst of times, it was  
the age of wisdom, it was the age of foolishness, it was  
the epoch of belief, it was the epoch of incredulity, it  
was the season of Light, it was the season of Darkness,  
it was the spring of hope, it was the winter of despair,  
we had everything before us, we had nothing before us,  
we were all going direct to Heaven, we were all going  
direct the other way—in short, the period was so far like  
the present period, that some of its noisiest authorities  
insisted on its being received, for good or for evil, in the  
superlative degree of comparison only.*

Charles Dickens

# 0

## Introduction

In 2012, the Higgs boson was discovered by the ATLAS and CMS experiment at the LHC. The particle physics community faces a period just like at the beginning of *A Tale of Two Cities*.

After Run 1 of the LHC, with the existence of the Higgs now firmly established, the focus shifted to searches for physics beyond the Standard Model.

In particular, searches for high mass resonances benefit from the LHC’s increase to  $\sqrt{s} = 13$  TeV in Run 2. The cross section for a generic gluon-initiated resonance with a mass of 2 TeV increases tenfold in Run 2, making searches for high mass resonances a high priority. The newly discovered Higgs can be used as a tool in these searches. After the discovery, the Higgs boson provides a large swath of unmeasured phase space where new physics could be discovered. Higgs pair production in the Standard Model has a low cross section that requires large datasets (on the order of the

LHC’s lifetime) for full measurement. However, new physics can modify this cross section, especially through new resonances which decay to two Higgs bosons. Such high mass resonances also produce difficult to recognize final state topologies due to the merging of decay products from high momentum Higgs bosons. A search for Higgs pair production in the  $HH \rightarrow b\bar{b}b\bar{b}$  final state was performed with  $3.2\text{fb}^{-1}$  collected with ATLAS at  $\sqrt{s} = 13$  TeV in 2015. The results are presented in this dissertation with a focus on a dedicated signal region for boosted final states. This signal region uses new techniques for recognizing jet substructure and  $b$ -tagging to the improve signal acceptance of high mass resonances.

The discovery of the Standard Model (SM) Higgs boson ( $h$ )<sup>2</sup> at the Large Hadron Collider (LHC) motivates searches for new physics using the Higgs boson as a probe. In particular, many models predict cross sections for Higgs boson pair production that are significantly greater than the SM prediction. Resonant Higgs boson pair production is predicted by models such as the bulk Randall–Sundrum model<sup>7,8</sup>, which features spin-2 Kaluza–Klein gravitons,  $G_{KK}^*$ , that subsequently decay to a pair of Higgs bosons. Extensions of the Higgs sector, such as two-Higgs-doublet models<sup>9,10</sup>, propose the existence of a heavy spin-0 scalar that can decay into  $hpairs$ . Enhanced non-resonant Higgs boson pair production is predicted by other models, for example those featuring light coloured scalars<sup>11</sup> or direct  $t\bar{t}hh$  vertices<sup>12,13</sup>.

Previous searches for Higgs boson pair production have all yielded null results. In the  $b\bar{b}b\bar{b}$  channel, ATLAS searched for both non-resonant and resonant production in the mass range of 400–3000 GeV using  $3.2\text{ fb}^{-1}$  of 13 TeV data<sup>14</sup> collected during 2015. CMS searched for the production of resonances with masses of 750–3000 GeV<sup>2</sup> using 13 TeV data and with masses 270–1100 GeV with 8 TeV data<sup>2</sup>. Using 8 TeV data, ATLAS has examined the  $b\bar{b}b\bar{b}$ <sup>15</sup>,  $b\bar{b}\gamma\gamma$ <sup>16</sup>,  $b\bar{b}\tau^+\tau^-$  and  $W^+W^-\gamma\gamma$  channels, all of which were combined in Ref.<sup>2</sup>. CMS has performed searches using 13 TeV data for the  $b\bar{b}\tau^+\tau^-$ <sup>2</sup> and  $b\bar{b}\ell\nu\ell\nu$ <sup>2</sup> final states, and used 8 TeV data to search for  $b\bar{b}\gamma\gamma$ <sup>2</sup> in addition to a search in multilepton and multilepton+photons final states<sup>2</sup>.

The analyses presented in this paper exploit the dominant  $b \rightarrow b\bar{b}$  decay mode to search for Higgs boson pair production in both resonant and non-resonant production. Two analyses are presented, which are complementary in their acceptance, each employing a unique technique to reconstruct the Higgs boson. The “resolved” analysis is used for  $hh$  systems in which the Higgs bosons have Lorentz boosts low enough that four  $b$ -jets can be reconstructed. The “boosted” analysis is used for those  $hh$  systems in which the Higgs bosons have higher Lorentz boosts, which prevents the Higgs boson decay products from being resolved in the detector as separate  $b$ -jets. Instead, each Higgs boson candidate consists of a single large-radius jet, and  $b$ -decays are identified using smaller-radius jets built from charged-particle tracks.

Both analyses were re-optimized with respect to the former ATLAS publication <sup>14</sup>; an improved algorithm to pair  $b$ -jets to Higgs boson candidates is used in the resolved analysis, and in the boosted analysis an additional signal-enriched sample is utilized. The dataset comprises the 2015 and 2016 data, corresponding to  $27.5 \text{ fb}^{-1}$  for the resolved analysis and  $36.1 \text{ fb}^{-1}$  for the boosted analysis, with the difference due to the trigger selections used. The results are obtained using the resolved analysis for a resonance mass between 260 and 1400 GeV, and the boosted analysis between 800 GeV and 3000 GeV. The main background is multijet production, which is estimated from data; the sub-leading background is  $t\bar{t}$ , which is estimated using both data and simulations. The two analyses employ orthogonal selections, and a statistical combination is performed in the mass range where they overlap. The final discriminants are the four-jet and dijet mass distributions in the resolved and boosted analyses, respectively. Searches are performed for the following benchmark signals: a spin-2 graviton decaying into Higgs bosons, a scalar resonance decaying into a Higgs boson pair, and SM non-resonant Higgs boson pair production.

This dissertation begins by discussing the status of di-Higgs. Chapter 1 gives an overview of double Higgs production in the Standard Model and beyond. Chapter 2 and 3 present details regarding

the Large Hadron Collider and the ATLAS experiment. Chapter 4 provides an overview of object reconstruction in ATLAS, with a focus on Muon Segment Seeding. A brief interlude in Chapter 5 on the ATLAS Muon Data Quality, as this has been a focus of my graduate work.

The rest of the dissertation presents a search for Higgs pair production in the  $HH \rightarrow b\bar{b}b\bar{b}$  channel. Chapter 6 presents an overview of physics object selection, where the Higgs pairs are the result of the decay of a heavy resonance. Chapter 7 discusses the background estimation techninics in detail, followed by Chapter 8, Systematics. Chapter 9 presents the results, and Chapter 10 shows the limits between the boosted regime and the resolved regime, which is sensitive to lower mass resonances and non-resonant Higgs pair production. Finally, the work is summarized a conclusion and brief outlook of future Higgs physics with ATLAS.

*Knowledge knows no bounds.*

Creator

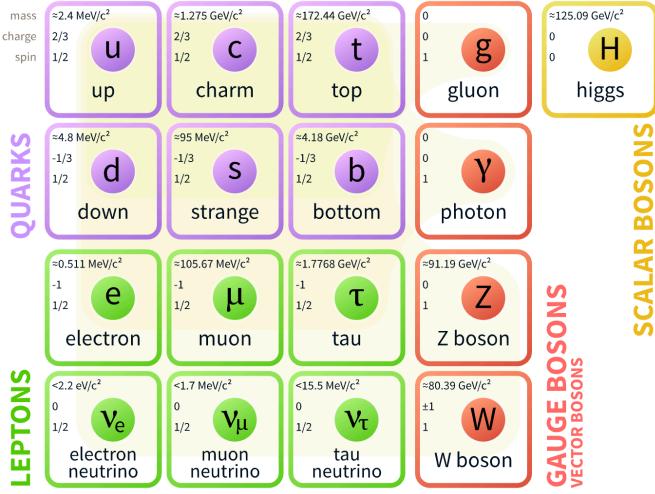
# 1

## Theory and Motivation

### 1.1 THE STANDARD MODEL AND THE HIGGS BOSON

The Standard Model(SM)<sup>1,17,18,19</sup> is a quantum field theory describing the interactions of fundamental particles. The particles are shown in Figure 1.1. So far, the SM predictions agree extremely well with experimental observations.

In the SM, the Higgs mechanism introduces a complex scalar Higgs field,  $\phi$ , with nonzero vacuum expectation values. The scalar Higgs potential is  $V(\phi) = -v^2\lambda_v\phi^\dagger\phi + \lambda_v(\phi^\dagger\phi)^2$ . Through spontaneous symmetry breaking,  $W^\pm$  and  $Z$  bosons acquire their masses. This process also predicts an extra scalar, the Higgs boson. The SM Lagrangian containing Higgs couplings,  $\mathcal{L}_{\text{Higgs}}$ , is shown



**Figure 1.1:** Fermions and bosons of the Standard Model and their properties<sup>1</sup>, where all the values are measured experimentally.

in Eq 1.1.

$$\mathcal{L}_{\text{Higgs}} = -\lambda_{h\bar{f}f} h\bar{f}f + \delta_V V_\mu V^\mu (\lambda_{hvv} b + \lambda_{hhvv} b^2) + \lambda_{hh} b^2 + \lambda_{hhh} b^3 + \lambda_{hhhh} b^4 \quad (1.1)$$

where

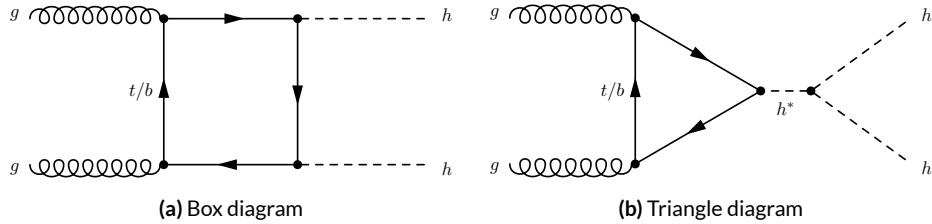
- $v \sim 246 \text{ GeV}$ , is the non-zero expectation value of the Higgs field;
- $m_h = \sqrt{2\lambda_v} v \sim 125 \text{ GeV}$ , is the Higgs mass; this is discovered in 2012<sup>20,21</sup>;
- $\lambda_v$ , coefficient for the quartic potential term, is constrained from Higgs mass, to be  $\sim -0.13$ ;
- $V = W^\pm$  or  $Z$ ,  $\delta_W = 1$ ,  $\delta_Z = \frac{1}{2}$ ;
- $\lambda_{h\bar{f}f} = \frac{m_f}{v}$ , is the Higgs to fermion coupling;  $m_f$  is the mass of the fermion;
- $\lambda_{hvv} = \frac{2m_v^2}{v}$ , is the Higgs to boson coupling;  $m_v$  is the mass of the boson;
- $\lambda_{hhvv} = \frac{m_v^2}{v^2}$ , is the Higgs-Higgs to boson-boson coupling;
- $\lambda_{hh} = \frac{m_h^2}{2}$ , is the Higgs mass term;

- $\lambda_{hhh} = \frac{m_h^2}{2\nu} = \lambda_{\nu\nu}$ , or  $\lambda_{hhh}$ , is the Higgs self-coupling;
- $\lambda_{hhh} = \frac{m_h^2}{8\nu^2}$ , is the Higgs quartic-coupling.

What's particularly interesting and has not been measured experimentally in Eq 1.1 is  $\lambda_{hhh}$ . SM predicts  $\lambda_{hhh} = \frac{m_h^2}{2\nu}$ , which is referred as  $\lambda_{SM}$  in this thesis. This term directly probes the Higgs potential. Also,  $\lambda_{hhh} h^3$  term shows one way for double Higgs production within the SM. Double Higgs production is also known as di-Higgs or Higgs pair production.

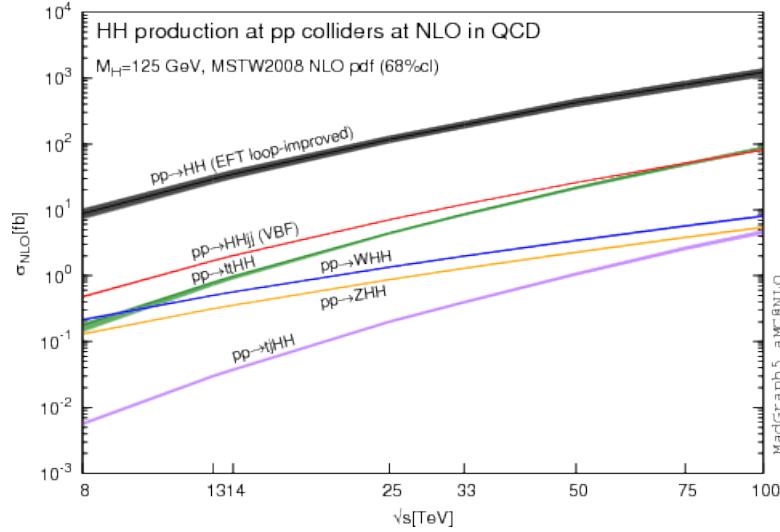
## 1.2 STANDARD MODEL DI-HIGGS PRODUCTION

There are two main production diagrams of di-Higgs at the LHC, shown in Figure 1.2. In the gluon-gluon fusion process, di-Higgs are produced through a box or a triangle loop. Only the triangle loop 1.2b probes the  $\lambda_{hhh}$ . In the triangle diagram, the middle Higgs boson acts as a propagator (off-shell), and the two Higgs boson in the final state are on-shell. An on-shell middle Higgs, with two off-shell Higgs bosons in the final state, is strongly disfavored<sup>1</sup>. The box and triangle diagrams interfere destructively, which makes the overall production rate smaller than what would be expected in the absence of a  $\lambda_{hhh}$  term.



**Figure 1.2:** Leading order Feynman diagrams contributing to di-Higgs production via gluon-gluon fusion, through the Higgs-fermion Yukawa interactions 1.2a and the Higgs boson self-coupling 1.2b. Only Figure 1.2b probes  $\lambda_{hhh}$ .

Many other different production modes of di-Higgs exist, but gluon-gluon fusion is the dominant one. Figure 1.3<sup>22</sup> compares the cross sections of gluon-gluon fusion, Vector Boson Fusion (VBF), and top-pair,  $W^\pm$ ,  $Z$  and single-top associated di-Higgs production.



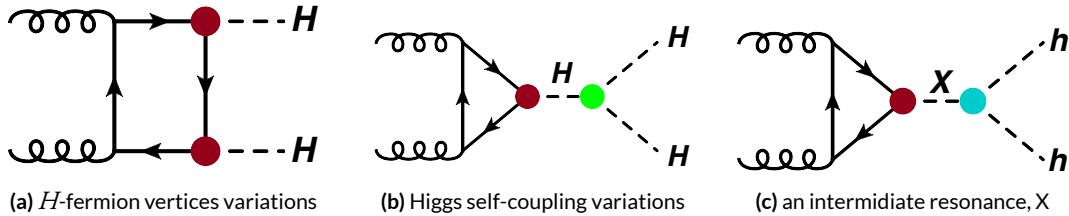
**Figure 1.3:** Total cross sections (y-axis) at the NLO in QCD for the six largest di-Higgs production channels at p-p colliders at different energy (x-axis). The thickness of the lines corresponds to the scale and PDF uncertainties added linearly.  $H$  refers to the SM Higgs.

For p-p collisions at  $\sqrt{s} = 13$  TeV, the total cross section for SM gluon fusion di-Higgs production<sup>23</sup>, evaluated at next-to-next-to-leading order (NNLO) with the summation of logarithms at next-to-next-leading-logarithm (NNLL) accuracy and including top-quark mass effects at NLO, is  $33.49^{+4.3\%}_{-6.0\%} \pm 2.1\% \pm 2.3\%$  fb. The cross section for the next dominant production, VBF, is  $1.62^{+2.3\%}_{-2.7\%} \pm 2.3\%$  fb. The estimated cross section for triple-Higgs production is  $0.06332^{+16.1\%}_{-14.1\%} \pm 3.4\%$  fb, which is negligible with current dataset. The uncertainties are Scale uncertainty, PDF uncertainty and  $\alpha_s$  uncertainty. This means inside 2015 and 2016  $\sqrt{s} = 13$  TeV ATLAS 36  $\text{fb}^{-1}$  data, there are only around one thousand SM di-Higgs events.

### 1.3 BEYOND THE STANDARD MODEL PHYSICS DI-HIGGS PRODUCTION

The SM works extremely well, yet the Higgs boson mass at 125 GeV requires extreme fine-tuning for radiative corrections. The presence of new physics at the TeV scale would help solve the naturalness problem.

BSM physics could significantly enhance the production of di-Higgs at the LHC. This is separated into two categories: non-resonant and resonant productions. The non-resonant production generally refers to modifications of the Higgs couplings, either the Higgs self-coupling or the Higgs-top couplings. Resonant production refers to a particle with invariant mass greater than twice the Higgs mass decays directly into two Higgs bosons. The difference also comes from the distribution of the di-Higgs invariant mass at the truth level. In the non-resonant case, the distribution has no clear peak, whereas in the resonant case, the invariant mass distribution usually forms a peak with model dependent width.



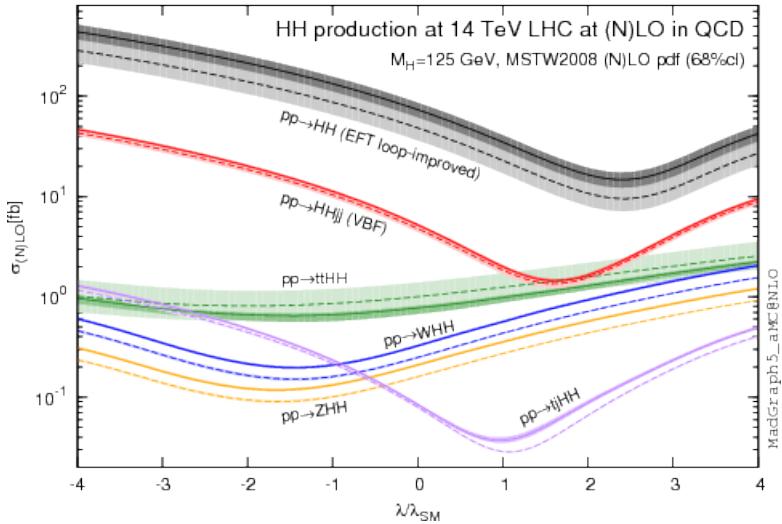
**Figure 1.4:** BSM Higgs boson pair production: non-resonant production proceeds through changes in the SM Higgs couplings in 1.4a and 1.4b, resonant production proceeds through 1.4c an intermediate resonance,  $X$ .  $H$  and  $h$  both refers to the SM Higgs.

#### 1.3.1 BSM NON-RESONANT DI-HIGGS

Enhanced non-resonant Higgs boson pair production is predicted by many models. Models featuring direct  $t\bar{t}hh$  vertices<sup>12,13</sup> or new light colored scalars<sup>11</sup> could change vertices shown as the red

dots in Figure 1.4. A direct modification of Higgs self-coupling term in Eq 1.1 to  $\lambda bhh$ , where  $\lambda$  is different from  $\lambda_{\text{SM}}$ , is also possible. This is shown as the green dot in Figure 1.4b.

The non-resonant di-Higgs enhancement is usually described by  $\frac{\lambda}{\lambda_{\text{SM}}}$ , which is the cross section ratio between  $\lambda$  and  $\lambda_{\text{SM}}$ . From the SM electroweak measurements, the self coupling term could be constrained to  $-14 \leq \frac{\lambda}{\lambda_{\text{SM}}} \leq 17.4$ <sup>24</sup>. Variations of  $\lambda$  have a non-trivial effect on di-Higgs production cross section, shown in Figure 1.5<sup>22</sup>. In the regime of relatively high trilinear coupling, the observation will be an excess of di-Higgs events with respect to the expected background. A simple limit can be set in this case.



**Figure 1.5:** Total cross sections (y-axis) at the LO and NLO in QCD for di-Higgs production channels, at the  $\sqrt{s} = 14$  TeV LHC as a function of the self-interaction coupling  $\lambda$  (x-axis). The dashed (solid) lines and light- (dark-) color bands correspond to the LO (NLO) results and to the scale and PDF uncertainties added linearly. The SM values of the cross sections are obtained at  $\frac{\lambda}{\lambda_{\text{SM}}} = 1$ .  $H$  refers to the SM Higgs.

### 1.3.2 BSM RESONANT DI-HIGGS

Resonant Higgs boson pair production is also predicted by many models. Extensions of the Higgs sector, such as two-Higgs-doublet models(2HDM)<sup>9,10</sup>, propose the existence of a heavy spin-0 scalar

$H$  that can decay into di-Higgs. The bulk Randall-Sundrum model<sup>7,8</sup>, which features spin-2 Kaluza-Klein gravitons,  $G_{\text{KK}}^*$ , could also subsequently decay to pairs of Higgs bosons. These proposed heavy particles, heavy CP-even scalar  $H$  and  $G_{\text{KK}}^*$ , are represented as X in Figure 1.4c.

The 2HDM is a simple extension of the SM which can exhibit large resonance effects<sup>23</sup>. The 2HDM has 5 physical Higgs bosons:  $h$  (light scalar Higgs),  $H$  (heavy scalar Higgs),  $A$  (heavy pseudoscalar Higgs), and  $H^\pm$  (two charged Higgs). The 2HDM can introduce tree level flavor changing neutral currents. To avoid this, models impose discrete symmetries in which the charged fermions only couple to one of the Higgs doublets. One version is type II 2HDM, in which all positively charged quarks couple to one doublet and the negatively charged quarks and leptons couple to the other. The type II model is the Minimal Supersymmetric Standard Model(MSSM)'s Higgs sector.

Resonant di-Higgs production in 2HDM models can proceed through decays of the heavy CP-even Higgs  $H \rightarrow hh$ . The branching ratio for  $H \rightarrow hh$  depends on the model type as well as the values of  $\tan \beta$  and  $\cos(\beta - \alpha)$ .  $\tan \beta = \frac{v_{\text{doublet}}}{v_{\text{SM}}}$  is the ratio of the vacuum expectation values of the two Higgs doublets.  $\alpha$  is the mixing angle between the heavy  $H$  and light  $h$  fields. The limit where  $\cos(\beta - \alpha) = 0$  is called the alignment limit, and in this limit the light Higgs  $h$  has the same couplings as a SM Higgs. Near the alignment limit there is some unprobed phase space depending on the exact models and values of  $\tan \beta$  being considered, and they are particularly interesting to be searched for at the LHC.

The Randall-Sundrum model proposes a five-dimensional warped spacetime that contains two manifolds: one where the force of gravity is very strong and a second manifold at the TeV scale corresponding to the known SM sector. The experimental consequence of this theory is a series of widely mass-spaced Kaluza-Klein graviton resonances,  $G_{\text{KK}}^*$ . In theories where the fermions are localized to the SM brane, production of gravitons from fermion pairs is suppressed and the primary

mode of production of  $G_{KK}^*$  is gluon fusion. These gravitons have a substantial branching fraction to di-Higgs, ranging from 6.43% for gravitons with a mass of 500 GeV to 7.66% at 3 TeV. Randall-Sundrum models have two free parameters - the mass of the graviton and  $c = k/\bar{M}_{\text{pl}}$ , where  $\bar{M}_{\text{pl}}$  is the reduced Planck mass and  $k$  is the curvature parameter. The width of the graviton increases with both mass and  $c$ .

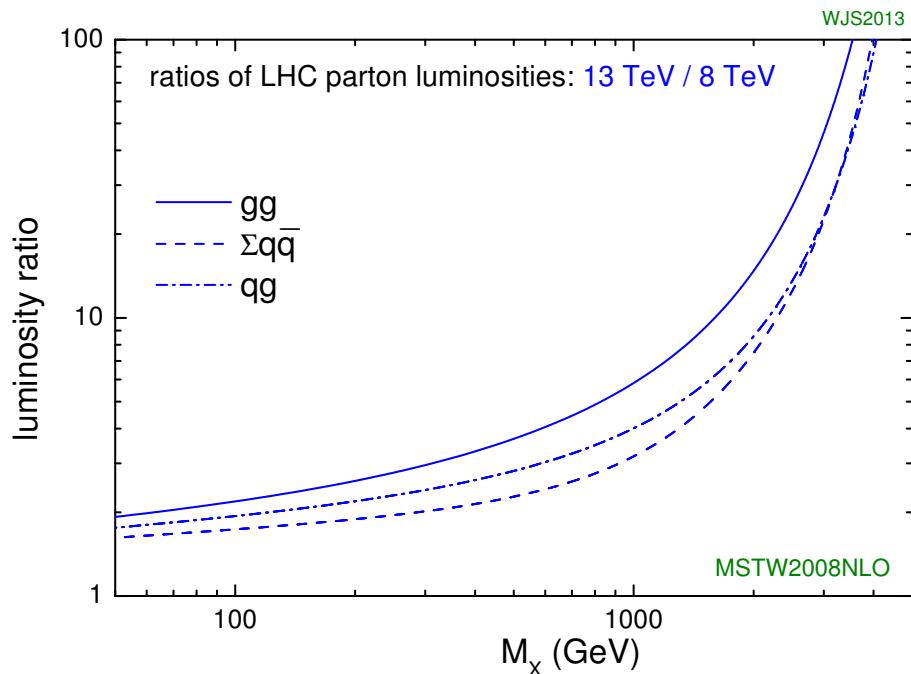


Figure 1.6: Parton luminosity ratios as a function of resonance mass  $M_X$  for 13/8 TeV<sup>2</sup>. For a 2 TeV  $X$ , the luminosity ratio is almost 10.

In model dependent searches, based on fixed assumptions of the resonance particles' branching ratio, other search channels like resonance  $VV$  or  $t\bar{t}$  are more sensitive compared to di-Higgs<sup>25</sup>. In order to constrain more BSM physics phase space, di-Higgs search results need to be interpreted in different baseline models, covering both narrow and wide resonances.

Generally, it is easy theoretically for new heavy resonance particles to interact with the SM through the Higgs as a portal, resulting in resonance di-Higgs production. With the increased center of mass collision energy from 8 TeV to 13 TeV, the production cross section through gluon-gluon fusion for heavy particles above TeV grows in LHC Run 2, as shown in Figure 1.6. Therefore, it is particularly important to focus on resonant searches above TeV region.

#### 1.4 DI-HIGGS DECAY AND LHC PREVIOUS SEARCH RESULTS

Di-Higgs decay is a combination of single Higgs decays. The coupling terms to fermions and bosons are shown in Eq 1.1. The branching ratio of the di-Higgs final state is shown in Figure 1.7.

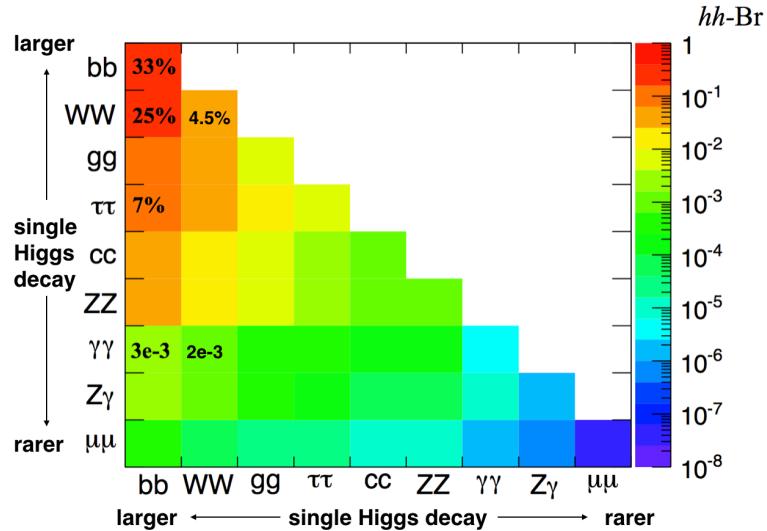
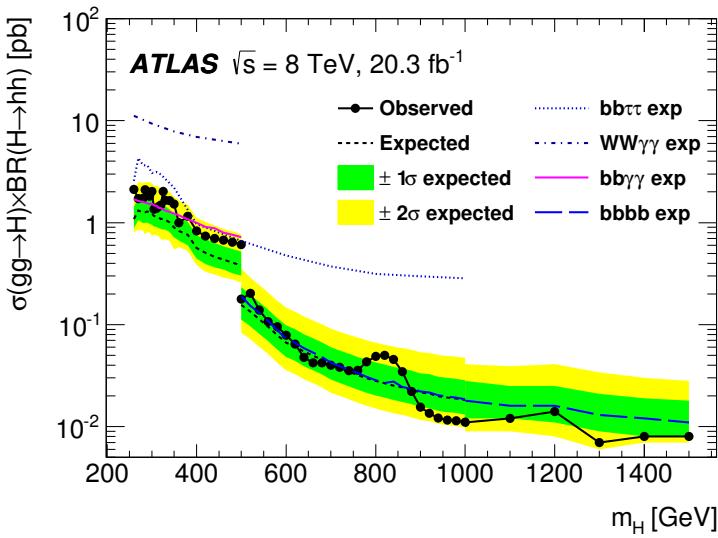


Figure 1.7: Summary of di-Higgs final states and their ratios. Top left,  $b\bar{b}b\bar{b}$ , has the largest branching ratio.

Previous searches for Higgs boson pair production have all yielded null results. Using 8 TeV data, ATLAS has examined the  $b\bar{b}b\bar{b}$ <sup>15</sup>,  $b\bar{b}\gamma\gamma$ <sup>16</sup>,  $b\bar{b}\tau^+\tau^-$  and  $W^+W^-\gamma\gamma$  channels, all of which were combined<sup>26</sup>. The resonant search combination result is shown in Figure 1.8. The best non-resonant  $\sigma(pp \rightarrow hh)$  cross section limit in Run 1 is the ATLAS combination, at 0.69 pb. This corresponds to



**Figure 1.8:** The observed and expected 95% CL upper limits of  $\sigma(gg \rightarrow H) \times BR(H \rightarrow hh)$  at  $\sqrt{s} = 8$  TeV as functions of the heavy Higgs boson mass  $m_H$ , combining resonant searches in Higgs boson pair to  $b\bar{b}\tau^+\tau^-$ ,  $W^+W^-\gamma\gamma$ ,  $bb\gamma\gamma$ , and  $bb\bar{b}\bar{b}$  final states. The expected limits from individual searches are also shown. The green and yellow bands represent  $\pm 1\sigma$  and  $\pm 2\sigma$  uncertainty ranges of the expected combined limits. The improvement above  $m_H = 500$  GeV reflects the sensitivity of the  $bb\bar{b}\bar{b}$  analysis. The results beyond 1 TeV are only from the  $bb\bar{b}\bar{b}$  final state alone.

$\frac{\lambda}{\lambda_{SM}} < 70$ . Different di-Higgs search challenges and perspectives are summarized below:

- $b\bar{b}b\bar{b}$ : Trigger limits the low mass resonance searches, but for high mass resonances above 500 GeV, the branching ratio of this channel provides a decisive advantage. Great for non-resonant searches.
- $b\bar{b}W^+W^-$ : Despite the second largest branching ratio, large background from  $t\bar{t}$  limits this search sensitivity.
- $b\bar{b}\gamma\gamma$ : Benefit from a good double photon trigger efficiency, a good photon reconstruction efficiency and a low SM background. Most sensitive at low mass  $m_X \leq 350$  GeV. At higher masses, the smaller branching ratio and the merging of photons hurt the search sensitivity. Great for non-resonant searches.
- $b\bar{b}\tau^+\tau^-$ : An intermediate choice between  $b\bar{b}b\bar{b}$  and  $b\bar{b}\gamma\gamma$  for resonance searches. Yet this channel contributes to the non-resonant result significantly.
- $W^+W^-\gamma\gamma$ : Suffers from much lower branching ratio and lower reconstruction efficiency of the  $W^+W^-$  compared to  $b\bar{b}$ .

- $W^+W^-\tau\tau, W^+W^-W^+W^-, b\bar{b}ZZ$ : There are no search results on these channels yet. But because of the relatively large branching ratio, it is likely that they would be explored in the future.

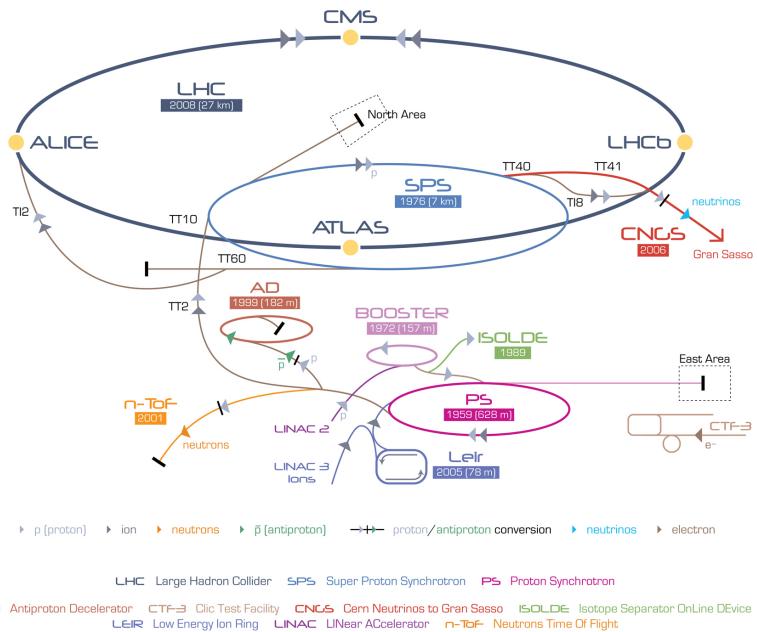
In summary, di-Higgs has a small production rate in the SM, but could be significantly enhanced in BSM scenarios. In particular, a heavy resonance spin-0 or spin 2 particle could decay into Higgs boson pair directly. The search sensitivity for massive resonances increases as the center of mass energy of the collision increases. For resonance signals above 1 TeV decaying into di-Higgs,  $b\bar{b}b\bar{b}$  channel has the best discovery potential in Run 2. Therefore, searching for TeV scale resonance production of di-Higgs  $\rightarrow b\bar{b}b\bar{b}$  is the goal of thesis.

*Pain teaches lessons no scholar can.*

# 2

## LHC and ATLAS

The Large Hadron Collider (LHC) is a proton-proton ( $p\bar{p}$ ) collider at the European Organization for Nuclear Research (CERN) laboratory in Geneva, Switzerland<sup>27</sup>. ATLAS (A Toroidal LHC ApparatuS), CMS (the Compact Muon Solenoid), ALICE (A Large Ion Collider Experiment), and LHC $b$  (Large Hadron Collider beauty experiment)<sup>28,29,30,31</sup> are the four main experiments. They are located at the Interaction points(IPs) of the accelerator. Figure 2.1 shows a schematic of the LHC ring and its experiments.



**Figure 2.1:** A schematic view of the LHC ring<sup>3</sup>. LINAC2, Booster, PS, SPS, and LHC accelerate the protons in order. Four main experiments are located at interaction points along the ring. ATLAS and CMS are general purpose experiments, while ALICE focuses on heavy ion collisions and LHC $b$  is dedicated to  $B$  physics.

## 2.1 THE LARGE HADRON COLLIDER

Protons accelerated in the LHC are from a red bottle of hydrogen gas. The whole acceleration takes around 25 minutes in multiple steps:

- An electric field strips the electrons from the hydrogen to create protons;
- A linear particle accelerator, Linac 2, accelerates the protons to 50 MeV;
- The Proton Synchrotron Booster (PSB) accelerates the protons to 1.4 GeV;
- The Proton Synchrotron (PS) accelerates the protons to 25 GeV;
- The Super Proton Synchrotron (SPS) accelerates the protons to 450 GeV;
- The 16.7 kilometers LHC accelerates the protons in a series of Radio Frequency cavities to the final TeV energies. The LHC uses 1232 Niobium Titanium magnetic dipole for steering

the protons. The magnets are cooled by superfluid helium to 1.9 Kelvin, and can generate 8.33 Tesla magnetic field.

In proton-proton collisions, the rate of a certain physics process  $R_{\text{phy}} = L\sigma$ , where  $L$  ( $\text{m}^{-2}\text{s}^{-1}$ ) is the instantaneous luminosity, and  $\sigma(\text{m}^2)$  the cross section of physics process (like di-Higgs'  $\sigma$ , 1.2). For a Gaussian beam profile, the instantaneous luminosity is defined in Eq2.1<sup>3</sup>:

$$L = \frac{n_b N_b^2 f_{\text{rev}} \gamma_r}{4\pi \epsilon_n \beta^*} F \quad (2.1)$$

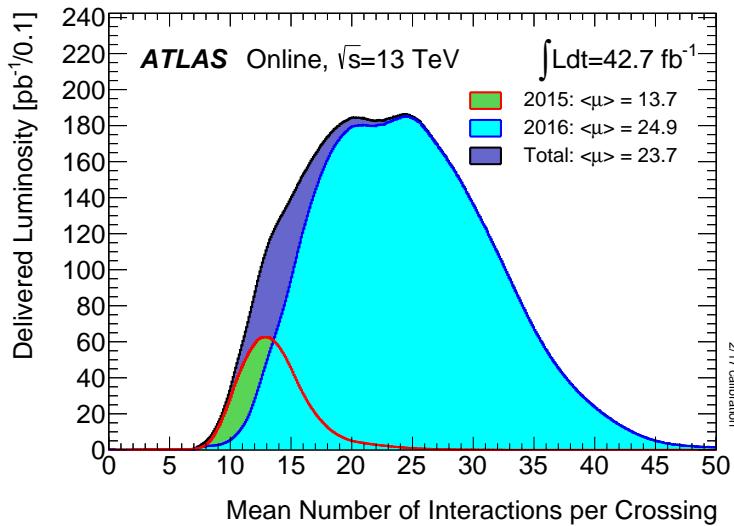
In the above Eq2.1:

- $n_b$  is the number of bunches per beam;  $n_b$  cannot be too large due to potential beam loss damages on the accelerator and detector;
- $N_b$  is the number of protons per bunch;
- $f_{\text{rev}}$  is the proton revolution frequency;
- $\gamma_r$  is the relativistic Lorentz factor for the protons;
- $\epsilon_n$  is the average beam spread length in the transverse plane;
- $\beta^*$  is the beam spread in the longitudinal direction; affected by focusing magnets;
- $F$  is a reduction factor for the angle beams are colliding; smaller crossing angles could cause larger spread in the longitudinal direction.

The instantaneous luminosity can also be written as the ratio of the rate of inelastic collisions to the inelastic cross section  $\sigma_{\text{inel}}$ <sup>32</sup>:

$$L = \frac{R_{\text{inel}}}{\sigma_{\text{inel}}} = \frac{\mu n_b f_{\text{rev}}}{\sigma_{\text{inel}}} \quad (2.2)$$

where,  $\mu$  is the number of interactions per bunch crossing. At each bunch crossing, multiple proton-proton collide, and the collisions without the highest center of mass energy are called “pileup” interactions. The target peak instantaneous luminosity for both the ATLAS and CMS experiments is  $L = 10^{34} \text{ cm}^{-2}\text{s}^{-1}$ <sup>27</sup>, which is already exceeded in 2016. This is partly due to the rising number of “pileup” interactions, shown in Figure 2.2. The main parameters of the LHC beam and performance are shown in Table 2.1.



**Figure 2.2:** The luminosity-weighted distribution of the mean number of interactions per crossing for the 2015 and 2016  $p\bar{p}$  collision data at  $13 \text{ TeV}$  centre-of-mass energy.<sup>4</sup>

## 2.2 A TOROIDAL LHC APPARATUS

The ATLAS experiment<sup>35</sup> at the LHC is a general-purpose particle detector with a near  $4\pi$  coverage in solid angle and a forward-backward symmetric cylindrical geometry. The ATLAS detector (Figure 2.3) consists of an inner tracking detector (ID) surrounded by a 2.3 m diameter thin superconducting solenoid providing a 2 T axial magnetic field, electromagnetic (EM) and hadronic calorimeters, and a muon spectrometer (MS). Three extra air-core toroid magnets generate the mag-

Parameter [unit]	Nominal design value	2015 Operating value	2016 Operating value
Beam Energy [TeV]	7	6.5	6.5
Peak L [ $10^{34} \text{ cm}^2 \text{ s}^{-1}$ ]	1	0.5	1.25
Bunch spacing [ns]	25	25	25
$f_{\text{rev}}$ [kHz]	11245	11245	11245
$n_b$ [ $10^{11}$ p/bunch]	1.15	1.15	1.12
$N_b$ [bunch]	2808	1825	2220
$\epsilon_n$ [mm mrad]	3.5	3.5	2
$\beta^*$ [cm]	55	20-40	40
$F$	0.84	0.84	0.59
$\langle \mu \rangle$	19	13	41

Table 2.1: LHC nominal<sup>27</sup> and operational parameters in 2015<sup>33</sup> and 2016<sup>34</sup>.

netic field in the MS.

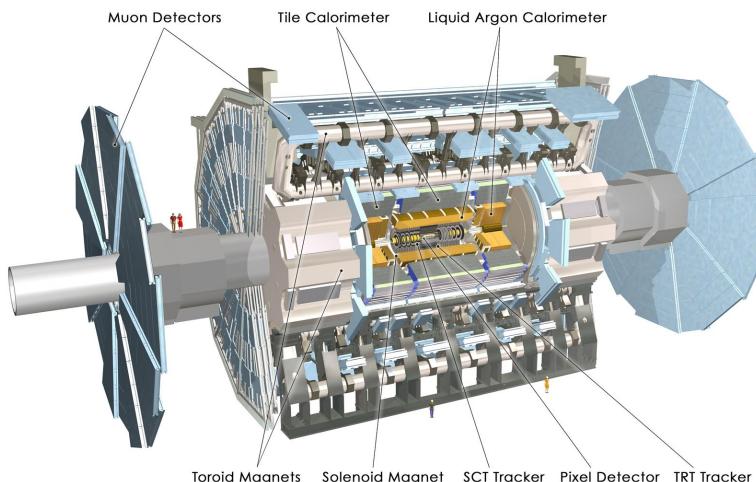


Figure 2.3: A detailed computer-generated image of the ATLAS detector and it's systems.

### 2.2.1 COORDINATE SYSTEM

ATLAS uses a right-handed coordinate system with its origin at the nominal IP in the center of the detector and the  $z$ -axis along the beam pipe. The  $x$ -axis points from the IP to the center of the

LHC ring, the  $y$ -axis points towards the sky, and the  $z$ -axis points straight (like bridges) towards the Geneva airport (A side), back from the Charlie pub in France (C side). Cylindrical coordinates  $(r, \varphi)$  are used in the transverse plane,  $\varphi$  being the azimuthal angle around the  $z$ -axis. The pseudorapidity is defined in terms of the polar angle  $\vartheta$  as  $\eta = -\ln(\tan(\vartheta/2))$ . It is the massless approximation of rapidity, the angle parameterizing special relativity's boosts. Most hadron productions are roughly constant in  $\eta$ , and for two massless particles traveling in different directions, their difference in  $\Delta\eta$  is invariant. Therefore, angular distance is measured in units of  $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2}$ .

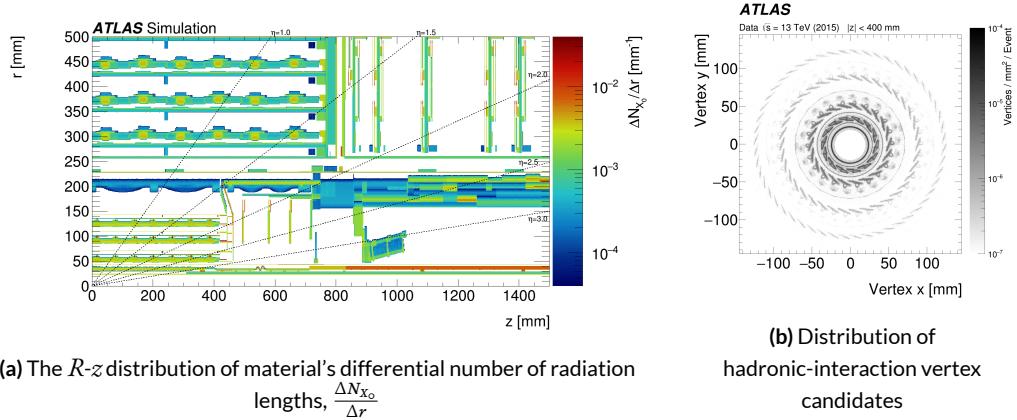
The region with  $|\eta| < 1.5$  is called “central”. It consists of the “barrel” elements, surrounding the beam line cylindrically. For  $|\eta| > 1.5$ , the region is called “endcap”, and the detector elements are arranged as disks perpendicular to the beam line. At high  $|\eta| > 2.5$ , the region is referred to as “forward”.

### 2.2.2 INNER DETECTOR

The ID covers the pseudorapidity range  $|\eta| < 2.5$ . It consists of three parts: silicon pixel (PIXEL), silicon microstrip (SCT), and straw-tube transition-radiation tracking (TRT) detectors. An additional pixel detector layer (IBL)<sup>36</sup>, inserted at a mean radius of 3.3 cm, is used in the Run-2 data-taking and improves the identification of  $b$ -jets<sup>37</sup>. A 10 GeV charged particle in the barrel region expect 1 + 3 IBL-Pixel hits, 8 SCT hits and 36 TRT hits.

ID is designed to provide charged particle momentum measurement with  $\sigma_{p_T}/p_T \sim 0.05\% p_T \oplus 1\%$  and vertex reconstruction. Because of this, each detector proves measurement accuracies of the order 10  $\mu\text{m}$  in  $R\text{-}\varphi$  and 100  $\mu\text{m}$  in  $z$ . Figure 2.4a<sup>38</sup> shows the  $R\text{-}z$  distribution of the material for a quadrant of the barrel region PIXEL and SCT. The intensity of a particle beam decreases exponentially in radiation length.  $I(x) = I_0 e^{-x/X_0}$ , where  $I$  is the intensity,  $x$  is the distance traveled, and  $X_0$

is the radiation length. Figure 2.4b shows the distribution of hadronic-interaction vertex candidates in  $|\eta| < 2.4$  and  $|z| < 400$  mm for 13 TeV data.



**Figure 2.4:** Geometry of IBL, PIXEL and SCT detectors in Run2.

### 2.2.3 CALORIMETER

Lead/liquid-argon (LAr) finely segmented sampling calorimeters provide EM energy measurements. A steel/scintillator-tile hadronic calorimeter covers the central pseudorapidity range ( $|\eta| < 1.7$ ). The endcap and forward regions are instrumented with copper/tungsten and LAr calorimeters for both the EM and hadronic energy measurements up to  $|\eta| = 4.9$ . The calorimeters also provide basic EM/Hadronic trigger information, with fast analogue summing in coarse granularity.

EM calorimeter (ECal) is designed to have  $> 22$  radiation lengths in the barrel and  $> 24$  in the endcap. It provides EM measurement with  $\sigma_E/E = 10\%/\sqrt{E} \oplus 0.7\%$ . The hadronic calorimeter (HCal) has approximately 9.7 interaction length in the barrel and 10 in the endcap. HCal provides hadronic measurement with  $\sigma_E/E = 50\%/\sqrt{E} \oplus 3\%$  in the Barrel and Endcap regions, and  $\sigma_E/E = 100\%/\sqrt{E} \oplus 10\%$  in the forward region.

#### 2.2.4 MUON SPECTROMETER

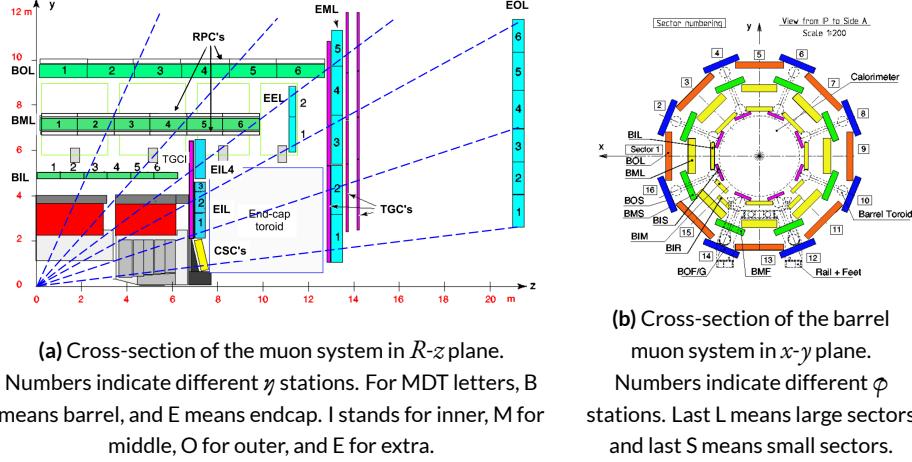


Figure 2.5: The overall layout of the ATLAS MuonSpectrometer.

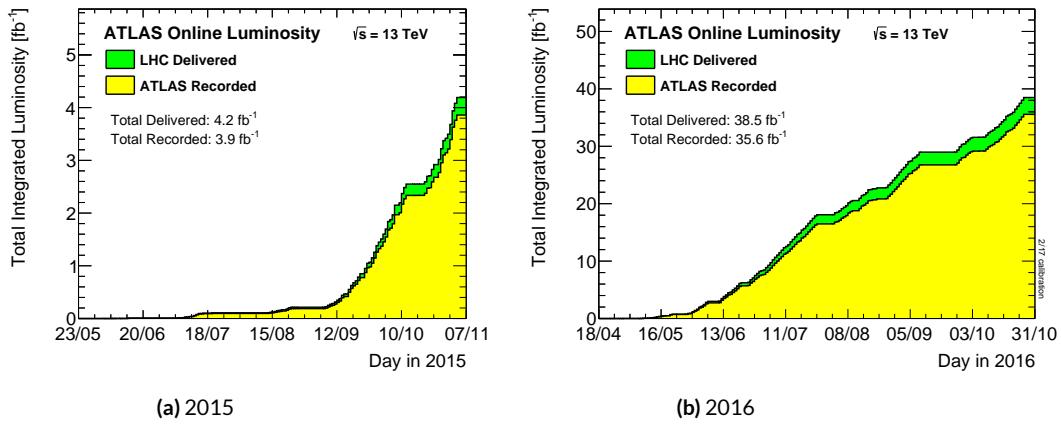
The muon spectrometer (Figure 2.5) surrounds the calorimeters and includes three large superconducting air-core toroids. The field integral of the toroids ranges between 2 and 6 T/m for most of the detector. Because of this bending power, the MS measures Muon momentum stand-alone, with  $\sigma_{p_T}/p_T \sim 10\%$  at  $p_T = 1\text{ TeV}$ . Muon Drift Tubes (MDT) and Cathode Strip Chambers (CSC) provide precision tracking. Each MDT has  $80\text{ }\mu\text{m}$  spacial resolution, with an alignment precision of  $30\text{ }\mu\text{m}$ . Resistive Plate Chambers (RPC) in the barrel and Thin Gap Chambers (TGC) provide triggering, with  $1.5\text{-}5\text{ ns}$  timing resolution. The muon spectrometer defines the overall dimensions of the ATLAS detector.

#### 2.2.5 TRIGGER AND DATA ACQUISITION

A dedicated trigger system is used to select events<sup>39</sup>. The first-level trigger ( $L_1$ ) is implemented in hardware and uses the calorimeter and muon detectors to seed regions of interest (RoI) and reduce

the accepted event rate to 100 kHz. This is followed by a software-based high-level trigger (HLT) that reduces the accepted event rate to 1 kHz on average. To avoid too high accept rates for certain triggers, the triggers are often prescaled, which means the accepted events get rejected at the prescale. For example, a prescale of two means only every second event passing all trigger conditions gets accepted.

Over 2015 and 2016, both the LHC and the ATLAS performed outstandingly <sup>4</sup>. The total data recording efficiency for ATLAS is around 92%, shown in Figure 2.6.



**Figure 2.6:** Cumulative luminosity vs. time delivered to (green) and recorded by ATLAS (yellow) during stable beams for  $p\bar{p}$  collisions at 13 TeV centre-of-mass energy.

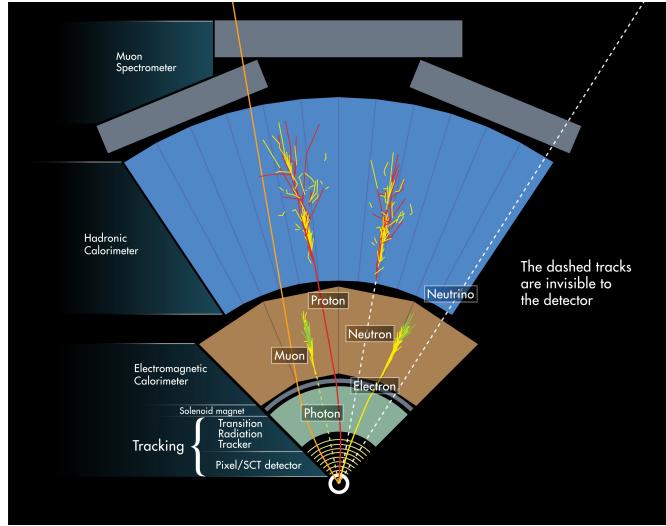
*“A picture is worth a thousand swords.”*

Tony

# 3

## Reconstruction and Objects

Reconstruction is the construction of particles from raw detector readouts. In each  $p\bar{p}$  collision recorded by ATLAS, charged particles bend in the magnetic field and leave tracks in the ID, electrons and photons deposit their energies in ECAL, hadrons are absorbed in HCAL, muons leave extra tracks in the MS, and neutrinos are inferred by the conservation of momentum in the transverse plane. Figure 3.1 gives an overview of the different sub-detectors that each type of particle will interact with in ATLAS. Quark reconstruction and identification is particularly important for this thesis, with di-Higgs decaying to  $b\bar{b}b\bar{b}$ .



**Figure 3.1:** Illustration of particle interactions in ATLAS.

### 3.1 ID TRACKS AND VERTICES

ID tracks originate from clusters based on PIXEL and SCT energy deposits. Three clusters form a seed, and the seeds are combined to build track candidates using a Kalman filter. After ambiguity solving, an artificial neural network is used to identify merged clusters. Finally, ID tracks are built from CPU intensive high resolution fits. The tracks are required to have  $p_T > 0.4 \text{ GeV}$  and  $|\eta| < 2.5$ .

Each  $p\bar{p}$  collision generates multiple vertices, and the vertices are reconstructed from the available ID tracks. The primary vertex (PV), or the hard-scatter vertex, is selected as the one with the largest  $\sum p_T^2$ , where the sum is over all tracks with transverse momentum  $p_T > 0.4 \text{ GeV}$  that are associated with the vertex. The ID tracks are usually required to have at least 1 PIXEL hit and 6 SCT hits, and to be tightly matched to the primary vertex.

The performance of track reconstruction is highly dependent on the momentum of the particle. With higher momentum, the decay tracks have smaller separations in the inner detector, hindering the resolving cluster process, and thus degrading the track identification efficiency. For a 1 TeV  $b$ -hadron, the reconstruction track efficiency is 83%, compared to 95% for a 200 GeV  $b$ -hadron <sup>40</sup>.

## 3.2 JETS

When a quark or gluon is produced in  $pp$  collisions, it produces a spray of hadrons, which is known as a jet. Jets are built from topological clusters of energy deposits in calorimeter cells <sup>41</sup>, using a four-momentum reconstruction scheme with massless clusters as input. The directions of jets are corrected to point back to the primary vertex. Typically, jets are reconstructed using the anti- $k_t$  algorithm with different values of the radius parameter  $R$ .  $R$  appears in the denominator of the clustering distance metric. It determines the radial size of the jet in  $\eta$ - $\phi$  plane.

### 3.2.1 SMALL- $R$ JETS

The jets with  $R = 0.4$  (“small- $R$  jets”) are reconstructed from clusters calibrated at the electromagnetic (EM) scale. The jets are corrected for additional energy deposited from pile-up interactions using an area-based correction <sup>42</sup>. They are then calibrated using  $p_T$ - and  $\eta$ -dependent calibration factors derived from simulation, before global sequential calibration <sup>43</sup> is applied, which reduces differences in calorimeter responses to gluon or quark-initiated jets. The final calibration is based on in situ measurements in collision data <sup>44</sup>.

“Small- $R$  jets” are required to be consistent with the primary vertex, in order to avoid contamination from pileup interactions. The jet vertex fraction (JVF) is a useful variable for this purpose. It is the ratio of tracks associated with a primary vertex to the total number of tracks inside a jet. Jets from the PV should have most tracks consistent with the PV and therefore have a large JVF value.

### 3.2.2 LARGE- $R$ JETS

The jets with  $R = 1.0$  (“large- $R$  jets”) are built from locally calibrated<sup>43</sup> topological clusters.

They are trimmed<sup>45</sup> to minimize the impact of energy deposits from pile-up interactions. Trimming proceeds by reclustering the jet with the  $k_t$  algorithm<sup>46</sup> into  $R = 0.2$  sub-jets and then removing those sub-jets with  $\hat{p}_T^{\text{subjet}}/\hat{p}_T^{\text{jet}} < 0.05$ , where  $\hat{p}_T^{\text{subjet}}$  is the transverse momentum of the sub-jet and  $\hat{p}_T^{\text{jet}}$  that of the original jet. The energy and mass scales of the trimmed jets are then calibrated using  $p_T$ - and  $\eta$ -dependent calibration factors derived from simulation<sup>47</sup>.

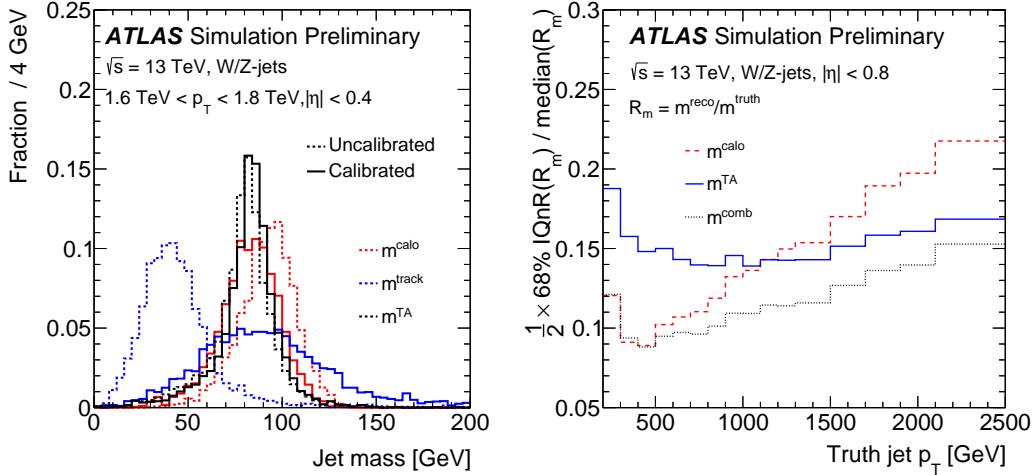
The calorimeter-based jet mass  $m^{\text{calo}}$  for a large-radius calorimeter jet  $J$  is computed from the calorimeter cell cluster constituents  $i$  with energy  $E_i$  and momentum  $\vec{p}_i$ :

$$m^{\text{calo}} = \sqrt{\left(\sum_{i \in J} E_i\right)^2 - \left(\sum_{i \in J} \vec{p}_i\right)^2}. \quad (3.1)$$

For a boosted massive particle, the angular spread in the decay products scales as  $\frac{1}{p_T}$ . For highly boosted cases, the spread is comparable with the  $\eta \times \varphi \sim 0.1 \times 0.1$  calorimeter granularity. Tracking information is used to maintain performance beyond the calorimeter granularity limit. The track-assisted jet mass,  $m^{\text{TA}}$ , is defined as:

$$m^{\text{TA}} = \frac{\hat{p}_T^{\text{calo}}}{\hat{p}_T^{\text{track}}} \cdot m^{\text{track}}. \quad (3.2)$$

where  $\hat{p}_T^{\text{calo}}$  is the transverse momentum of the large- $R$  calorimeter jet,  $\hat{p}_T^{\text{track}}$  is the transverse momentum of the four-vector sum of tracks associated to the large- $R$  calorimeter jet, and  $m^{\text{track}}$  is the invariant mass of this four-vector sum. This ratio corrects for charged-to-neutral hadron fluctuations, and therefore improves the resolution with respect to track-only jet mass.



(a) Uncalibrated (dashed line) and calibrated (solid line) jet mass distribution.

(b) The fractional jet mass resolution vs. the truth jet mass transverse momentum.

**Figure 3.2:** Uncalibrated (dashed line) and calibrated (solid line) reconstructed jet mass distribution 3.2a, and the jet mass resolution vs jet  $p_T$  3.2b for calorimeter-based jet mass,  $m_{calo}$  (red), track-assisted jet mass  $m_{TA}$  (black) and the invariant mass of four-vector sum of tracks associated to the large-radius calorimeter jet  $m_{track}$  (blue) for W/Z-jets<sup>5</sup>.

The above two mass definitions are only weakly correlated with each other, so they can be linearly combined to the combined mass,  $m^{comb}$ , by weighting the components with  $w$ :

$$m^{comb} = w \cdot m^{calo} + (1 - w) \cdot m^{TA}. \quad (3.3)$$

where  $w$  is determined for each large- $R$  jet from the resolution functions of the calibrated track and calo mass terms. This results in a smaller mass resolution and better estimate of the median mass value than obtained using only calorimeter energy clusters, as shown in Figure 3.2.

### 3.2.3 TRACK JETS

Track jets are essentially clustered charged hadron tracks. They are reconstructed from ID tracks using the anti- $k_t$  algorithm with a fixed  $R = 0.2$ . Once the track jet axis is determined, an extra step of track association is performed to select tracks with looser impact parameter requirements,

in order to collect the tracks needed for effectively running the  $b$ -tagging algorithms. Only track jets with at least two tracks are kept. Track jets are also required to have  $p_T > 10 \text{ GeV}$  and  $|\eta| < 2.5$ , in order to suppress track jets from light flavors.

Track jets are associated to large- $R$  jets using Ghost-Association. “Ghosts” are track jet 4-vectors, with each track jet’s  $p_T$  set to an infinitesimal amount, essentially keeping only the direction of the 4-vector. This ensures that Large- $R$  jets reconstruction is not altered by the ghosts when the calorimeter clusters plus ghosts are reclustered. Reclustering is then performed using the anti- $k_t$  algorithm with  $R = 1.0$ . The calorimeter jets after reclustering are identical to the parents of the trimmed jets used in this analysis, with the addition of the associated track jets retained as constituents. In addition, the track jets corresponding to the ghosts that survive the trimming procedure (and thus are clustered into one of the surviving sub-jets) are the track jets ghost-associated to the trimmed jet. The small radius parameter of the track-jets enables two nearby  $b$ -hadrons to be identified when their  $\Delta R$  separation is small, which is beneficial when reconstructing high- $p_T$  Higgs boson candidates.

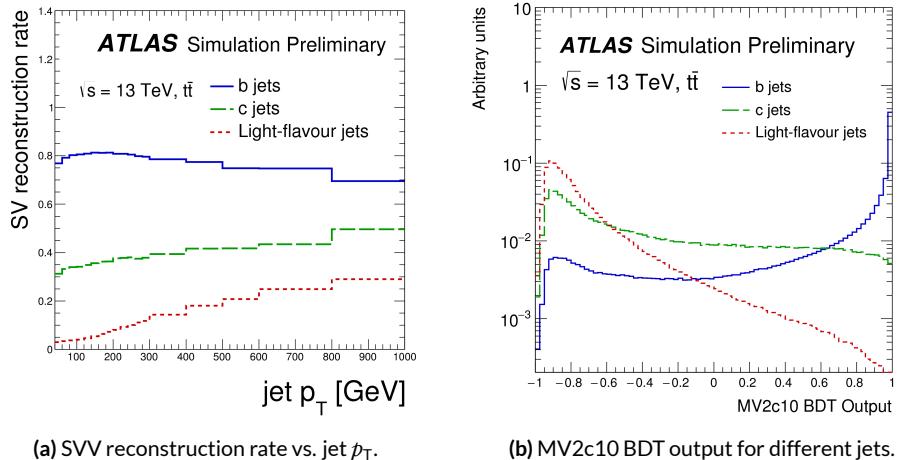
### 3.3 FLAVOR TAGGING

Some jets are formed from quarks with different flavors. Identifying the flavor of a jet is called flavor tagging. Jets originating from a  $b$ -quark is referred as a  $b$ -jet, from a  $c$ -quark is referred as a  $c$ -jet, and from a other quarks other than the  $t$ -quark is referred as a light jet.  $b$ -tagging is particularly useful for this analysis, since there are four  $b$ s in the final state.

$B$ -hadrons have a lifetime on the order of  $10^{-12}$  seconds, which makes a flight distance of  $0.5 \text{ mm}^{\textcolor{red}{1}}$ . This results in a displaced decay vertex that can be identified in vertex reconstruction. This allows some  $b$ -jets to be distinguished from other flavors of jets theoretically.

ATLAS uses three different basic  $b$ -tagging algorithms, which provide complementary information:

- Impact parameter (IP) based algorithm: it uses the transverse and longitudinal impact parameters  $d_o$  and  $z_o$  of the tracks inside a jet to determine their consistency with the primary vertex. The algorithm uses two or three dimensional templates for light,  $c$ , and  $b$  jets and evaluates the likelihood of the jet coming from each of these types.
- Inclusive secondary vertex (SV) reconstruction algorithm: it uses tracks inside the jet to fit for vertices that are displaced from the primary vertex. The algorithm provides information on the invariant mass of tracks pointing to same vertex, the number of two track vertices, and the angular separation between the jet and the PV  $\rightarrow$  SV direction.
- Decay chain multi-vertex reconstruction algorithm, or JetFitter (JF): it reconstructs the full flight path of the  $b$  by looking for multiple displaced vertices along the same direction. A Kalman filter is used to find common line for the  $b$  and  $c$  vertices, and hence the algorithm exploits the topology of the weak  $b/c$ -hadron decay chain.



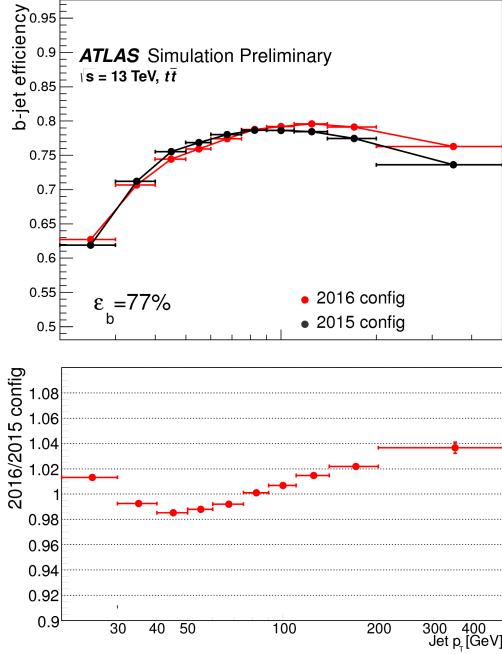
**Figure 3.3:** Secondary vertex reconstruction rate and MV2c10 output for  $b$ -jets (solid blue),  $c$ -jets (dashed green) and light-jets (dotted red) evaluated with simulated  $t\bar{t}$  events.

Jets containing  $b$ -hadrons are identified using a score value computed from a boosted decision tree(BDT) algorithm  $MV2c10$ <sup>6,48</sup>, which makes use of observables provided by three algorithms above. The  $MV2c10$  algorithm is trained on a sample with charm composition of 7%. It is applied to a set of charged-particle tracks that satisfy quality and impact parameter criteria and are matched to each jet. Hence either a small- $R$  jet or a track jet can be  $b$ -tagged.

The  $b$ -tagging working point (wp) is a fixed cut on the  $MV2c10$  value that lead to an efficiency of 70% for  $b$ -jets with  $p_T > 20$  GeV when evaluated in a sample of simulated  $t\bar{t}$  events. This working point corresponds to a rejection rate of jets originating from  $u, d$  or  $s$ -quarks or gluons of 380 for the jets with  $R = 0.4$  and 120 for the track jets. The rejection of jets from  $c$ -quarks is 12 for the  $R = 0.4$  jets and 7.1 for the track jets.

In this thesis, the track-jets have a wider  $p_T$  range, between 50 – 400 GeV, and the same working point leads to  $b$ -tagging efficiencies varying from 40% at low  $p_T$ , to 80% for  $p_T$  values of about 150 GeV, to 60% at high  $p_T$ . This can be seen in Figure 3.4. The increase of tracks from fragmentation in the high jet  $p_T$  region is the main reason for the performance degradation. As the jet  $p_T$  increases, the number of fake vertices is increasing, while the secondary vertex reconstruction efficiency for  $b$  and  $c$  jets decreases with jet  $p_T$ . This is shown in Figure 3.5. This non-trivial jet  $p_T$  dependence of  $b$ -tagging performance is one of the major challenge of this analysis.

Correction factors are applied to the simulated MC samples to compensate for differences between data and simulation in the  $b$ -tagging efficiency for  $b, c$  and light-jets. The correction for  $b$ -jets is derived from  $t\bar{t}$  events with final states containing two leptons, and the corrections are consistent with unity with uncertainties at the level of a few percent over most of the jet  $p_T$  range.

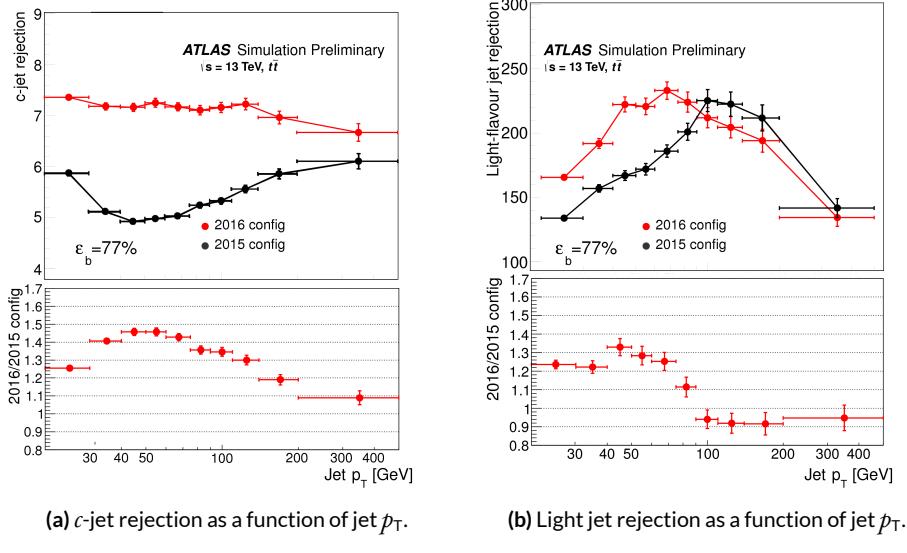


**Figure 3.4:**  $b$ -jet efficiency for the fixed cut working point with a  $b$ -jet efficiency of 77% as a function of the jet  $p_T$  for the comparison between the MV2c10  $b$ -tagging algorithm employed for the 2016 analyses (2016 config) and the previous version of the tagger, MV2c20 (2015 config), which has 15%  $c$ -fraction in the training

### 3.4 LEPTONS

Electron and photon identification is based on matching tracks to energy clusters in the ECAL and relying on the longitudinal and transverse shapes of the EM shower<sup>49</sup>. Well-reconstructed ID tracks matched to EM clusters are classified as electron candidates, while EM clusters without matching tracks are classified as unconverted photon candidates. Clusters matched to a reconstructed conversion vertex or to pairs of tracks consistent with the conversion hypothesis are classified as converted photon candidates. Electrons and photons are not used in this thesis.

Hadronic tau decays into a tau neutrino and one or three charged pions and up to two neutral pions<sup>50</sup>. Hence tau reconstruction is seeded by jets, and is matched to one or three associated tracks,



**Figure 3.5:** Light-flavour jet and  $c$ -jet rejection as a function of jet  $p_T$  for the previous (2015 config) MV2c20 and the current MV2c10 configuration (2016 config). A fixed cut at 77%  $b$ -jet efficiency operating point is used<sup>6</sup>.

with a total electric charge of  $\pm 1$ . A Boosted Decision Tree identification procedure, based on calorimetric shower shapes and tracking information is used to reject fakes from jets. Hadronic taus are not used in this thesis.

Neutrinos are inferred from the missing transverse momentum (MET), or  $E_T^{miss}$ . Neutrinos do not interact with the ATLAS. Their presence can only be deduced from the conservation of transverse momentum in each collision, as the incoming protons have no net momentum in the transverse plane. MET is calculated as the negative vectorial sum of the  $p_T$  of all fully reconstructed and calibrated physics objects. This procedure includes a soft term, which is calculated using the ID tracks that originate from the primary vertex but are not associated with reconstructed objects. MET is not used in this thesis.

Muons are identified by matching ID tracks with reconstructed MS tracks<sup>5</sup>. For this thesis, muons must have  $p_T > 4 \text{ GeV}$ ,  $|\eta| < 2.5$  and to satisfy “medium” muon identification criteria<sup>5</sup>.

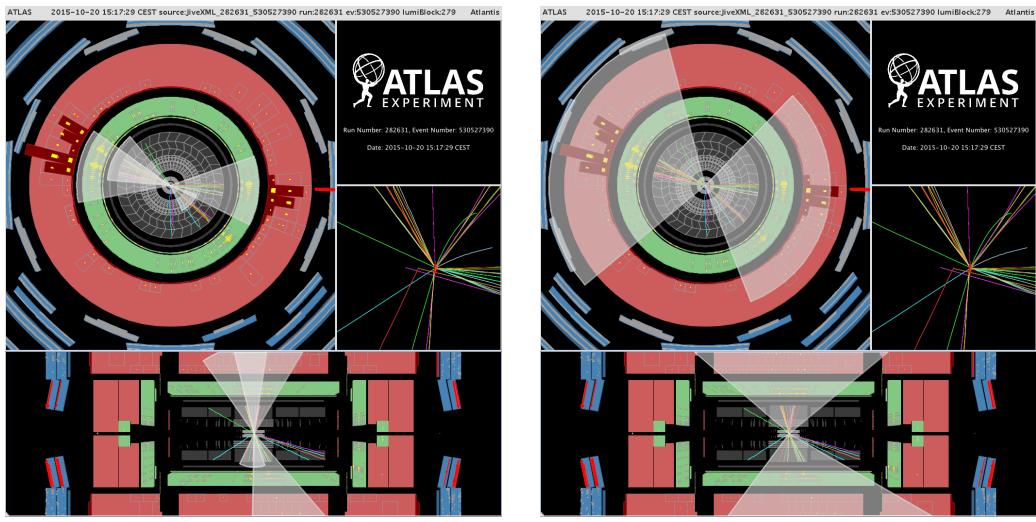
Muons are used in this thesis, because  $b$  hadrons decay to muons with  $\sim 20\%$  probability. This will be demonstrated in the later chapters.

### 3.5 RESOLVED AND BOOSTED

The thesis focuses on searching for TeVscale resonance decaying into di-Higgs and then to  $b\bar{b}b\bar{b}$ . It is important to fully reconstruct the out-coming  $b$  quarks. The angular separation between the  $b\bar{b}$  from one Higgs with momentum  $p_H$ ,  $\Delta R_{bb}$ , scales roughly as  $\frac{2m_H}{p_H}$ . This means for a 1.5 TeV resonance  $G_{KK}^*$ , produced roughly at rest, the two Higgs with each  $\sim 625$  GeV momentum, and the  $\Delta R_{bb}$  is around 0.4. Similarly, for a 3 TeV resonance, the 2.75 TeV resonance,  $\Delta R_{bb}$  is around 0.2.

The  $\Delta R_{bb}$  defines the reconstructed object choice in this analysis. Two different methods are used to reconstruct the Higgs bosons. The *resolved analysis* is used for di-Higgs systems in which the Higgs bosons have Lorentz boosts low enough that four  $R = 0.4$   $b$ -jets can be reconstructed. The *boosted analysis* is used for di-Higgs systems in which the Higgs bosons have higher Lorentz boosts, which prevents the Higgs boson decay products from being resolved in the detector as separate  $b$ -jets. Instead, each Higgs boson candidate consists of a single large-radius jet, and the presence of  $b$ -quarks is inferred using smaller-radius track jets built from charged-particle tracks.

Sometimes one event can be reconstructed in both the resolved method as four small- $R$  jets and the boosted method as two large- $R$  jets with track jets. Figure 3.6 shows an example event display of collision data recorded in 2015. To solve this ambiguity, the boosted selection vetos events passing the resolved selection, which will be introduced later.



**Figure 3.6:** Event display of the same event using 3.6a resolved and 3.6b boosted topologies. ID is in grey, ECAL is in green, HCAL is in red and MS is in blue. Jets are gray cones, and ID tracks are colored lines in the ID. The resolved reconstruction ends up with a  $m_{4J}$  of 873 GeV, and the boosted reconstruction gives  $m_{2J}$  of 852 GeV.

*Ugliness is in a way superior to beauty because it lasts.*

Serge Gainsbourg

# 4

## Data and Simulation

### 4.1 DATA

This analysis uses 2015 and 2016 LHC  $p\bar{p}$  collision datasets at  $\sqrt{s} = 13$  TeV recorded by the ATLAS experiment. Data were collected during stable beam conditions and when all relevant detector systems were functional. A Good Run List (GRL) is generated after gathering online and offline data quality reviews of the dataset after reconstruction. Typically, any  $> 10\%$  defect in any detector subsystem makes the corresponding Lumiblocks (LB) fail the GRL requirement. The integrated luminosity of the 2015 dataset passing the GRL is  $3.2 \text{ fb}^{-1}$ , and the 2016 dataset passing a different GRL is  $32.9 \text{ fb}^{-1}$ . These values are about 82% and 92% of the data ATLAS recorded, in Figure 2.6.

In the resolved analysis, a combination of  $b$ -jet triggers is used. Events are required to feature either one  $b$ -tagged jet with transverse momentum  $p_T > 225$  GeV, or two  $b$ -tagged jets, either both satisfying  $p_T > 35$  GeV or both satisfying  $p_T > 55$  GeV, with different requirements on the  $b$ -tagging. Some triggers require additional non- $b$ -tagged jets. Due to a change in the online  $b$ -tagging algorithm between 2015 and 2016, the two datasets are treated independently until they are combined in the final statistical analysis. After the selection described later, this combination of triggers is estimated to be 65% efficient for simulated signals with a Higgs boson pair invariant mass,  $m_{HH}$ , of 280 GeV, rising to 100% efficiency for resonance masses greater than 600 GeV.

During 2016 data-taking, a fraction of the data was affected by a bug. The movement of beam spot was not accounted for in the online vertex reconstruction. This reduced the efficiency of the algorithms used to identify  $b$ -jets. This reduces the integrated luminosity of the 2016 dataset for the resolved analysis to 24.3  $\text{fb}^{-1}$ .

In the boosted analysis, events were selected from the 2015 dataset using a trigger that required a single anti- $k_t$  jet with radius parameter  $R = 1.0$  and with  $p_T > 360$  GeV. In 2016, a similar trigger was used but with a higher threshold of  $p_T > 420$  GeV. The efficiency of these triggers is 100% for simulated signals passing the jet requirements as described later, so the 2015 and 2016 datasets were combined into one dataset.

The data is further skimmed into the Derived Analysis Object Data (DAOD). The ATLAS offline software 20.7.8.7 derivation cache is used, with version  $p$ -tags  $p2950$ . The boosted slimming keeps events with at least two large- $R$  jets with  $p_T > 200$  GeV. The final input data file has name format: `dataYR_13TeV.periodPR.physics_Main.PhysCont.DAOD_EXOT8.grpYR_v01_p2950`, where YR is 15 and PR is DEFGHJ for 2015 data, and YR is 16 and PR is ABCDEFGIKL for 2016 data.

## 4.2 MC

All Monte Carlo (MC) samples used in this analysis are produced with full simulation. For all simulated samples, charm-hadron and bottom-hadron decays were handled by `EVTGEN 1.2.0`<sup>3</sup>. To simulate the impact of multiple  $p\bar{p}$  interactions that occur within the same or nearby bunch crossings (pile-up), minimum-bias events generated with `PYTHIA 8`<sup>52</sup> using the A2 set of tuned parameters<sup>53</sup> were overlaid on the hard-scatter event. The detector response was simulated with `GEANT 4`<sup>54,55</sup> and the events were processed with the same reconstruction software as that used for the data. Simulated data samples from the ATLAS MC15c campaign are used, corresponding to  $p$ -tags p2952–p2949.

### 4.2.1 BACKGROUNDS

A very small fraction of the background arises from  $Z + \text{jets}$  events. The  $Z+\text{jets}$  sample was generated using `PYTHIA 8.186` with the NNPDF2.3 LO PDF set.

The  $t\bar{t}$  background is modeled using large all-hadronic and non-all-hadronic samples that have both been generated with `POWHEG-BOX v1`<sup>56</sup> using the CT10 PDF set. The parton shower, hadronization, and the underlying event were simulated using `PYTHIA 6.428` with the CTEQ6L1 PDF set and the corresponding Perugia 2012 set of tuned underlying-event parameters<sup>57</sup>. The  $t$ -quark mass in both samples is set to 172.5 GeV. Higher-order corrections to the  $t\bar{t}$  cross section were computed with `Top++ 2.0`<sup>58</sup>. These incorporate NNLO corrections in QCD, including resummation of NNLL soft gluon terms. The  $t\bar{t}$  MC samples are normalized to the NNLO+NLL predicted inclusive  $t\bar{t}$  cross-section of 1821.87 pb multiplied by the all-hadronic branching ratio of 0.457 and non-all-hadronic of 0.543.

In order to keep statistical fluctuations small across the dijet invariant mass spectrum, especially for large values of  $m_{tt}$ , additional  $t\bar{t}$  samples are generated in slices of  $t\bar{t}$  invariant mass. The cross-section of the  $t\bar{t}$  process is normalized to NNLO+NNLL in QCD, as calculated by Top++ 2.0. Overlap with the inclusive  $t\bar{t}$  samples is removed by a fixed cut on the truth value of  $m_{tt}$  at 1100 GeV.

A PYTHIA dijet sample is used to understand the physical processes contributing to the multi-jet background and characteristics of the event selection. This MC sample is generated without a heavy flavor filter, hence is limited by the total generated number of events, given the high background rejection factors of the analysis selection.

#### 4.3 SIGNAL

In all signal samples, the mass of the Higgs boson ( $m_H$ ) was set to 125 GeV. The signal MC contains truth information, like the two Higgs and the b quark four-momentum before detector interactions. This enables a  $\Delta R$  truth matching between the reconstructed objects and the Higgs.

SM non-resonant production of Higgs boson pairs via the gluon–gluon fusion process was simulated at NLO with MG5\_aMC@NLO, using form factors for the top-quark loop from HPAIR<sup>66,67</sup>. The simulated events were reweighted to reproduce the  $m_{hh}$  spectrum obtained<sup>68,69</sup>, which calculated the process at NLO in QCD while fully accounting for the top-quark mass. Interference effects between di-Higgs resonant production and SM non-resonant di-Higgs production are not included in the simulated samples.

Signal  $G_{KK}^* \rightarrow bb \rightarrow b\bar{b}b\bar{b}$  events were generated at leading order (LO) with MG5\_aMC@NLO 2.2.2<sup>59</sup> interfaced with PYTHIA 8.186 for parton-showering, hadronization and underlying-event simulation. The NNPDF2.3 LO parton distribution function (PDF) set<sup>60</sup> was used for both MG5\_aMC@NLO and PYTHIA. The A14 set of tuned underlying-event parameters was used. These signal samples

were generated with  $k/\bar{M}_{\text{Pl}} = 1$  or  $2$ . Relative to the resonance mass, widths of the graviton signals range from  $3\%$  (at low mass) to  $13\%$  (at the highest mass) for  $k/\bar{M}_{\text{Pl}} = 1$ , and  $6\%$  to  $25\%$  for  $k/\bar{M}_{\text{Pl}} = 2$ . The graviton samples were normalized using fixed cross sections<sup>61</sup>.

Signal 2HDM Scalar  $\rightarrow HH \rightarrow b\bar{b}b\bar{b}$  events were generated at LO in QCD with MG5\_aMC@NLO 2.2.3 interfaced with HERWIG++<sup>62</sup> for parton-showering, hadronization and simulation of the underlying event. CT10<sup>63</sup> PDF sets were used for MG5\_aMC@NLO and CTEQ6L1<sup>64</sup> for HERWIG++. The UE-EE-5-CTEQ6L1 set of tuned underlying-event parameters<sup>65</sup> was used. The scalar signals were generated with a width of  $1$  GeV, which represent generic narrow-width scalar signals. Because the width and branching ratios depend on 2HDM parameters, each mass point generated with this fixed width corresponds to a different point in the 2HDM parameter phase space.

Resonant signal samples for the scalar and  $k/\bar{M}_{\text{Pl}} = 1$  models were produced in  $10$  GeV steps between  $260$  and  $300$  GeV, in  $100$  GeV steps up to  $1600$  GeV, in  $200$  GeV steps up to  $2000$  GeV, and in  $250$  GeV steps up to  $3000$  GeV. Signal samples for the  $k/\bar{M}_{\text{Pl}} = 2$  model were produced with the same spacings but omitting the masses of  $270$  GeV,  $290$  GeV and  $2750$  GeV due to the larger generated width. Unless specified, the MC signal sample used as benchmark is  $k/\bar{M}_{\text{Pl}} = 1$   $G_{\text{KK}}^*$ , due to its width is medium among the three signal models.

*You gotta have a swine to show you where the truffles are.*

Edward Albee

# 5

## Event Selection

### 5.1 DATA CLEANING

The following data cleaning requirements are made:

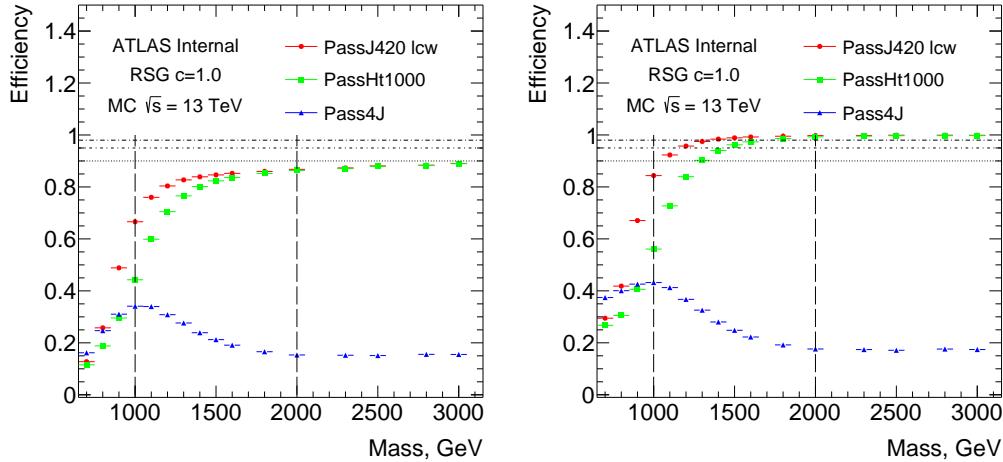
- Events with problems in TileCal/LAr are removed.
- Events that are affected by the recovery procedure for single event upsets in the SCT are removed.
- Events that fail the jet cleaning procedure are removed. This is designed to exclude jets caused by detector noise, non-collision backgrounds and cosmic rays.
- Incomplete events are removed.

The analysis also runs over the debug stream, which contains events recorded that couldn't be reconstructed online due to CPU time constraints. No event passing the full signal selection is found.

## 5.2 TRIGGER

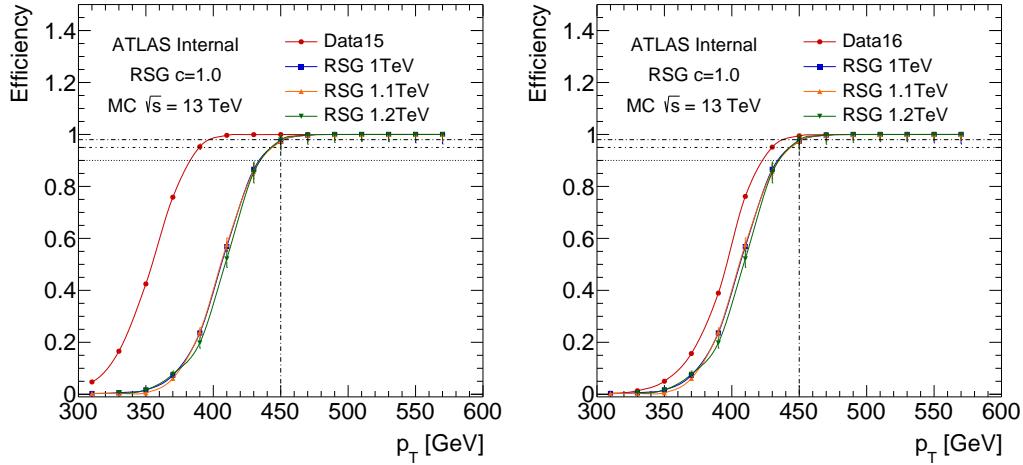
Events in data and MC are required to pass the lowest unprescaled large- $R$  jet trigger:

`HLT_j360_a10_lcw` in 2015 and `HLT_j420_a10_lcw` in 2016. The triggered jets are topo-cluster jets with local calibration weights and pile-up subtraction. They are seeded by the lowest unprescaled L1 jet trigger, `L1_J100`. LCW cluster trigger is chosen, because the other option, reclustered large- $R$  jet trigger, has slower turn-on in multi-jet events. Other options such as the lowest unprescaled HT trigger, `HLT_ht1000`, has a much slower turn-on compared to large- $R$  jet triggers. Another trigger option is `HLT_4j100`, but because of the boosted jets merging, the trigger efficiency decreases rapidly as the signal mass increases. The results are shown in Figure 5.1.



**Figure 5.1:** Different trigger efficiencies as a function of the signal resonance mass with respect to all events with no selection (left) and with respect to events passing the two large- $R$  jets  $p_T > 400 \text{ GeV}$  and leading/subleading jet  $p_T > 250 \text{ GeV}$  (right). For  $1.4 \text{ TeV}$  signal, the trigger efficiency is about 98%.

The selected large- $R$  triggers are found to have  $> 98\%$  efficiency for signals with mass above 1400 GeV. The trigger turn-on curve in 2015 and 2016 data, as a function of leading jet  $p_T$ , is shown in Figure 5.2.



**Figure 5.2:** Large- $R$  jet trigger efficiencies, defined as the fraction of events fired trigger with a given highest large- $R$  jet  $p_T$ , measured in 2015 Data (HLT\_j360\_a10\_lcw, left) and 2016 Data (HLT\_j420\_a10\_lcw, right) and MC.

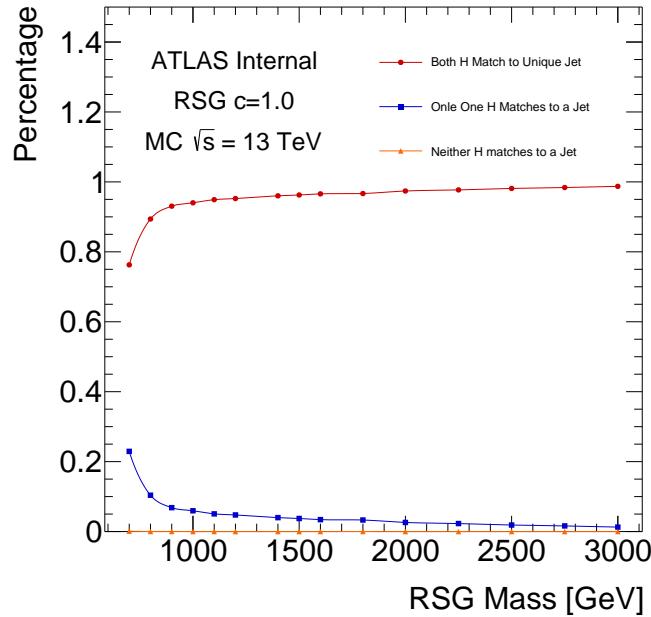
### 5.3 OBJECT SELECTION

The specific physics objects used in the boosted analysis are described in previous sections and reiterated in Table 5.1.

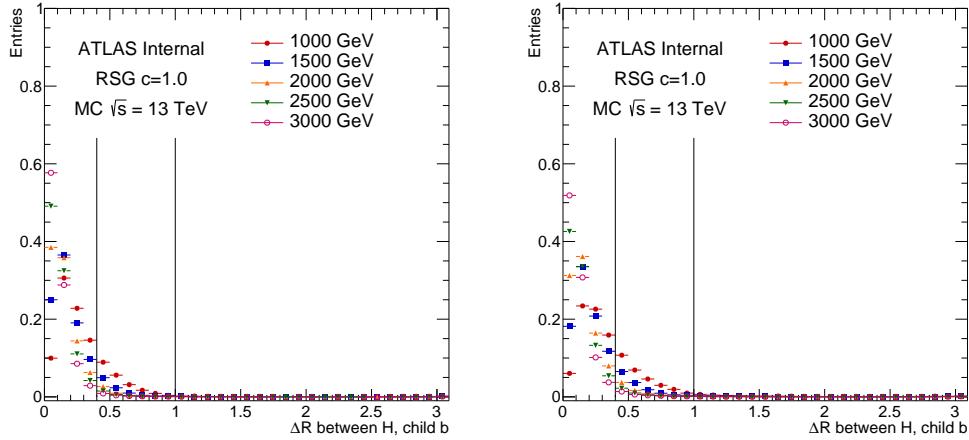
**Table 5.1:** Physics objects and their technical names in the boosted analysis.

object	technical name
large- $R$ calorimeter jets	AntiKt1oLCTopoTrimmedPtFrac5SmallR2oJets
small- $R$ track jets	AntiKt2PV0TrackJets
b-tagging	on track jets, MV2c10, 70% b-tagging wp

Each event must have at least two high momentum large- $R$  jets, each with at least one ghost associated track jets for  $b$ -tagging. They are sorted by  $p_T$ , and the highest  $p_T$  one is named as the leading large- $R$  jet, or the leading Higgs Candidate. The second highest  $p_T$  large- $R$  jet is the subleading large- $R$  jet, or the subleading Higgs Candidate. The large- $R$  jets are required to have  $p_T > 250$  GeV,  $|\eta| < 2$  to guarantee a good overlap with the tracking acceptance, and mass  $> 50$  GeV to avoid oversized derivations. The leading large- $R$  jet is also required to have  $p_T > 450$  GeV to be above the trigger turn on threshold. Only the leading and subleading large- $R$  jets are considered in the rest of this thesis. This selection is  $\sim 95\%$  efficient for 1.2 TeV signals, as shown in Figure 5.3.  $R = 1.0$  ensures that the two  $b$  quarks and their decay products are very likely to be contained within the large- $R$  jet, as shown in Figure 5.4.

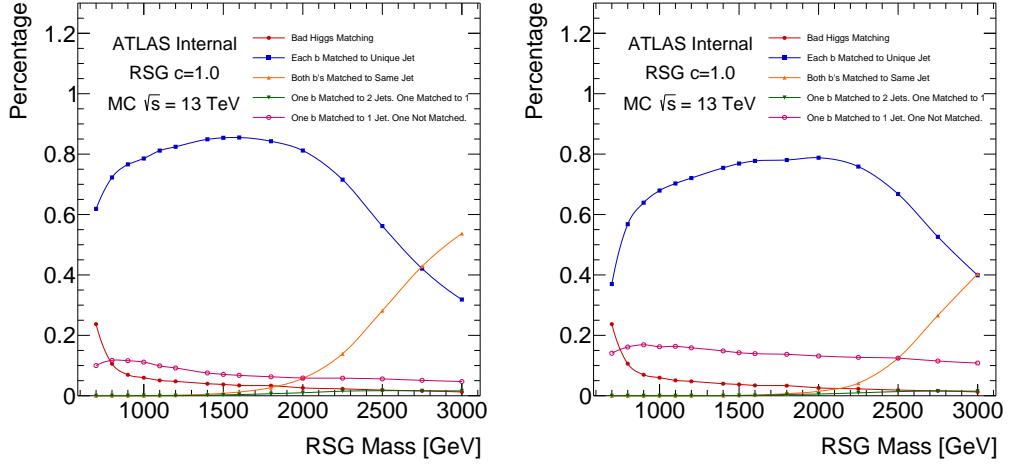


**Figure 5.3:** Percentages of truth Higgs to large- $R$  jet  $\Delta R < 1.0$  matching as a function of  $G_{KK}^*$  mass. Both Higgs almost never match to the same large- $R$  jet.



**Figure 5.4:** Normalized  $\Delta R$  between the truth Higgs (leading on left, subleading on right) and the truth children  $b$ -quarks for  $G_{KK}^*$  MCs. Lines are drawn at  $\Delta R = 0.4$  ( $R$  of small- $R$  jets) and  $\Delta R = 1.0$  ( $R$  of large- $R$  jets).

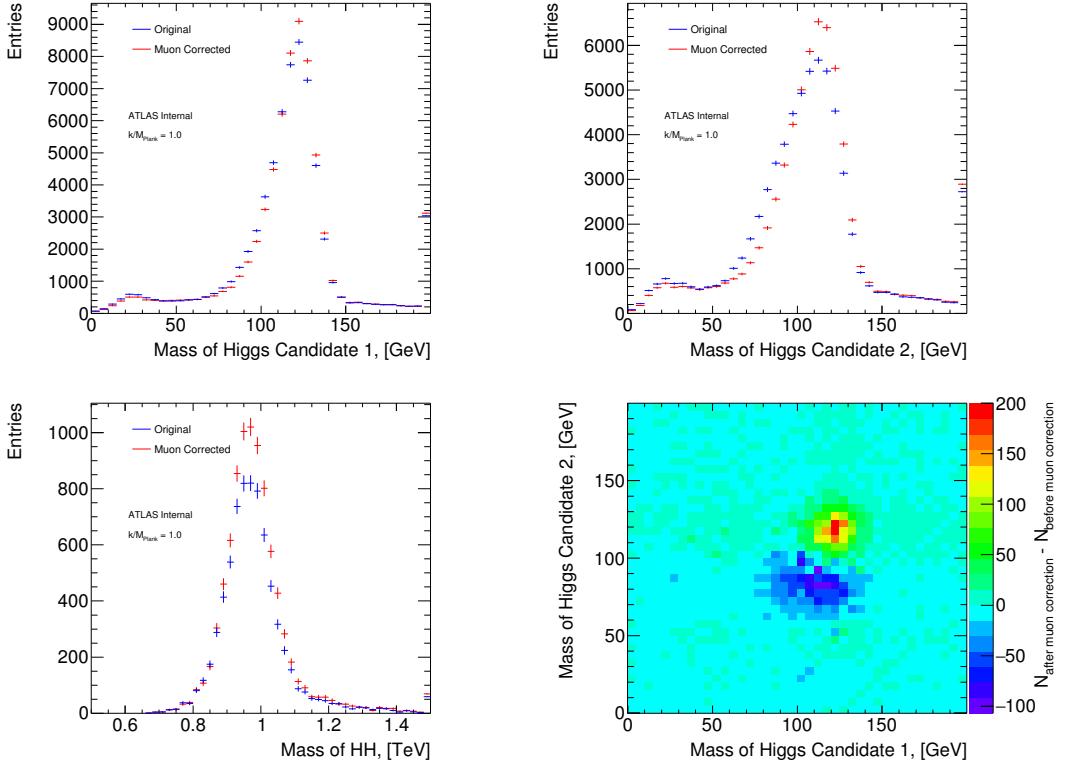
The track jets are required to have  $p_T > 10 \text{ GeV}$ ,  $|\eta| < 2.5$  and at least two tracks associated with it. A track jet is considered  $b$ -tagged if it has  $\text{MV}_{2\text{C}10} > 0.6455$ , see Section 3.3 for details. Each large- $R$  jet is required to have at least one, not two, track jet Ghost associated with it. This accounts for the  $R = 0.2$  track jets merging at really high boost, from signals above 2.5 TeV. If there are more than one track jet contained in the large- $R$  jet, they are also sorted by  $p_T$ . The highest  $p_T$  track jet is named as the leading track jet, and the second highest  $p_T$  one is named as the subleading track jet. Only the two highest  $p_T$  track jet is considered in this thesis. The one or two track jets  $\Delta R$  match to the truth  $b$  quarks 80%, as shown in Figure 5.5. In the Figure, red means the truth Higgs doesn't match the large- $R$  jet. Blue means both truth  $b$  matches the two track jets. Orange means two truth  $b$  matches one same track jet. Green indicates one  $b$  quark has  $\Delta R < 0.2$  for both leading and sub-leading track jet, and the other  $b$  is matched to one of the two track jets. Pink means one  $b$  quark matches to one of the two track jets, the other  $b$  doesn't match to the two leading track jets.



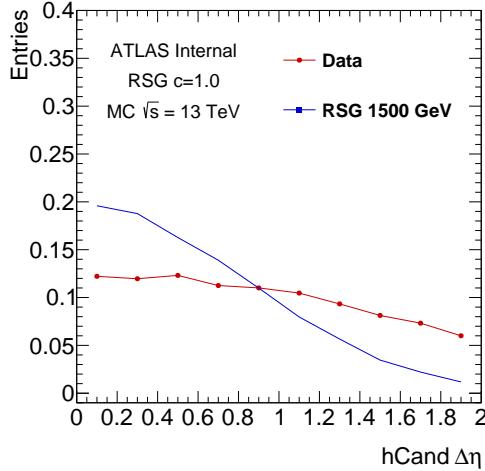
**Figure 5.5:** Percentage of  $\Delta R < 0.2$  matching truth  $b$ 's to track jets (leading Higgs on the left, subleading Higgs on the right) for different  $G_{KK}^*$  mass. The cases listed in the legend are orthogonal to each other. The cases not listed on the legend (including when a truth  $b$  is not contained in the large- $R$  jet) happen in total at most 1.6% of the time for a given  $G_{KK}^*$  mass.

A further muon correction accounts for energy loss due to leptonic  $b$ -hadron decays with a muon in the final state. The muon-in-jet corrections are applied only after the fiducial large- $R$  jet requirements on  $p_T$  and  $\eta$ . The muons are required have  $\Delta R < 0.2$  with the  $b$ -tagged track jets within each large- $R$  jet. In case more than one muon is found within a track jet, only the muon with the smallest  $\Delta R$  is considered. If two  $b$ -tagged track jets are found to have muons, both corrections are considered. The four-momenta of the matched muon is added to the large- $R$  jet four-momentum, with the muon calorimeter energy deposits subtracted. This correction is only applied to the calorimeter mass portion of the combined mass. The muon-in-jet correction improves the large- $R$  jet mass resolution by approximately 5%, and Figure 5.6 shows the impact of this correction on the 1 TeV  $G_{KK}^*$ .

Finally, the Higgs candidates (large- $R$  jets) are also required to have  $|\Delta\eta| = |\eta_{\text{leadJ}} - \eta_{\text{subJ}}| < 1.7$ . This is because the spin 2  $G_{KK}^*$  are produced mostly through s-channel, while the multijet events could also be produced through t-channels or u-channel. Figure 5.7 shows the distribution for sig-



**Figure 5.6:** Kinematics of 1 TeV  $G_{KK}^*$  before and after muon-in-jet corrections. The reconstructed Higgs masses (top row, left for leading large- $R$  jet, right for subleading large- $R$  jet) are closer to 125 GeV after the correction, which improves the signal efficiency for the signal region selection by  $\sim 10\%$  (bottom row, left for  $m_{jj}$ , right for event distribution differences on the leading-subleading large- $R$  jet mass plane.).



**Figure 5.7:** After large- $R$  jet requirements, normalized  $\Delta\eta_{JJ}$  distribution in  $1.5 \text{ TeV } G_{KK}^*$  and data, where the data consists of mostly multijet events ( $> 90\%$ ). The background multijet event is flatter in  $\Delta\eta_{JJ}$  distribution.

nal sample and data inclusive  $b$ -tag region  $\Delta\eta_{JJ}$  distribution. This cut is not entirely optimal for Scalar signals due to the different spin, yet it is fixed for both  $G_{KK}^*$  and Scalar selections.

#### 5.4 RESOLVED VETO

In order to avoid events being reconstructed by both the resolved and the boosted analysis, events that pass the resolved signal region selections are vetoed in the boosted analysis. This is a political decision, and the effect of vetoing boosted event selection in the resolved analysis is never tested. The gain is a full statistical combination of the resolved and boosted result. For boosted analysis, it hurts the sensitivity up to  $1.5 \text{ TeV}$  for resonance signals. Hence it is necessary to introduce the resolved selection.

For resolved analysis, four small- $R$  jets with the highest  $b$ -tagging score are paired to construct two Higgs boson candidates. Each jet must have  $p_T > 40 \text{ GeV}$ ,  $|\eta| < 2.5$ ,  $\text{MV2c10} > 0.8244$  (small- $R$  jet 70%  $b$ -tagging working point). Pairings of jets into Higgs boson candidates are only accepted if they

satisfy the following requirements, where  $m_{4j}$  is expressed in GeV:

if  $m_{4j} < 1250$  GeV:

$$\frac{360 \text{ GeV}}{m_{4j}} - 0.5 < \Delta R_{jj}^{lead} < \frac{653 \text{ GeV}}{m_{4j}} + 0.475; \quad \frac{235 \text{ GeV}}{m_{4j}} < \Delta R_{jj}^{subl} < \frac{875 \text{ GeV}}{m_{4j}} + 0.35 \quad (5.1)$$

if  $m_{4j} > 1250$  GeV:

$$0 < \Delta R_{jj}^{lead} < 1; \quad 0 < \Delta R_{jj}^{subl} < 1 \quad (5.2)$$

In these expressions,  $\Delta R_{jj,lead}$  is the angular distance between jets in the leading Higgs boson candidate and  $\Delta R_{jj,subl}$  for the sub-leading candidate. The leading Higgs boson candidate is defined to be the candidate with the highest scalar sum of jet  $p_T$ . This requirement efficiently rejects jet-pairings where one of the  $b$ -tagged jets is not consistent with that originating from a Higgs boson decay. The specific cut values in this and the following requirements were chosen to maximize the sensitivity to the signal.

Also, mass-dependent requirements are made on the leading Higgs boson candidate  $p_T$ , and the sub-leading Higgs boson  $p_T$ :

$$p_T^{lead} > 0.5m_{4j} - 105 \text{ GeV}, \quad p_T^{subl} > 0.33m_{4j} - 75 \text{ GeV} \quad (5.3)$$

where  $m_{4j}$  is again expressed in GeV.

A further ( $m_{4j}$ -independent) requirement is placed on the pseudorapidity difference between the two Higgs boson candidates,  $|\Delta\eta_{bb}| < 1.5$ , which rejects multijet events.

$$|\Delta\eta_{bb}| < 1.1 \quad \text{if } m_{4j} < 850 \text{ GeV}, \quad |\Delta\eta_{bb}| < 2 \times 10^{-3}m_{4j} - 0.6 \quad \text{if } m_{4j} > 850 \text{ GeV} \quad (5.4)$$

Events that have multiple Higgs boson candidates satisfying these requirements (which happens often when  $m_{4j} < 500$  GeV) necessitate an algorithm to choose the correct pairs. In the absence of energy loss through semi-leptonic decays, the optimal choice would be the combination most consistent with the decays of two particles of equal mass. To account for energy loss, the requirement of equal masses is modified. The distance,  $D_{hh}$ , of the pairing's leading and subleading Higgs boson candidate masses,  $(m_{2j}^{\text{lead}}, m_{2j}^{\text{subl}})$  from the line connecting (0 GeV, 0 GeV) and (120 GeV, 110 GeV) is computed, and the pairing with the smallest value of  $D_{hh}$  is chosen. The values of 120 GeV and 110 GeV are chosen because they correspond to the median values of the narrowest intervals that contain 90% of the signal in simulations.  $D_{hh}$  can be expressed as follows:

$$D_{hh} = \frac{|m_{2j}^{\text{lead}} - \frac{120}{110} m_{2j}^{\text{subl}}|}{\sqrt{1 + \left(\frac{110}{120}\right)^2}}. \quad (5.5)$$

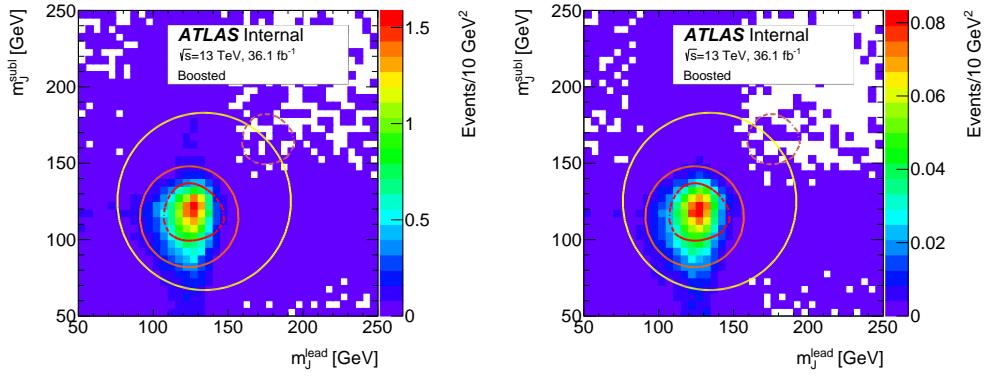
A requirement on the Higgs boson candidates' masses is used to define the resolved signal region:

$$X_{hh-\text{resolved}} = \sqrt{\left(\frac{m_{2j}^{\text{lead}} - 120 \text{ GeV}}{0.1m_{2j}^{\text{lead}}}\right)^2 + \left(\frac{m_{2j}^{\text{subl}} - 110 \text{ GeV}}{0.1m_{2j}^{\text{subl}}}\right)^2} < 1.6, \quad (5.6)$$

where the  $0.1m_{2j}$  terms represent the widths of the leading and sub-leading Higgs boson candidate mass distributions, derived from simulation. The signal region is shown as the inner region of Figure ???. In summary, for any event can make a Higgs candidate through the  $D_{hh}$  minimization and passing through the resolved signal region  $X_{hh-\text{resolved}}$  cut, it is rejected in the boosted selection.

## 5.5 2D HIGGS MASS CUT

To separate di-Higgs decays from background productions like QCD multi-jets and top, requirements on the leading and subleading large- $R$  jet masses are imposed. The signal region is defined



**Figure 5.8:** For RSG  $c = 1.0$  samples, number of events as a function of leading Higgs candidate mass and subleading Higgs candidate mass, for 1.2 TeV (left) signal and 2 TeV (right) signal samples. The red dotted line in the center correspond to the signal region, passing  $X_{hh} < 1.6$ .

using the expression 5.7:

$$X_{hh} = \sqrt{\left(\frac{m_J^{\text{lead}} - 124 \text{ GeV}}{0.1(m_J^{\text{lead}})}\right)^2 + \left(\frac{m_J^{\text{subl}} - 115 \text{ GeV}}{0.1(m_J^{\text{subl}})}\right)^2} \quad (5.7)$$

The denominator of each term in  $X_{hh}$  can be interpreted as a resolution on the reconstructed mass of 10% for the leading and subleading jets, hence  $X_{hh}$  can be interpreted as a  $\chi^2$  compatibility with the di-Higgs hypothesis. The subleading jet mass value of 115 GeV is chosen after investigating the signal jet masses in MC. The subleading large- $R$  jet typically has a reconstructed mass which is biased downward. This is partly due to the ordering of the large- $R$  jets in  $p_T$ , which makes the subleading jet towards lower energy. The energy losses from neutrinos in leptonic  $b$  decays, cracks in the calorimeter, and other effects also contributes. The signal region requires  $X_{hh} < 1.6$ . This cut gives nearly optimal performance. A more optimal signal region definition, using asymmetric signal jet mass resolution and a momentum dependent cut accounting for the higher mass resolution of larger  $p_T$  jets, can improve the overall sensitivity by 2 – 8%. Since the gain in sensitivity is small, the signal region is kept to be consistent with the  $X_{hh}-resolved$ . Figure 5.8 shows the  $G_{KK}^*$  MC  $m_{z_j}^{\text{lead}}$ -

$m_{2j}^{\text{subl}}$  2D distribution.

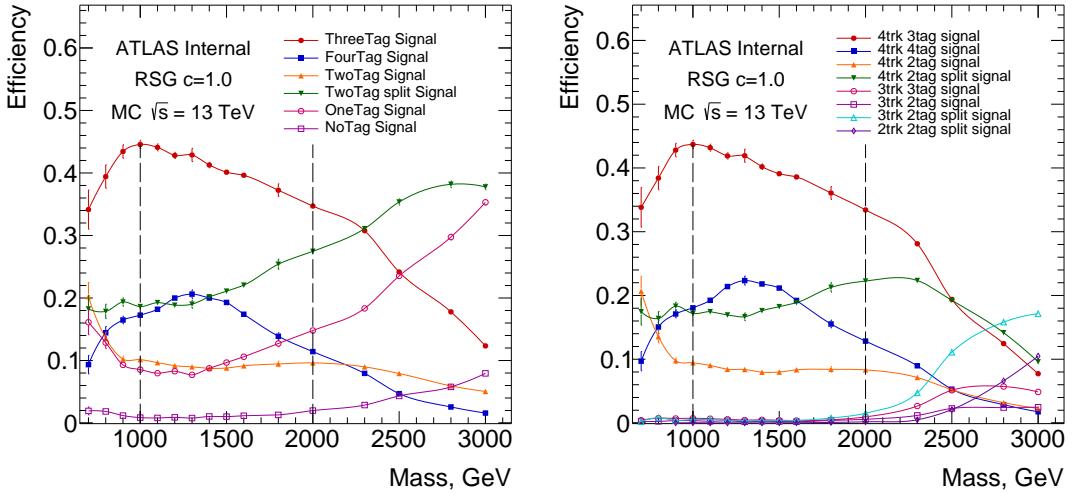
## 5.6 NUMBER OF $b$ -TAGGING REQUIREMENT

Passing basic object selection and  $X_{hh}$  cut, the signal region selection is defined by requiring multiple  $b$ -tags which are consistent with the di-Higgs decay. The presence of two  $b \rightarrow b\bar{b}$  decays in the final state naturally suggests requiring 4 track jets passing  $b$ -tagging requirements, and this is defined as the  $4b$  selection.

The  $4b$  requirement has an overall efficiency of roughly  $\epsilon^4$ , where  $\epsilon$  is the  $b$ -tagging efficiency chosen to be 70%. This means an overall  $0.7^4 \sim 0.24$  probability, but having one actual  $b$ -jet failing while the other three pass has probability  $3 \times 0.7^3 \times (1 - 0.7) \sim 0.31$ . Therefore, a  $3b$  selection is also introduced to recover the signal efficiency. An event with  $3b$ -tags must have at least  $3b$ -tagged track jets, but can have any number of additional un-tagged track jets. In  $4b$  and  $3b$ , each Higgs candidate can have at most two  $b$ -tagged track jets, hence  $\geq 3b$ -tagged trackjets cannot be in the same large- $R$  jet.

At the highest resonance mass, the Lorentz boost of the Higgs boson can be large enough to collimate the daughter  $b$ -quarks below the distance scale resolvable by the track jets ( $R = 0.2$ ). A third signal region is denoted by two-tag-split or simply  $2bs$ . It requires exactly one  $b$ -tagged track jet is found in each Higgs candidate, plus an arbitrary number of track jets that must fail the  $b$ -tag.

The other  $b$ -tagging situations are also sorted and studied, but not as signal regions due to their relatively small acceptance.  $2b$  region is defined as one large- $R$  jet has two  $b$ -tagged track jets, and the other large- $R$  jet has no  $b$ -tagged track jet.  $1b$  region is defined as one large- $R$  jet has one and only one  $b$ -tagged track jets, and the other large- $R$  jet has no  $b$ -tagged track jet.  $0b$  region is defined as both large- $R$  jets have no  $b$ -tagged track jet.

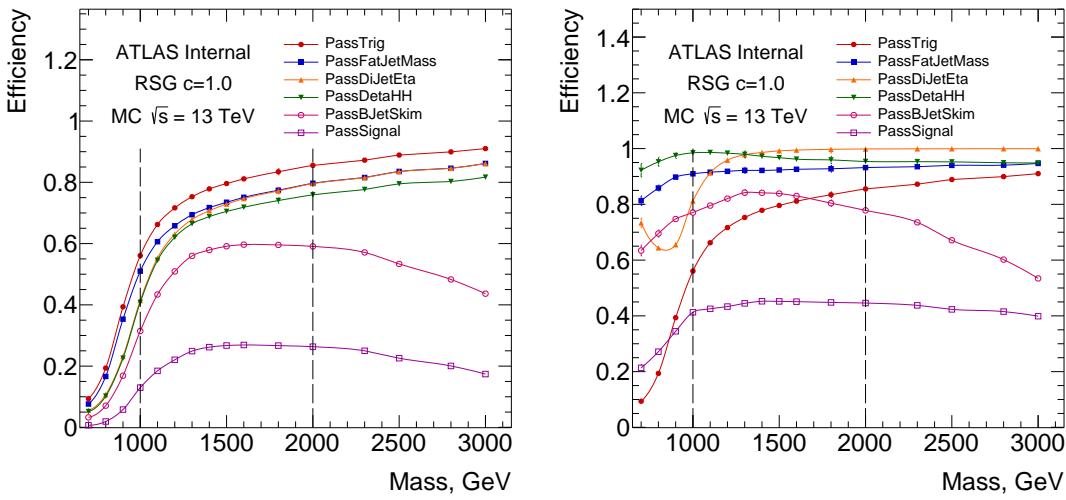


**Figure 5.9:** Signal fraction in different  $b$ -tag categories (left) and detailed fraction in different number of track jet and  $b$ -tag categories (right) as a function of signal resonance mass hypothesis for selection cuts. The efficiencies are relative to the total number of events passing the 2D mass cut.

The MC events passing all signal region selections are sorted into different  $b$ -tagging categories. This is shown in Figure 5.9. For masses above 2.5 TeV, the  $2bs$  region (where each large- $R$  jet has exactly one  $b$ -tagged track jet) significantly improves the acceptance.

## 5.7 SIGNAL EFFICIENCY AND CUTFLOW

Acceptance means purely geometric fiducial volume of the detector. Efficiency refers to purely detector effectiveness in finding objects. The signal efficiency as a function of  $C_{KK}^*$  resonance mass is shown in Figure 5.10, both for the absolute signal efficiency and for the efficiency relative to the previous cut in the selection. Above a mass of  $\sim 1$  TeV, the reconstruction of high momentum large- $R$  jets with small  $\Delta\eta$  is efficient. Across the mass range considered, the signal jet masses requirement ( $X_{hh}$ ) and  $b$ -tagging requirements are  $\mathcal{O}(20\%)$  efficient relative to the previous cuts.



**Figure 5.10:** Absolute (left) and relative (right) signal efficiency as a function of RSG  $c=1.0$  signal resonance mass hypothesis for selection cuts. The relative efficiency is defined from the previous cut, where the order of cuts is given by the legend. PassTrig means the event passes the trigger selection; PassDiJetPt means the event passes the leading and sub-leading jet  $p_T$  cuts; PassDiJetEta means the event passes the leading and sub-leading jet  $\eta$  cuts; PassDeltaH means the events passes the  $|\Delta\eta| < 1.7$  cut; PassBJetSkim means the event contains at least two  $b$ -tagged track jets, inclusive of  $2b$ ,  $2bs$ ,  $3b$  and  $4b$  configurations; PassSignal means the event passes the signal region cut  $X_{bb} < 1.6$ .

The selection efficiency at various stages for  $G_{KK}^*$  with  $c = 1.0$ ,  $G_{KK}^*$  with  $c = 2.0$ , and Heavy Scalar signal samples of all mass points can be found in Table 5.2, 5.3 and 5.4.

**Table 5.2:** The selection efficiency for  $G_{KK}^* \rightarrow hh \rightarrow b\bar{b}b\bar{b}$  events ( $c = 1.0$ ) at each stage of the event selection. Uncertainties are the MC stat uncertainty only.

Resonance Mass [GeV]	Mini-ntuple Skimming	2 large-R jets	$\Delta\eta$	Xhh < 1.6	2bs SR	3b SR	4b SR
500	317.31 ± 6.0	295.75 ± 5.79	164.5 ± 4.32	8.45 ± 0.99	1.08 ± 0.37	2.14 ± 0.52	0 ± 0
600	269.07 ± 3.64	247.94 ± 3.5	136.31 ± 2.59	11.31 ± 0.76	2.57 ± 0.37	3.84 ± 0.45	0.66 ± 0.19
700	253.68 ± 3.35	226.93 ± 3.16	124.83 ± 2.35	16.79 ± 0.86	3.74 ± 0.42	6.99 ± 0.56	1.91 ± 0.29
800	286.26 ± 2.28	245.36 ± 2.11	129.2 ± 1.53	24.41 ± 0.67	5.11 ± 0.31	11.27 ± 0.46	4.13 ± 0.27
900	306.51 ± 1.61	275.57 ± 1.52	158.03 ± 1.15	40.72 ± 0.59	8.81 ± 0.28	19.76 ± 0.41	7.5 ± 0.25
1000	238.2 ± 0.98	226.98 ± 0.96	165.2 ± 0.82	52.86 ± 0.47	10.87 ± 0.22	26.0 ± 0.33	10.07 ± 0.2
1100	164.5 ± 0.63	160.94 ± 0.63	132.53 ± 0.57	45.26 ± 0.34	9.55 ± 0.16	21.88 ± 0.23	9.03 ± 0.14
1200	109.24 ± 0.41	107.92 ± 0.4	93.45 ± 0.38	33.53 ± 0.23	6.96 ± 0.11	15.8 ± 0.16	7.38 ± 0.1
1300	72.72 ± 0.59	72.2 ± 0.59	63.74 ± 0.56	24.19 ± 0.35	5.02 ± 0.17	11.33 ± 0.24	5.45 ± 0.16
1400	48.83 ± 0.17	48.61 ± 0.17	42.96 ± 0.16	16.62 ± 0.1	3.72 ± 0.052	7.61 ± 0.07	3.68 ± 0.046
1500	33.13 ± 0.12	33.02 ± 0.12	29.25 ± 0.11	11.31 ± 0.07	2.67 ± 0.036	5.08 ± 0.047	2.44 ± 0.031
1600	22.81 ± 0.08	22.75 ± 0.08	20.16 ± 0.075	7.74 ± 0.048	1.93 ± 0.025	3.48 ± 0.032	1.53 ± 0.02
1800	11.2 ± 0.1	11.18 ± 0.1	9.93 ± 0.094	3.71 ± 0.059	1.1 ± 0.034	1.6 ± 0.038	0.6 ± 0.022
2000	5.72 ± 0.021	5.71 ± 0.021	5.07 ± 0.019	1.83 ± 0.012	0.6 ± 0.0072	0.76 ± 0.0076	0.25 ± 0.0041
2250	2.61 ± 0.0088	2.61 ± 0.0088	2.32 ± 0.0083	0.78 ± 0.005	0.31 ± 0.0032	0.3 ± 0.003	0.078 ± 0.0014
2500	1.24 ± 0.0054	1.24 ± 0.0054	1.11 ± 0.0051	0.33 ± 0.0028	0.16 ± 0.002	0.11 ± 0.0016	0.021 ± 0.00066
2750	0.6 ± 0.0026	0.6 ± 0.0026	0.54 ± 0.0025	0.14 ± 0.0013	0.081 ± 0.00099	0.038 ± 0.00065	0.0055 ± 0.00024
3000	0.3 ± 0.0011	0.3 ± 0.0011	0.27 ± 0.0011	0.058 ± 0.00051	0.039 ± 0.00041	0.013 ± 0.00023	0.0016 ± 8e-05

**Table 5.3:** The selection efficiency for  $G_{KK}^* \rightarrow hh \rightarrow b\bar{b}b\bar{b}$  events ( $c = 2.0$ ) at each stage of the event selection. Uncertainties are the MC stat uncertainty only.

Resonance Mass [GeV]	Mini-ntuple Skimming	2 large-R jets	$\Delta\eta$	Xhh < 1.6	2bs SR	3b SR	4b SR
500	3705.15 ± 40.86	3479.44 ± 39.59	2325.18 ± 32.37	568.04 ± 16.17	122.56 ± 7.76	253.49 ± 10.78	100.7 ± 6.53
600	2549.14 ± 22.55	2374.01 ± 21.76	1591.92 ± 17.82	396.96 ± 9.03	89.01 ± 4.46	178.63 ± 6.05	74.31 ± 3.71
700	1928.4 ± 13.57	1782.85 ± 13.04	1183.86 ± 10.63	320.53 ± 5.62	71.38 ± 2.76	148.41 ± 3.81	59.31 ± 2.31
800	1595.14 ± 8.82	1457.71 ± 8.43	958.89 ± 6.84	268.75 ± 3.67	64.14 ± 1.86	123.48 ± 2.47	49.43 ± 1.51
900	1264.78 ± 5.77	1179.88 ± 5.58	819.75 ± 4.65	251.29 ± 2.61	55.63 ± 1.27	119.44 ± 1.79	48.72 ± 1.11
1000	891.0 ± 3.66	856.95 ± 3.59	662.54 ± 3.15	219.04 ± 1.84	49.45 ± 0.91	104.31 ± 1.26	42.97 ± 0.78
1100	595.58 ± 2.98	581.72 ± 2.95	481.67 ± 2.68	167.96 ± 1.61	37.64 ± 0.79	78.28 ± 1.09	34.59 ± 0.7
1200	390.84 ± 1.69	385.41 ± 1.68	330.23 ± 1.55	118.0 ± 0.94	26.23 ± 0.46	54.18 ± 0.64	25.34 ± 0.42
1300	257.66 ± 0.94	255.35 ± 0.94	222.37 ± 0.88	82.11 ± 0.54	19.04 ± 0.27	37.8 ± 0.37	16.99 ± 0.23
1400	172.09 ± 0.72	171.02 ± 0.71	150.22 ± 0.67	56.23 ± 0.42	13.6 ± 0.22	25.36 ± 0.28	11.78 ± 0.18
1500	116.25 ± 0.41	115.72 ± 0.41	101.92 ± 0.39	38.5 ± 0.24	9.94 ± 0.13	17.04 ± 0.16	7.64 ± 0.1
1600	80.09 ± 0.28	79.82 ± 0.28	70.48 ± 0.26	26.24 ± 0.16	7.01 ± 0.09	11.62 ± 0.11	4.92 ± 0.067
1800	38.99 ± 0.14	38.9 ± 0.14	34.39 ± 0.13	12.65 ± 0.081	3.82 ± 0.047	5.46 ± 0.053	2.02 ± 0.03
2000	19.94 ± 0.088	19.91 ± 0.088	17.68 ± 0.083	6.17 ± 0.05	2.15 ± 0.031	2.52 ± 0.032	0.85 ± 0.017
2250	9.02 ± 0.031	9.01 ± 0.031	7.99 ± 0.029	2.62 ± 0.017	1.03 ± 0.011	1.02 ± 0.01	0.28 ± 0.0051
2500	4.28 ± 0.016	4.28 ± 0.016	3.8 ± 0.015	1.13 ± 0.0083	0.52 ± 0.0058	0.4 ± 0.0048	0.098 ± 0.0022
3000	1.07 ± 0.004	1.07 ± 0.004	0.96 ± 0.0038	0.23 ± 0.0019	0.13 ± 0.0015	0.062 ± 0.00097	0.013 ± 0.00043

**Table 5.4:** The selection efficiency for  $H \rightarrow hh \rightarrow b\bar{b}b\bar{b}$  events at each stage of the event selection.

Resonance Mass [GeV]	Mini-ntuple Skimming	2 large-R jets	$\Delta\eta$	Xhh < 1.6	2bs SR	3b SR	4b SR
500	1557.94 ± 136.12	1022.77 ± 110.29	95.14 ± 33.64	11.69 ± 11.69	0 ± 0	11.69 ± 11.69	0 ± 0
600	3289.78 ± 123.99	2542.11 ± 108.99	485.99 ± 47.66	54.55 ± 15.77	18.73 ± 9.39	9.17 ± 6.49	0 ± 0
700	4655.21 ± 94.59	3855.64 ± 86.09	1237.8 ± 48.78	142.42 ± 17.03	28.55 ± 7.74	52.6 ± 10.73	7.69 ± 3.85
800	7506.31 ± 81.79	6020.56 ± 73.25	2150.64 ± 43.78	320.63 ± 17.02	67.63 ± 7.83	139.97 ± 11.23	47.57 ± 6.75
900	9732.89 ± 61.17	8400.91 ± 56.83	3574.63 ± 37.07	866.13 ± 17.76	188.71 ± 8.71	377.92 ± 12.12	127.7 ± 6.99
1000	7516.07 ± 37.72	7033.18 ± 36.49	4496.85 ± 29.18	1351.1 ± 16.2	303.71 ± 7.88	650.89 ± 11.19	234.2 ± 6.57
1100	4731.39 ± 21.54	4563.4 ± 21.15	3485.58 ± 18.49	1135.18 ± 10.7	251.77 ± 5.18	539.51 ± 7.36	215.39 ± 4.5
1200	2853.51 ± 12.23	2782.42 ± 12.07	2253.95 ± 10.87	786.92 ± 6.53	175.53 ± 3.21	366.01 ± 4.44	158.29 ± 2.8
1300	1700.83 ± 7.01	1668.05 ± 6.94	1362.91 ± 6.27	494.36 ± 3.84	107.0 ± 1.86	224.19 ± 2.58	107.55 ± 1.7
1400	1016.14 ± 4.03	999.98 ± 4.0	802.44 ± 3.59	296.46 ± 2.22	65.49 ± 1.1	133.86 ± 1.49	65.58 ± 0.99
1500	621.34 ± 2.41	613.46 ± 2.39	484.92 ± 2.13	179.29 ± 1.32	42.75 ± 0.68	79.32 ± 0.88	36.77 ± 0.56
1600	386.3 ± 1.46	382.44 ± 1.46	297.63 ± 1.28	109.99 ± 0.8	27.68 ± 0.42	49.3 ± 0.53	20.92 ± 0.32
1800	154.8 ± 0.58	153.5 ± 0.57	116.52 ± 0.5	42.24 ± 0.31	12.41 ± 0.18	18.2 ± 0.2	6.66 ± 0.11
2000	65.4 ± 0.24	65.02 ± 0.24	48.57 ± 0.2	16.89 ± 0.12	5.64 ± 0.076	7.01 ± 0.079	2.18 ± 0.041
2250	23.84 ± 0.085	23.73 ± 0.085	17.44 ± 0.073	5.57 ± 0.042	2.25 ± 0.028	2.11 ± 0.025	0.52 ± 0.012
2500	9.2 ± 0.032	9.17 ± 0.032	6.72 ± 0.028	1.9 ± 0.015	0.92 ± 0.011	0.64 ± 0.0086	0.11 ± 0.0034
2750	3.73 ± 0.013	3.73 ± 0.013	2.71 ± 0.011	0.63 ± 0.0054	0.37 ± 0.0042	0.17 ± 0.0027	0.021 ± 0.00093
3000	1.59 ± 0.0054	1.59 ± 0.0054	1.15 ± 0.0046	0.22 ± 0.0021	0.15 ± 0.0017	0.044 ± 0.0009	0.0038 ± 0.00025

*Even in the darkest night. Stars and angels still shine  
bright.*

Bear

# 6

## Background Estimation

### 6.1 BACKGROUND ESTIMATION

A similar circular variable can be defined in the two-dimensional mass plane,  $R_{hh}$ . The circular region  $R_{hh}$  has the same central values as  $X_{hh}$ , but without resolution terms in the denominators and is defined as:

$$R_{hh} = \sqrt{(m_j^{\text{lead}} - 124 \text{ GeV})^2 + (m_j^{\text{subl}} - 115 \text{ GeV})^2} \quad (6.1)$$

The region defined by  $1.6 < X_{hh}$  and  $R_{hh} < 33$  is the control region. It will be discussed in Section 7.1.3. The cut value was optimized to allow for a reasonable sized sample (twice the statistics as the signal region) in the control region with kinematics similar to the signal region, and avoiding the large contributions of the  $t\bar{t}$  sample when the large- $R$  jets have a mass near the top quark mass (with

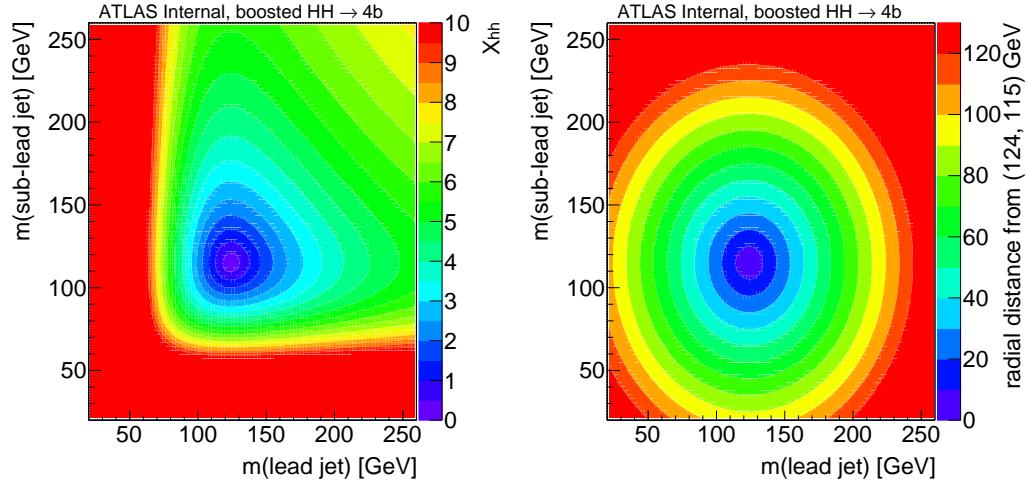
$m_J > 160 \text{ GeV}$ ).

Similarly,  $R_{hh}^{\text{high}}$ , the circular region that has the shifted central values up by 10  $\text{GeV}$  is defined using the variable:

$$R_{hh}^{\text{high}} = \sqrt{(m_J^{\text{lead}} - 134 \text{ GeV})^2 + (m_J^{\text{subl}} - 125 \text{ GeV})^2} \quad (6.2)$$

In Run-I, and in the the 2015 analysis, the sideband was defined to be all events not in the signal or control regions. However, the kinematics of events with very large and very small large- $R$  jet masses may not be the same as those within the signal region. To avoid biasing effects from extremely low mass or extremely high mass large- $R$  jets, the sideband region is also redesigned to be like the control region, but at  $R_{hh} > 33$  and  $R_{hh}^{\text{high}} < 58$ . The shift upwards helps to capture enough  $t\bar{t}$  events in the normalization estimates, as described in Section 7.1.5.

The values of the  $X_{hh}$  and  $R_{hh}$  variables can be seen graphically in Figure 6.1.



**Figure 6.1:** Values of the  $X_{hh}$  and  $R_{hh}$  variables, which are shapes in the two-dimensional plane of the large- $R$  jet masses used to define signal, control, and sideband regions. For both variables, a smaller value indicates the jets are closer to the Higgs mass.

The number of events in the control region and sideband region as a function of Resonance mass

is shown in Figure 6.2. For  $2b$ s,  $3b$  and  $4b$ , each region has the number of events decrease from signal region to control region to sideband regions.

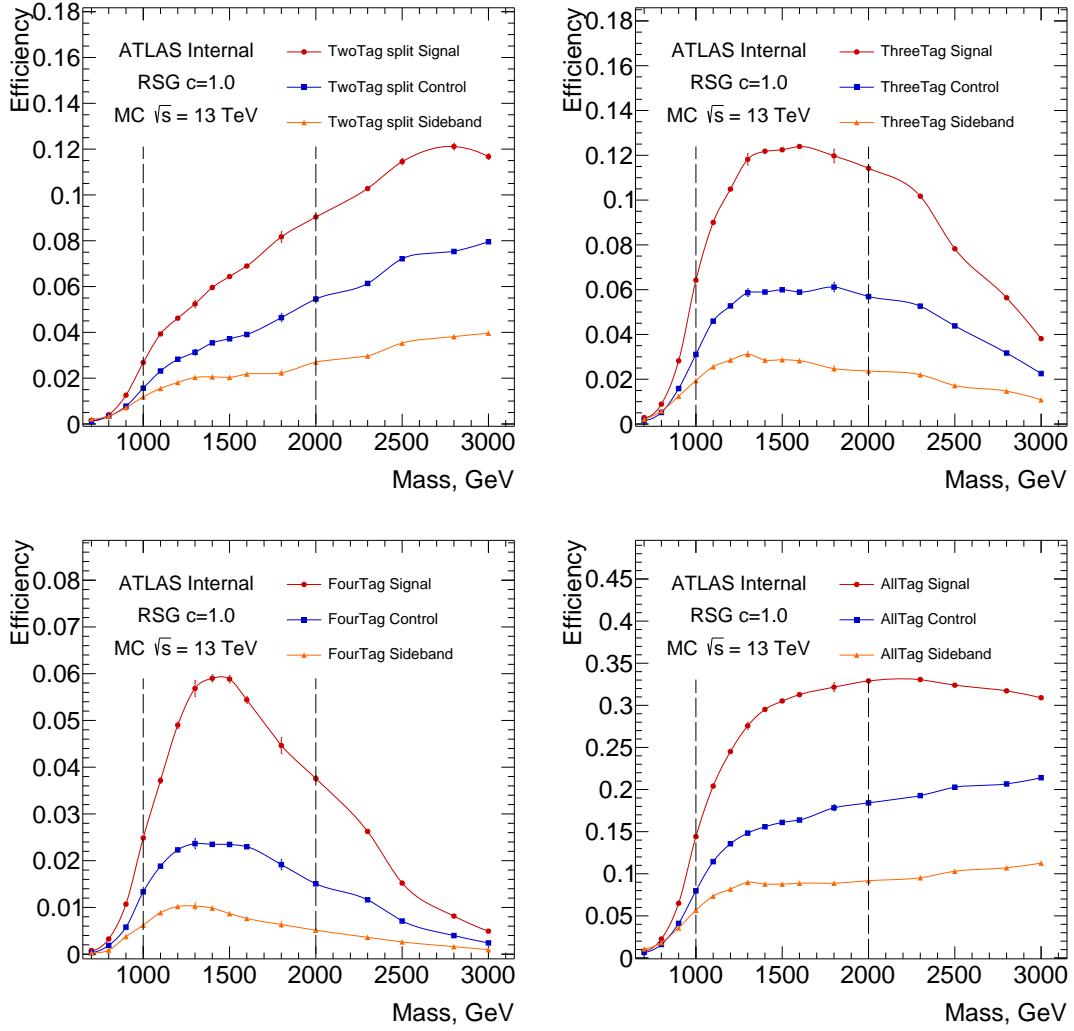
### 6.1.1 OVERVIEW

The primary backgrounds to this analysis in the four  $b$ -jet signal region, in order of size, are QCD multi-jet production ( $\sim 95\%$ ),  $t\bar{t}$  ( $\sim 5\%$ ), and  $Z+jets$  ( $< 1\%$ ), where the percentages are the expected fraction of the background coming from each source. In the three  $b$ -jet signal region, the fractions are QCD ( $\sim 90\%$ ),  $t\bar{t}$  ( $\sim 10\%$ ), and  $Z+jets$  ( $< 1\%$ ). In the two  $b$ -jet split signal region, the fractions are QCD ( $\sim 80\%$ ),  $t\bar{t}$  ( $\sim 20\%$ ), and  $Z+jets$  ( $< 1\%$ ).

QCD is by far the dominant background. However, there is no reliable, high-statistics Monte Carlo simulation sample in this region of phase space (i.e with three or four  $b$ -jets collected into two high- $p_T$  large radius jets) and thus a data-driven background estimation is needed. (See Appendix ??.) For the  $t\bar{t}$ background, Monte Carlo simulation samples of reasonable size are available, and thus can be used to guide an estimation of this background. The  $Z+jets$  background is small enough that we will rely on the Monte Carlo simulation of  $Z+heavy$  flavor jets.  $ZZ \rightarrow b\bar{b}b\bar{b}$  has been estimated to be completely negligible using a particle-level analysis, with less than one event expected after three  $b$ -tags are required, which will be further heavily suppressed by the  $X_{hh}$  requirement.

The QCD background prediction relies on finding a region which is similar enough in event properties that it can be used to estimate the shapes of the expected QCD background. This region is identical to the signal region defined by the full selection, with the exception that the events must have less  $b$ -tagged track jets:

- For the  $2b$ s category, the  $1b$  sample is used for modeling.
- For the  $3b$  and  $4b$  categories, the  $2b$  sample - where the two  $b$ -tagged trackjet are in the same large- $R$  jet - is used for modeling.



**Figure 6.2:** Detailed signal efficiency in different signal/control/sideband regions as in  $2b$ s (top left,  $3b$  (top right),  $4b$  (bottom left) and inclusive b-tagged regions, which include  $2b$ ,  $1b$  and  $0b$  as well, (bottom right) as a function of signal resonance mass hypothesis for selection cuts. The efficiencies are relative to the total number of events in the preselection.

To prevent differences in the number of track jets from biasing the dijet mass distribution, the  $1b$ -tagged region requires that each large- $R$  jet has at least one track jet (to model  $2bs$ , ie.  $2b$  tag split). Similarly, the  $2b$ -tagged region requires that one large- $R$  jet has at least one track jet and the other one has at least two track jets (to model  $3b$ ), and each large- $R$  jet has at least two track jets (to model  $4b$ ).

However, this less  $b$ -tagged region only supplies the shapes of the expected background and not the total yield, and a second control sample, which we denote the *Sideband* region, is used to estimate the yield. The Sideband is obtained by doing the full analysis selection, except instead of the  $X_{hh}$  cut an alternative criteria on the large radius jet masses is used, that  $33 < R_{hh} \text{ and } R_{hh}^{\text{high}} < 58 \text{ GeV}$ . To validate this approach, a third region, which we denote the *Control* region, is centered on the signal region in the plane of the two large radius jet masses but does not include the signal region, such that  $R_{hh} < 33 \text{ GeV}$ . The control region is used to validate the background estimations before unblinding. The control and sideband regions are optimized, as shown in the following sections, to accurately estimate the rate of the QCD background (and thus allow for an extrapolation from the  $1b/2b$  estimate to a prediction in the  $4b/3b/2bs$  signal regions), whilst giving a control region which has kinematic properties similar to that of the signal region.

The  $t\bar{t}$  background shape is taken from MC. A data-based estimation of the  $t\bar{t}$  background yield is performed simultaneously with the QCD background yield estimation, by means of a binned likelihood fit. In the plane of the two leading large radius jet masses, the main contribution of the  $t\bar{t}$  background lies in the Sideband region. The data distribution in the Sideband region of the leading- $p_T$  large radius jet mass is fit simultaneously with the QCD shape estimate (from the less  $b$ -tagged sample) and with the  $t\bar{t}$  Monte Carlo shape. This fit is done separately in the  $4b$ ,  $3b$ , and  $2bs$  Sideband regions. From this fit, two terms are determined simultaneously:  $\mu_{QCD}$  and  $\alpha_{tt}$ .  $\mu_{QCD}$  is the ratio of the QCD event yield in the  $2bs/3b/4b$  regions to the amount in each corresponding less  $b$ -tagged region.  $\alpha_{tt}$  is the ratio of the fitted ttbar event yield to the yield predicted from ttbar MC.

These two numbers are then used as multiplicative constants in other regions of the mass plane (i.e. the Control or Signal regions) to extrapolate from the rates of the less  $b$ -tagged regions to predictions of rates in the  $2bs/3/4$   $b$ -tagged regions, for estimating the amount of QCD, and to correct the rate of  $t\bar{t}$  production wrt. MC. Hence, the underlying assumption is that these scale factors are roughly constant over the 2D large- $R$  jet mass plane for Sideband/Control/Signal regions, which has been verified by performing these fits in small bins across the 2D mass plane. This is shown in Appendix ???. The correction factors are derived separately for the  $4b$ ,  $3b$ , and  $2bs$  regions.

In this section, we describe this approach in more detail and show its validation in data.

### 6.1.2 DEFINITION OF THE SIDEBAND AND CONTROL REGIONS (SB, CR)

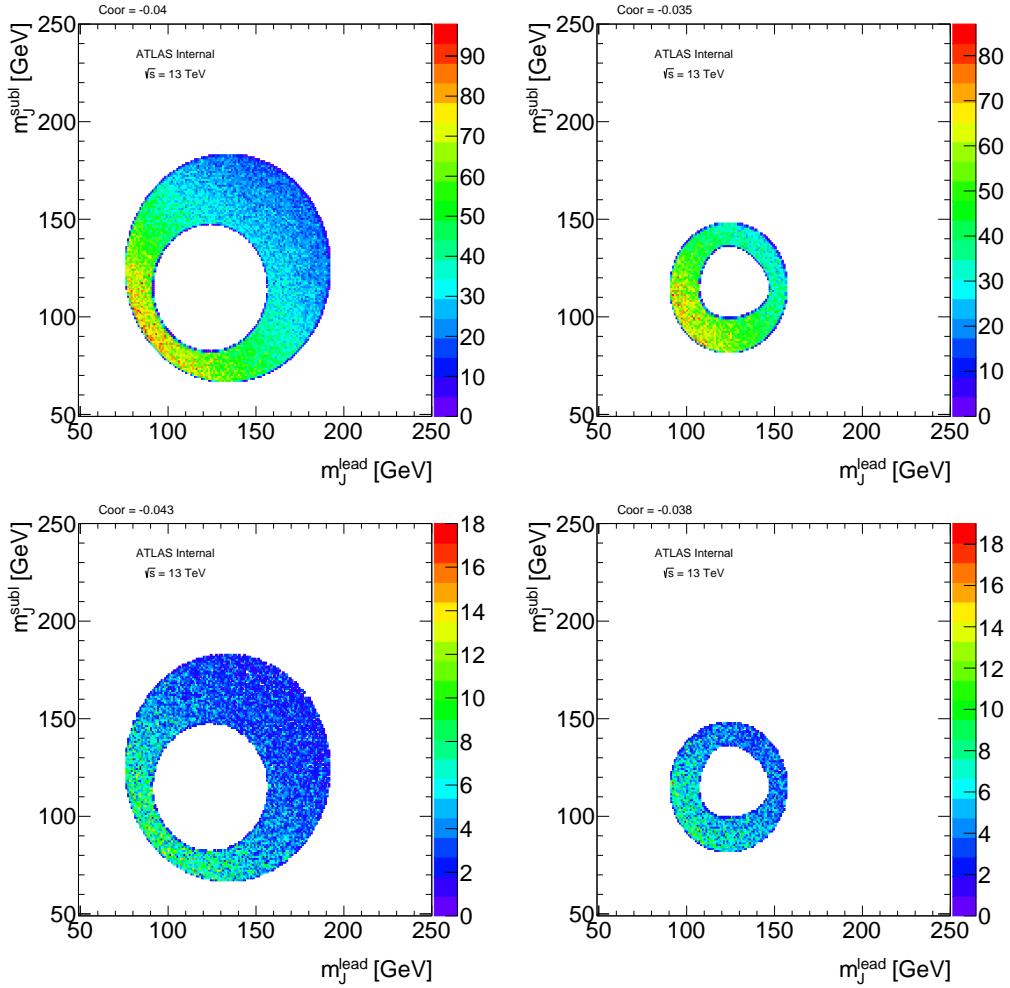
The definitions of the SB, CR, and SR in the leading ( $m_J^{\text{lead}}$ ) and sub-leading ( $m_J^{\text{subl}}$ ) large- $R$  jet mass plane are found in Table 7.1. These regions can be seen in the leading and sub-leading large- $R$  jet mass plane in Figure 7.1. As a reminder, the definition of  $X_{hh}$ ,  $R_{hh}$  and  $R_{hh}^{\text{high}}$  are:

$$X_{hh} = \sqrt{\left(\frac{m_J^{\text{lead}} - 124 \text{ GeV}}{\sigma(m_J^{\text{lead}})}\right)^2 + \left(\frac{m_J^{\text{subl}} - 115 \text{ GeV}}{\sigma(m_J^{\text{subl}})}\right)^2} \quad R_{hh} = \sqrt{(m_J^{\text{lead}} - 124)^2 + (m_J^{\text{subl}} - 115)^2} \quad R_{hh}^{\text{high}} = \sqrt{(m_J^{\text{lead}} - 134)^2 + (m_J^{\text{subl}} - 125)^2} \quad (6.3)$$

Region	Definition
Signal Region (SR)	$X_{hh} < 1.6$
Control Region (CR)	$R_{hh} < 33 \text{ GeV}$ and $X_{hh} > 1.6$
Sideband Region (SB)	$33 \text{ GeV} < R_{hh} \text{ and } R_{hh}^{\text{high}} < 58 \text{ GeV}$

**Table 6.1:** Definitions of the Signal (SR), Sideband (SB) and Control (CR) regions.

The CR is chosen to be as close as possible to the signal region, thus allows a good test for the background predictions, avoids the top mass peak around 175 GeV, and still gives reasonably statistics. The SB definition is optimized so as to also be a reasonable proxy for the events contained in



**Figure 6.3:**  $m_J^{\text{lead}}$  vs.  $m_J^{\text{subl}}$  in data in the  $1b$ -tag (top) and  $2b$ -tag (bottom) selection, the plots show the boundary between the Sideband (left) and Control (right) regions.

the CR and SR. Being farther from the SR means that the exact kinematics will not be the same, but one can avoid very large and very small mass jets not present in the SR by appropriate choice of SB. The optimization of teh CR and SB definitiona can be found in Appendix ???. The choice of SB's impact on the predicted QCD background normalization, which is derived from the SB, can be found in Appendix ??.

### 6.1.3 QCD multi-jets

The QCD multi-jets prediction relies on finding a region which is similar enough in event properties so that it can be used to estimate the shapes of the expected background. This region is defined to be identical to the signal region except requiring both of the large- $R$  jets to pass the  $\geq 2$  (like in  $4b$ ) or  $\geq 1/2$  (like in  $3b$ ) or  $\geq 1$  (like in  $2b$  split) track jet requirement, but have two or one associated  $b$ -tagged track jets only on one of the large- $R$  jets. However, this  $1b/2b$  region only provides shapes of the expected background and not the total yield.

It should be noted that the  $1b/2b$  region is orthogonal to the  $4b/3b/2bs$  signal regions. In addition, the MC predicted  $t\bar{t}$  events in the  $1b/2b$  regions are subtracted from the data to produce the  $1b/2b$  QCD estimation. This procedure follows closely the method used in Run 1 and also used in the resolved analysis, but requiring  $1b$ -tag for the  $2bs$  background estimation.

It should also be noted that the resolved veto will impact the  $4b$  background estimation. Specifically,  $2b$  events are excluded when they have at least two resolved jets that are  $b$  tagged (passing resolved 70% working point) and passing the resolved  $X_{hh} < 1.6$  cut if using two other non  $b$ -tagged resolved jets to make the Higgs candidates. This ensures that a similar sculpting effect is reflected in the background estimation, and a check for this can be found in Appendix ??.

Given the  $1b/2b$  samples, which predict shapes for the QCD background, the normalization of the QCD background is determined in the sideband by fitting the leading jet mass distribution simultaneously with QCD and  $t\bar{t}$  background templates, as described in section 7.1.5. This fit gives a scaling factor for QCD, called  $\mu_{QCD}$  (and for  $t\bar{t}$ , called  $\alpha_{t\bar{t}}$ ) which can be applied to scale the  $1b/2b$  predictions in the CR or SR to the predicted normalizations in those regions.

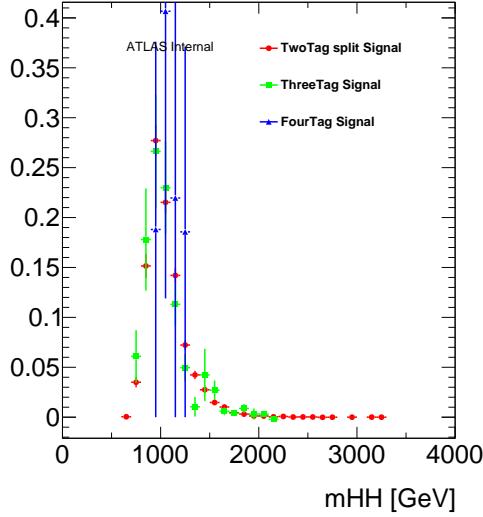
It should be noted that there can be kinematic differences between the  $1b/2b$  samples and the  $4b/3b/2bs$  regions. Thus a kinematic reweighting is applied to correct for such differences, as described in Section 7.1.6.

#### 6.1.4 $t\bar{t}$ BACKGROUND

The number of  $t\bar{t}$  events in the signal region coming mainly from the all-hadronic decay mode (with a smaller contribution from the leptonic + jets decay mode) comprises of around 5-20% of the inclusive total background in the  $4b/3b/2bs$  regions due to the high  $p_T$  threshold imposed on the leading *large – R*calorimeter jet. In addition, the normalization and the shape of  $t\bar{t}$  events in the sideband region can affect the QCD estimate described in the previous section.

For the normalization of the  $t\bar{t}$  background, we start with the MC prediction estimated by scaling the MC sample by the cross section and luminosity, and then applying the boosted event selection. To account for possible differences between data and MC, a normalization scaling factor is derived from a fit to data in the sideband region. Although estimated in the SB, this normalization scaling factor will be applied also in the CR and SR. Separate fits are done in the  $4b/3b/2bs$  SB, thus deriving separate normalization scaling factor for the  $4b/3b/2bs$  samples. For the shape of the  $t\bar{t}$  background, no data driven methods were identified, and thus the MC shape is used.

However, it should be noted that in the  $4b$  and  $3b$  signal region, there were not sufficient MC statistics to get a reasonable shape estimate. As a result, in the  $4b$  and  $3b$  signal region, the  $2bs$  shapes will be used (but the normalization will still be that estimated for the  $4b/3b$  sample). Since the same shape is used for the  $4b/3b/2bs$  SR predictions of  $t\bar{t}$ , the shape systematics are considered correlated in the final results and limit setting. A comparison between the  $4b/3b/2bs$  shapes for the di-large- $R$ -jet mass distributions (the final discriminant) in the SR can be found in Figure 7.2. As we can see, the shapes are compatible, with the  $4b$  having much larger statistical uncertainties. Differences between these distributions will be used as a systematic, as described in Section 8.



**Figure 6.4:** comparison between the  $2b$ ,  $3b$ , and  $4b$  shapes for the di-large- $R$ -jet mass distributions (the final discriminant) in the SR.

### 6.1.5 FITTING PROCEDURE FOR QCD AND $t\bar{t}$ NORMALIZATION

The number of  $4b/3b/2bs$  events in data observed in a given region (SB / CR / SR) can be described as:

$$N_{\text{data}}^b = \mu_{\text{qcd}}^b N_{\text{qcd}}^{xb} + \alpha_{t\bar{t}}^b N_{t\bar{t}}^{b} + N_{Z+jets}^b \quad (6.4)$$

where  $\nu_b$  is the number of  $b$ -tagged track jets required,  $x$  is 1 for  $2bs$  and 2 for  $3b$  and  $4b$ .  $\mu_{\text{multijet}}$  is essentially an estimate of the ratio of the number of QCD events with  $\nu_b$   $b$ -tagged track jets, to the number of  $1b/2b$  QCD events, while the  $t\bar{t}$  normalization parameter  $\alpha_{t\bar{t}}$  applied after the  $t\bar{t}$  is scaled to the total integrated luminosity, is a correction to the MC prediction in this phase space. The same equation can be applied to the  $4b/3b/2bs$  region (replacing  $\nu_b$  by  $4b/3b/2s$   $b$  in Equation 7.2).

In order to constrain the QCD and  $t\bar{t}$  background normalizations using data, a simultaneous fit

is applied to extract both the  $t\bar{t}$  normalization with respect to the yields from simulation and the number of  $1b/2b$  data events for the QCD background. These scaling parameters are determined independently for the  $4b/3b/2s$  signal regions. But as the procedure is the same for those three signal regions, we denote these scaling factors simply  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$  in the following text.

A binned maximum likelihood fit is employed to find the values of  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$ , as well as the correlation between the two parameters. The fit is performed on the leading- $p_T$  jet mass spectrum in the sideband region, as it has the best separation between QCD and  $t\bar{t}$  shapes. Due to the  $p_T > 450$  GeV cut imposed on the leading  $large - R_{\text{jet}}$ , the hadronic top quark is likely to be fully reconstructed inside of the  $large - R_{\text{jet}}$  and the leading jet mass in the  $t\bar{t}$  sample has a clean peak around  $M = 170$  GeV in the sideband region.

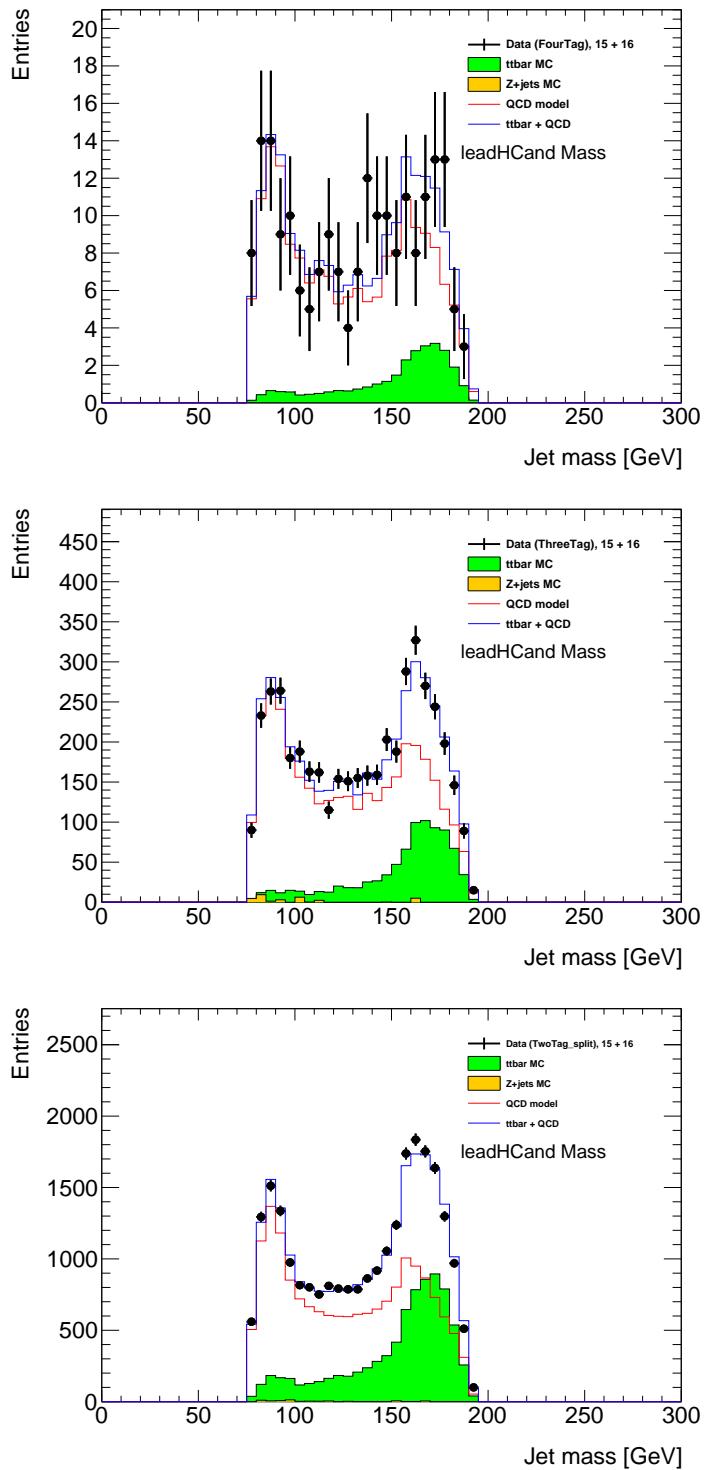
The values of  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$  as estimated by the fits in the  $4b/3b/2bs$  sideband regions can be found in Table 7.2, along with the correlation of the fitted parameters  $\xi(\mu_{qcd}, \alpha_{t\bar{t}}) = \frac{\text{Cov}(\mu_{qcd}, \alpha_{t\bar{t}})}{\sqrt{\text{Cov}(\mu_{qcd}) \text{Cov}(\alpha_{t\bar{t}})}}$ .  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$  are approximately 70% negatively correlated, which is not surprising as they are the only two components fit to the data distribution and their sum needs to predict the SB total event count.

Sample	$\mu_{qcd}$	$\alpha_{t\bar{t}}$	$\xi(\mu_{qcd}, \alpha_{t\bar{t}})$
FourTag	$0.0332 \pm 0.00428$	$0.891 \pm 0.599$	-0.785
ThreeTag	$0.163 \pm 0.00434$	$0.8 \pm 0.0733$	-0.72
TwoTag split	$0.0627 \pm 0.000573$	$0.986 \pm 0.0186$	-0.47

**Table 6.2:** Background scaling parameters ( $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$ ) estimated from fits to the leading jet mass distributions in  $4b/3b/2bs$  sideband regions.  $\xi(\mu_{qcd}, \alpha_{t\bar{t}}) = \frac{\text{Cov}(\mu_{qcd}, \alpha_{t\bar{t}})}{\sqrt{\text{Cov}(\mu_{qcd}) \text{Cov}(\alpha_{t\bar{t}})}}$

Figure 7.3 shows the post-fit spectrum of the leading  $large - R_{\text{calorimeter}}$  jet mass in the  $4b/3b/2bs$  sideband regions. The normalization of  $t\bar{t}$  is constrained by the top quark mass peak around 170 GeV. The shapes of the data is also well modeled by the predicted background. The fitting errors on

$\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$  are applied as systematic uncertainties taking into account their correlation. This will be explained in more detail in the systematics section.



**Figure 6.5:** Simultaneous fit of  $\mu_{\text{mlijet}}$  and  $\alpha_{\bar{t}t}$  in  $4b$  (top) and  $3b$  (middle) and  $2b$  (bottom) sideband region using leading large –  $R$  calorimeter jet mass spectrum.

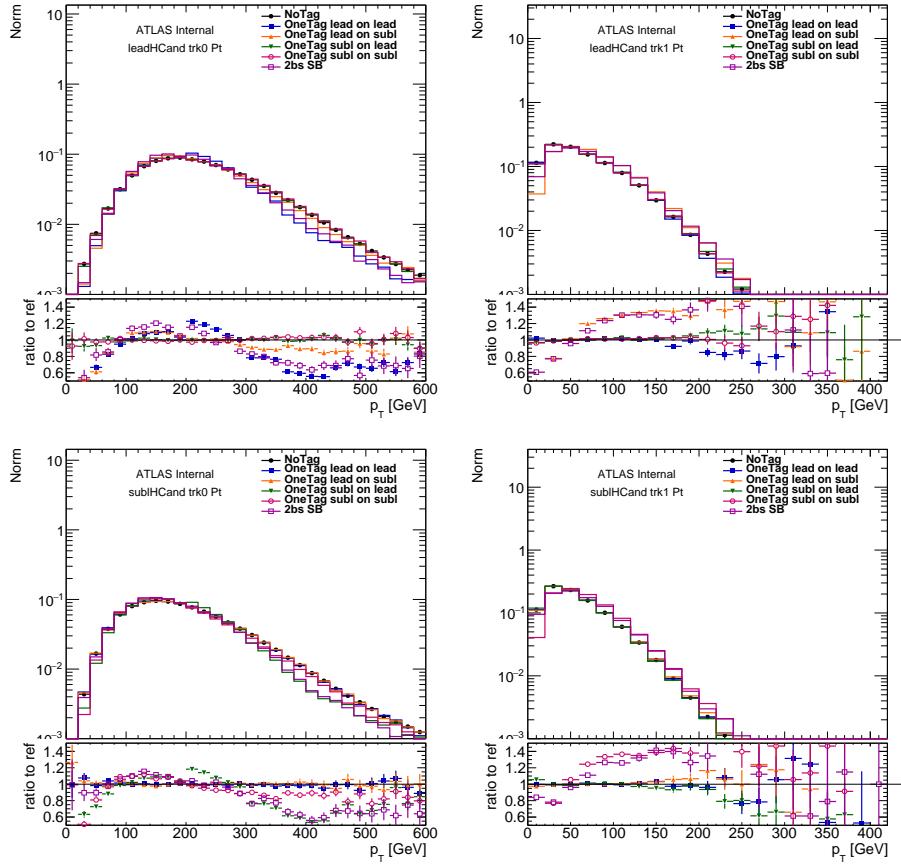
### 6.1.6 KINEMATIC REWEIGHTING

Due to the large contribution from the completely data-driven QCD background, it is important to model this background as good as possible in all regions of the analysis. Using the  $1/2b$  region to model the  $2bs$ ,  $3b$ , and  $4b$  regions can introduce discrepancies in the modeling of the estimated QCD background versus the real  $nb$  data. These discrepancies arise possibly from the non-trivial effect that  $b$ -tagging has on jet kinematics. The natural choice of reweighting variable is the  $p_T$  of the track jets in the event, since these are the objects that we apply the  $b$ -tagging to. Also, large- $R$  jet  $p_T$  is also reweighted to account for the effect from light and charm quark composition difference at different energy scales. The three chosen variables are the leading large- $R$  jet  $p_T$ , leading large- $R$  jet leading trackjet  $p_T$  and subleading large- $R$  jet leading trackjet  $p_T$ .

In order to account for the  $b$ -tagging effect, a reweighting on the  $1/2b$  data is adopted. The basic idea is to reweight the non  $b$ -tagged Higgs candidate to have kinematic distributions just like a  $b$ -tagged Higgs candidate. The idea is demonstrated in Figure 7.4. It shows that  $2bs$  has very similar kinematic distributions on the trackjet  $p_T$  as the  $1b$  sample, when the variable is the  $b$  tagged trackjet.

For  $2bs$ , the  $1b$  non-tagged Higgs candidate is reweighted to be like a  $1b$  tagged Higgs candidate; for  $3b$ , the  $2b$  non-tagged Higgs candidate is reweighted to be like a  $1b$  tagged Higgs candidate; for  $4b$ , the  $2b$  non-tagged Higgs candidate is reweighted to be like a  $2b$  tagged Higgs candidate. For each category, the events are split into two orthogonal subgroups, depending on whether leading/subleading Higgs candidate is  $b$ -tagged, the event is then reweighted such that the untagged Higgs candidate's distribution behaves like the corresponding  $b$ -tagged Higgs candidate's.

To avoid potential biases in the final distributions used for the analysis, a reweighting technique is applied to the  $1/2b$  data only. Since each signal region is modeled by a different  $1/2b$  tag category:  $2bs$  by  $1b$  tag events with at least 1 track jets on both large- $R$  jets,  $3b$  by  $2b$  tag events with at least one track jets on one large- $R$  jet and at least two track jets on the other large- $R$  jet, and  $4b$  by  $2b$  tag events



**Figure 6.6:** Comparison of different trackjet  $p_T$  distributions. Top row is for leading  $p_T$  Higgs candidate, and bottom row is for subleading  $p_T$  Higgs candidate. Left column is for the leading  $p_T$  trackjet of the Higgs candidate, and right column is for the subleading  $p_T$  trackjet of the Higgs candidate. Shown in the plot are just data distributions, inclusive of SB, CR, and SR regions for  $0b$  and  $1b$ , while for  $2bs$  only the SB region is shown.  $1b$  sample is further split into four subcategories, depending on which trackjet gets  $b$  tagged. OneTag lead on lead means the  $b$  tagged trackjet is the leading trackjet of the leading Higgs candidate, OneTag lead on subl means the  $b$  tagged trackjet is the subleading trackjet of the leading Higgs candidate, OneTag subl on lead means the  $b$  tagged trackjet is the leading trackjet of the subleading Higgs candidate, and OneTag subl on subl means the  $b$  tagged trackjet is the subleading trackjet of the subleading Higgs candidate. At the bottom ratio plot, all the ratio are taken with respect to the  $0b$  tagged distribution.

with at least two track jets on both large- $R$  jets, the reweighting procedure is the same but orthogonal for the three different channels. Note the  $2b$  sample is already split into separate parts, as described in paragraph 7.1.3.

The detailed procedure is listed as follows:

- Subtracting  $1/2b$  tag  $t\bar{t}$  and  $Z+jets$  samples in the sideband from the  $1/2b$  tag data in the Sideband + Control + Signal regions to get the  $1/2b$  QCD inclusive estimate.
- Separate the  $1/2b$  tag sample further to sample A. that has the  $b$ -tagged Higgs is the leading  $p_T$ Higgs candidate, and B. that the  $b$ -tagged Higgs is the subleading  $p_T$ Higgs candidate.
- For each variable, i.e. the large- $R$  jet  $p_T$ , normalize sample A to sample B total number of events, take the ratio of sample A distribution over sample B distribution, and fit the ratio with a spline function. (TSpline3)
- Use this functional form to extract reweighting values for each variable that is considered. The reweighting value for each variable is also constrained to be within a  $-30\%$  to  $+40\%$  range compared to one, to avoid over corrections and failed fit situations.
- For each event, all the weights are multiplied together to change the  $1/2b$  tag data event weight. Another constraint is applied, such that each total reweighting value is constrained to be within a  $10\%$  to  $+1000\%$  range compared to one, again to avoid over corrections.
- The reweighting is done on the three variables: large- $R$  jet  $p_T$  and the two track jet  $p_{T\gamma}$ s, which is counted as one iteration of reweighting.
- A total of ten iterations are used to stabilize the reweighting. The reweighting is roughly converging after three iterations.

For reweighting method comparisons and validations in data and Dijet MC, see Appendix ??.

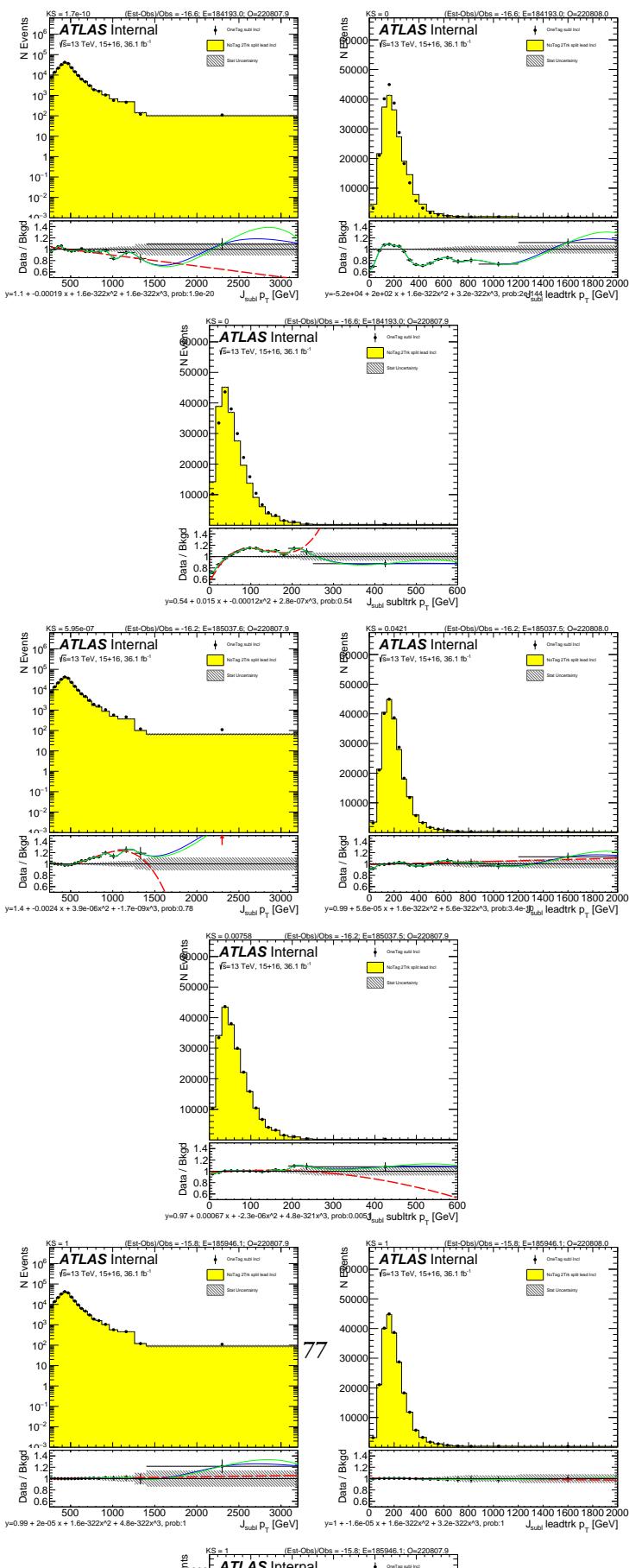
## REWEIGHTING FITS

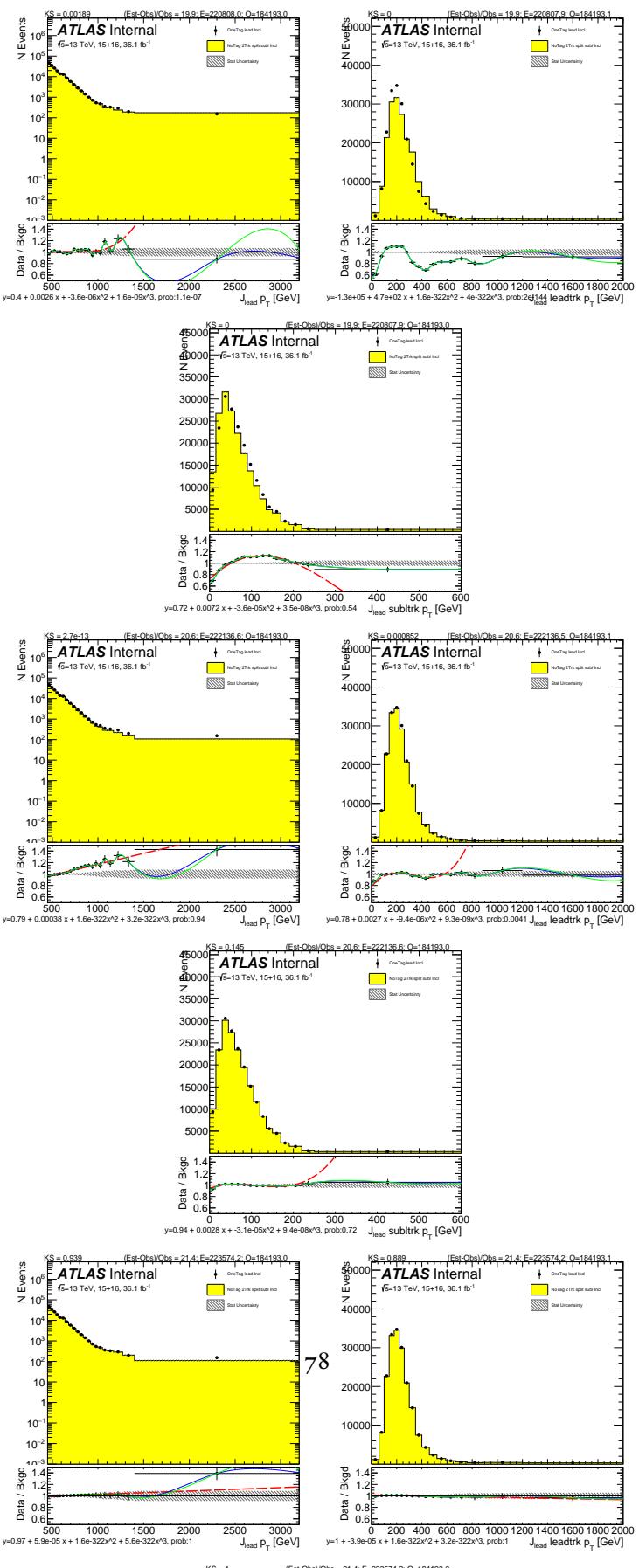
The first iteration, second iteration, and last iteration of fits for  $2b$ s, where in  $1b$  data, the non-tagged Higgs candidate are reweighted to be like a  $1b$  tagged Higgs candidate, can be seen in Figure 7.5 and 7.6. Similar distributions for  $3b$ , where in  $2b$  data, the non-tagged Higgs candidate are reweighted to be like a  $1b$  tagged Higgs candidate, are shown in Figure 7.7 and 7.8. Similar distributions for  $4b$ , where in  $2b$  data, the non-tagged Higgs candidate are reweighted to be like a  $2b$  tagged Higgs

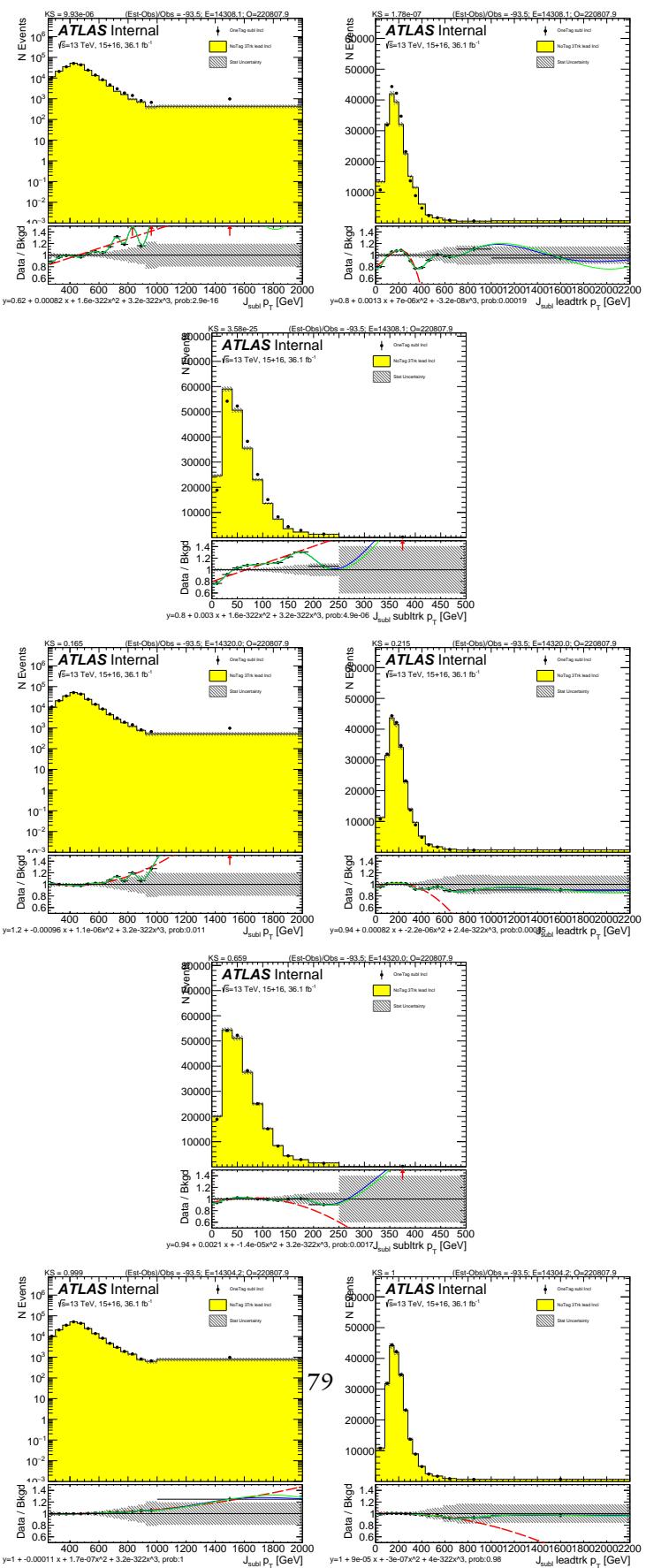
candidate, are shown in Figure 7.9 and 7.10. The before reweighting distribution (first row), the reweighting result after the first iteration (second row), and the final distribution after reweighting (last row) are presented.

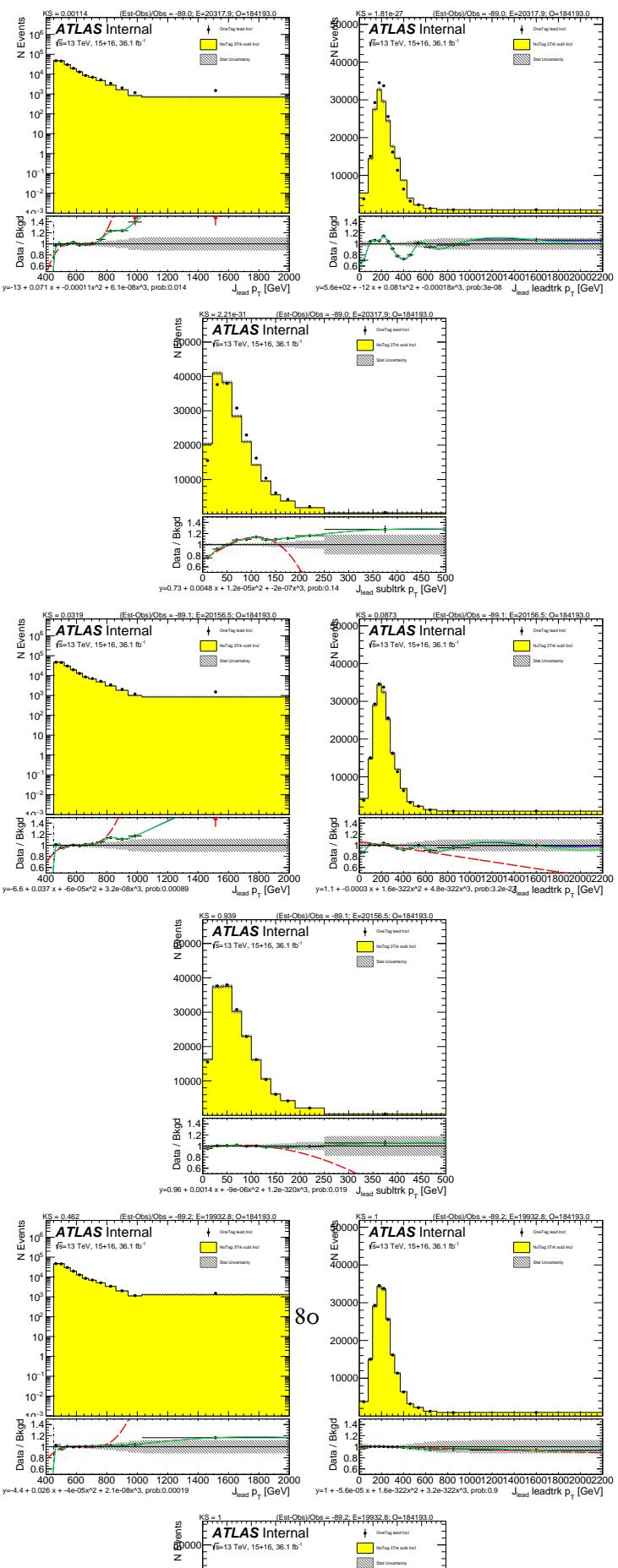
It should be noted that in some plots, like Figure 7.9 and 7.10, the last ratio bin sometimes still doesn't converge to unity. This is a feature from the limited statistics from the last bin, especially in the  $4b$  case, where only 20% number of events in  $2b$  is used for background prediction and therefore reweighted. One could choose a different binning and use more iterations to help this converge to one, yet the last bin's few event will also likely to end up with a large unphysical weight and therefore harm the background prediction later.

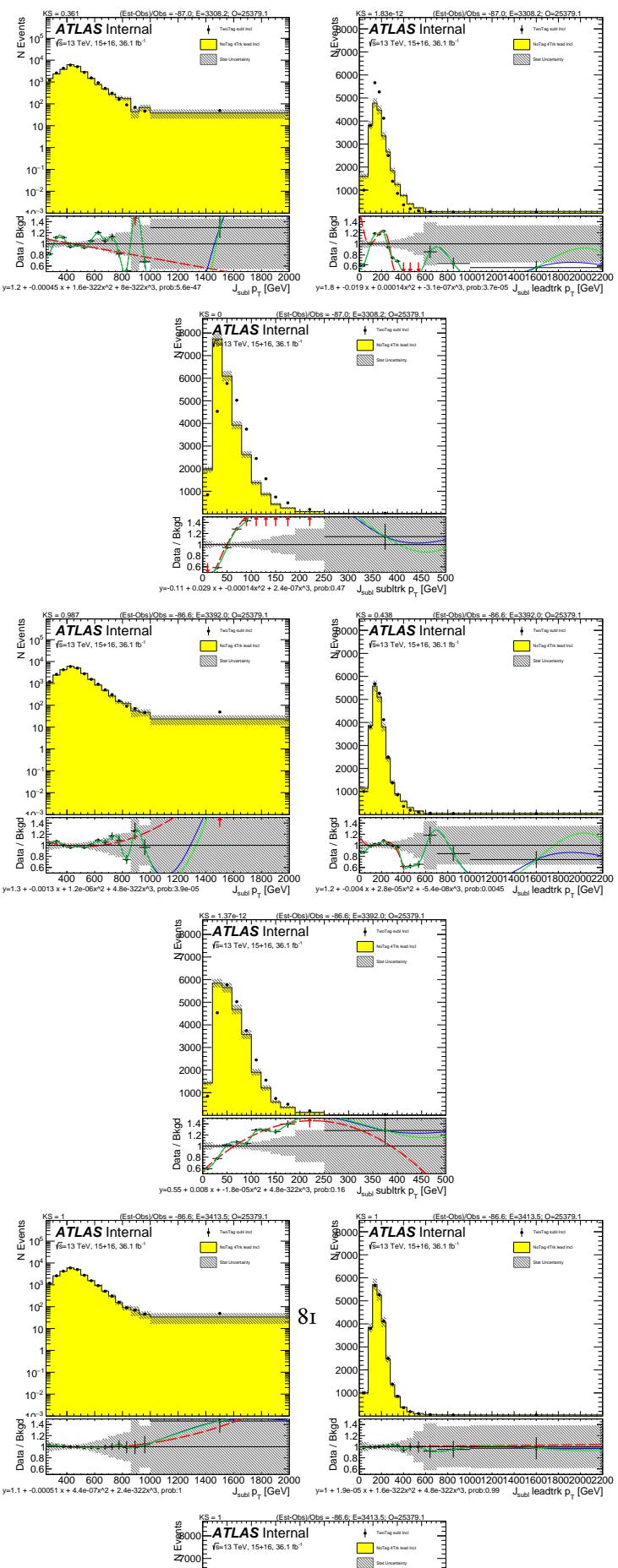
For the distribution of weights and the weight as a function of different kinematic ranges, see Appendix ??.

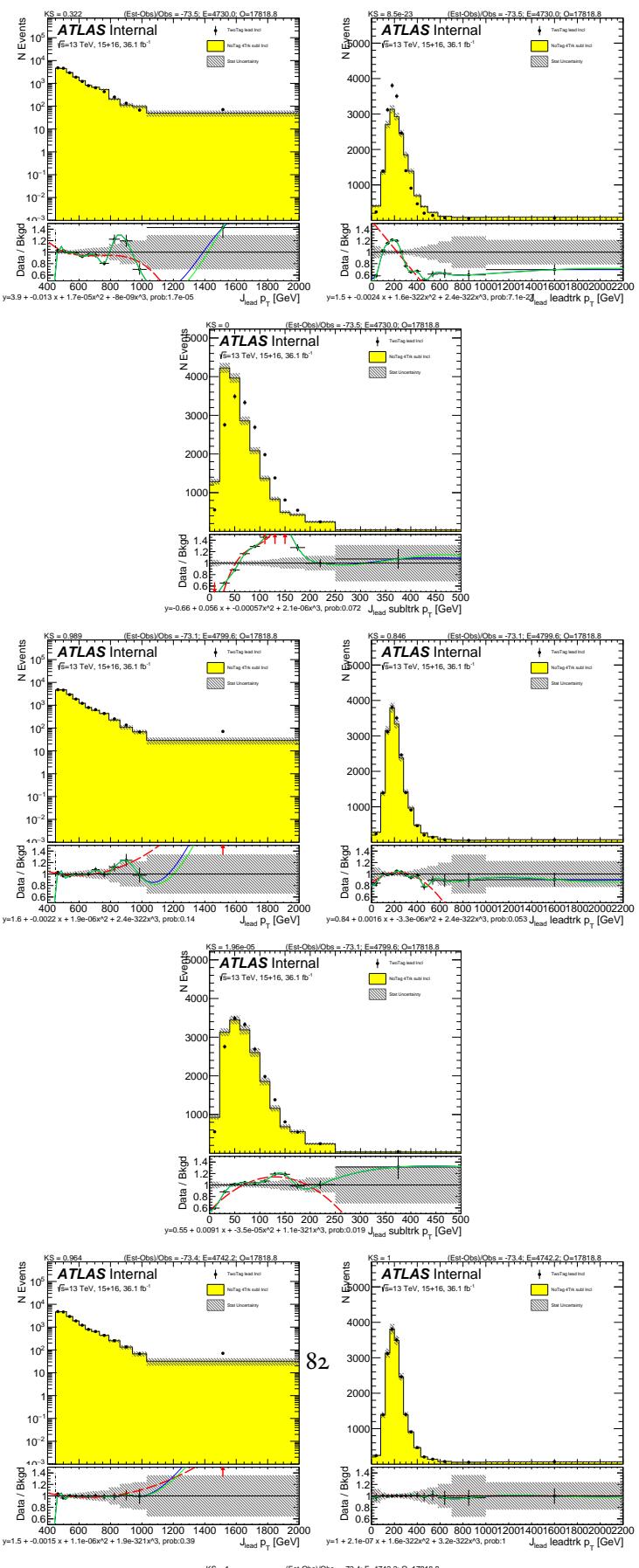






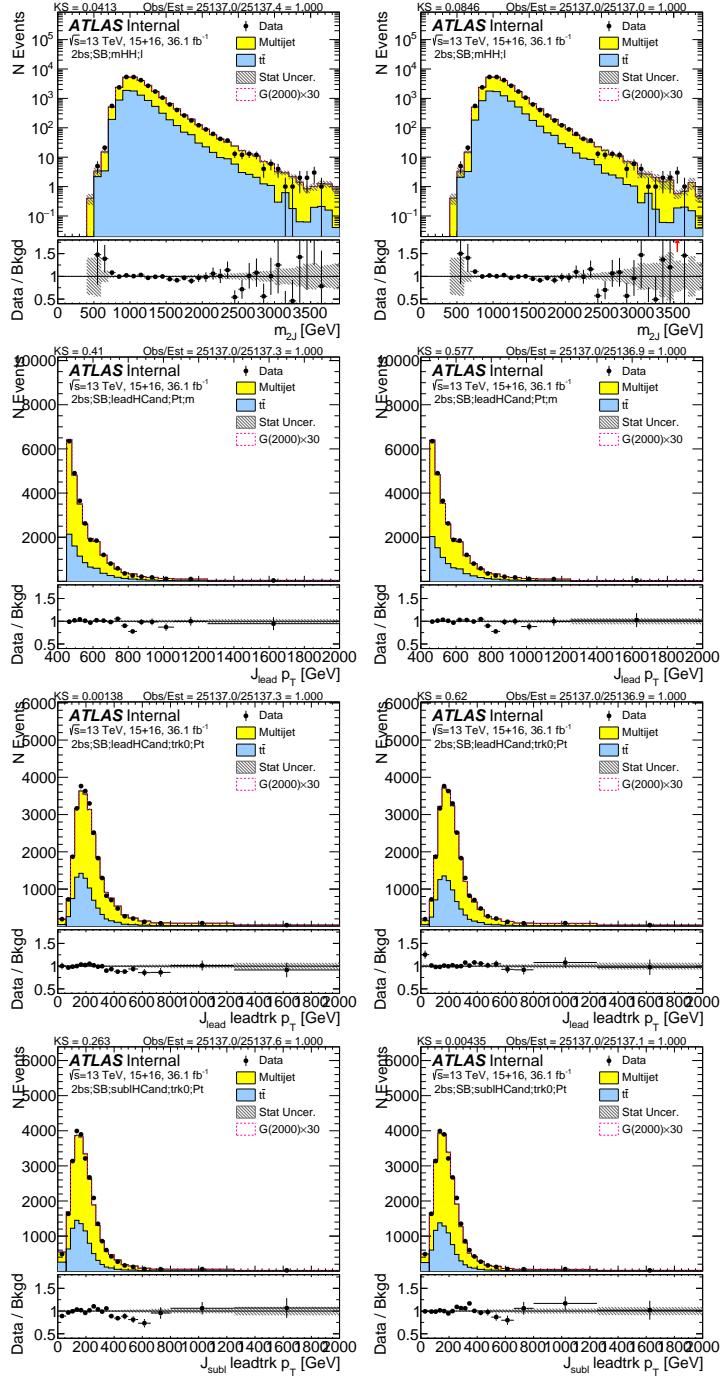




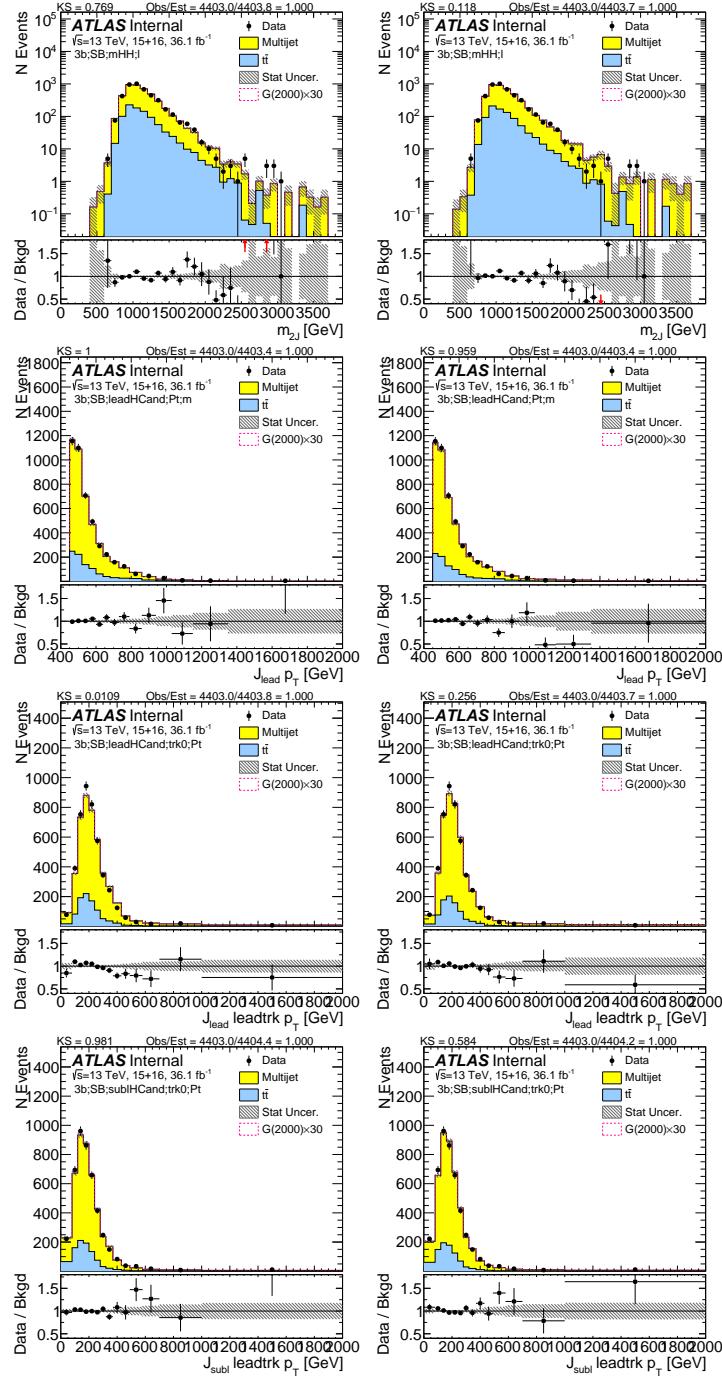


## REWEIGHTING RESULTS COMPARISON IN SIDEBAND AND CONTROL REGION

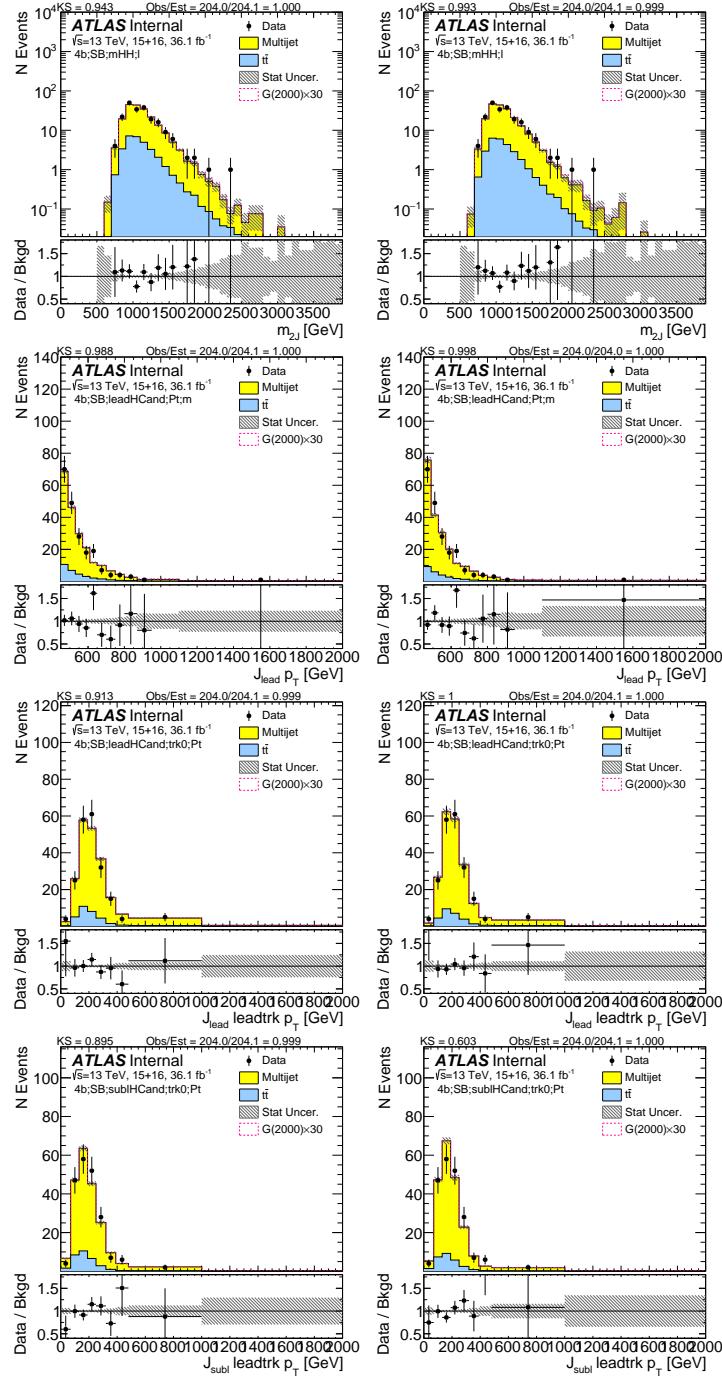
A comparison of the Sideband shapes before and after reweighting for  $2bs$ ,  $3b$  and  $4b$  can be seen in Figures 7.11, 7.12, and 7.13. Also, a comparison of the Control Region shapes before and after reweighting for  $2bs$ ,  $3b$  and  $4b$  can be seen in Figures 7.14, 7.15, and 7.16. In almost all cases, both the reweighted/non-reweighted prediction agrees fairly well with the data, and the reweighted plots' KS score improved from non-reweighted distributions.



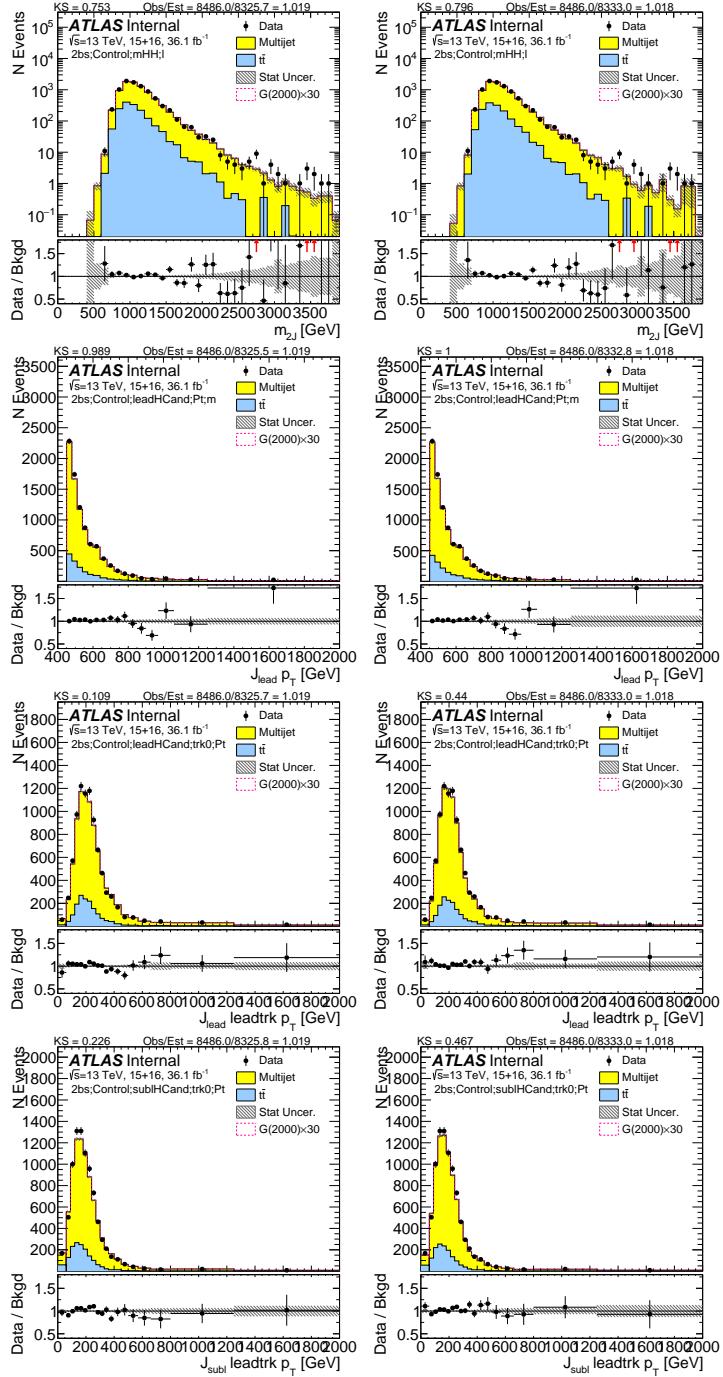
**Figure 6.13:** Reweighted 2bs Sideband region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



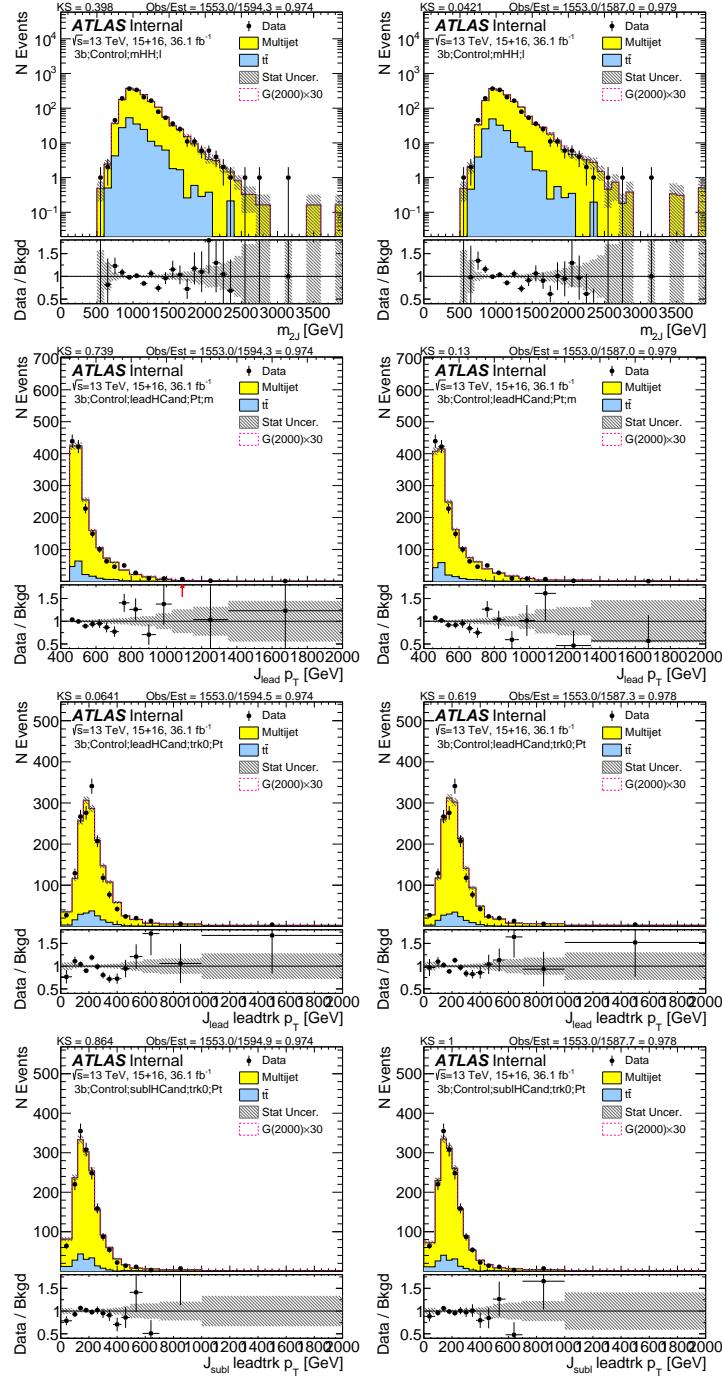
**Figure 6.14:** Reweighted 3 $b$  Sideband region predictions comparison. Top row is the dijet Mass, second row is lead large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



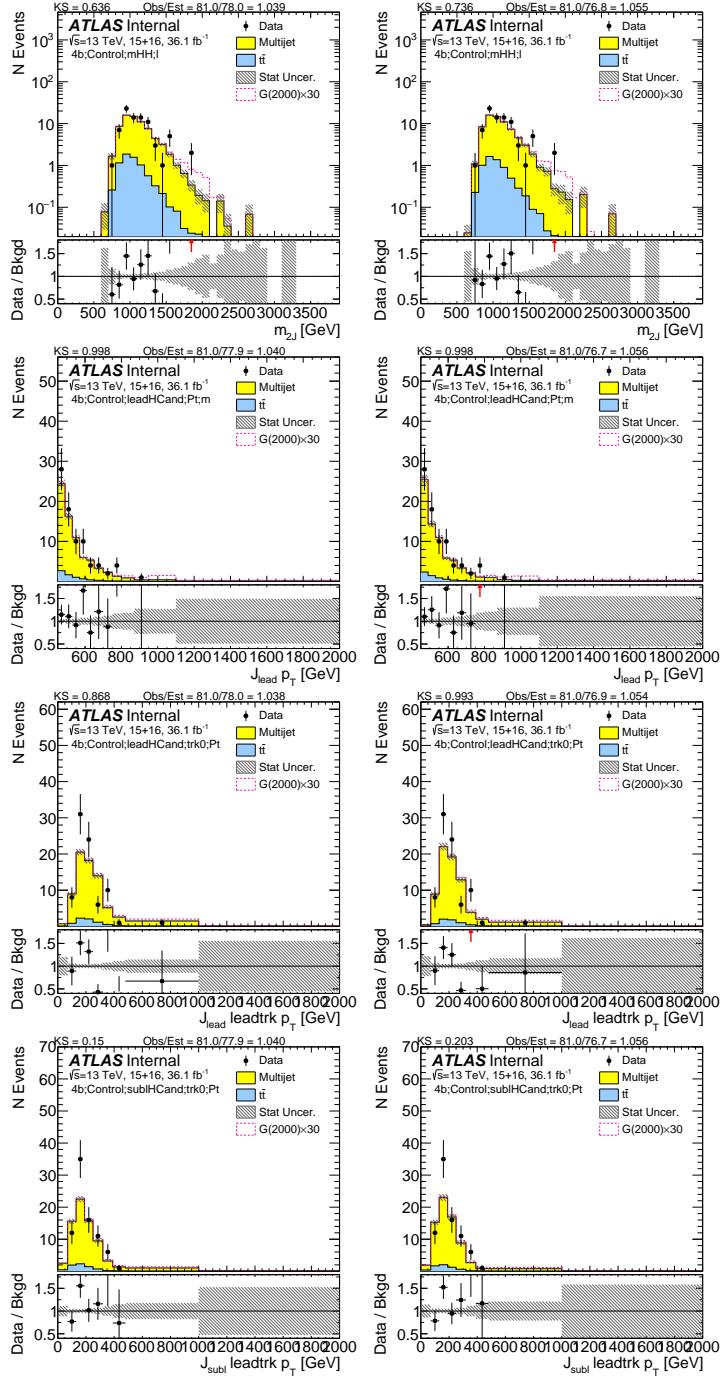
**Figure 6.15:** Reweighted  $4b$  Sideband region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



**Figure 6.16:** Reweighted  $2bs$  Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



**Figure 6.17:** Reweighted  $3b$  Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



**Figure 6.18:** Reweighted 4*b* Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.

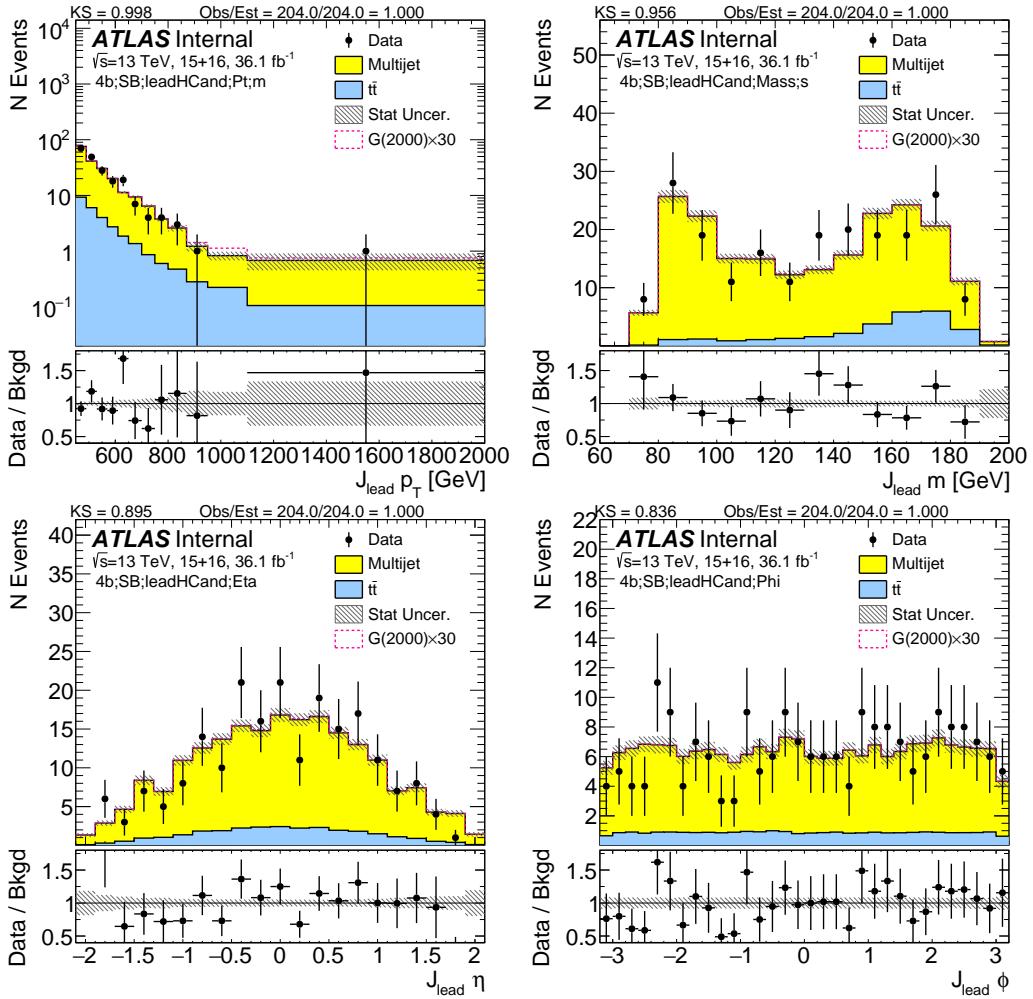
### 6.1.7 PREDICTIONS IN THE SIDEband REGION (SB)

This section shows comparisons of data with the prediction of QCD multi-jets and  $t\bar{t}$  in the sideband region (SB), which is identical to the signal region (SR) except the large- $R$  jets are required to have masses far from the Higgs mass. The definition of the sideband and control regions is discussed in Section 7.1.2. The predicted and observed event yields are summarized in Tables ?? and ??.

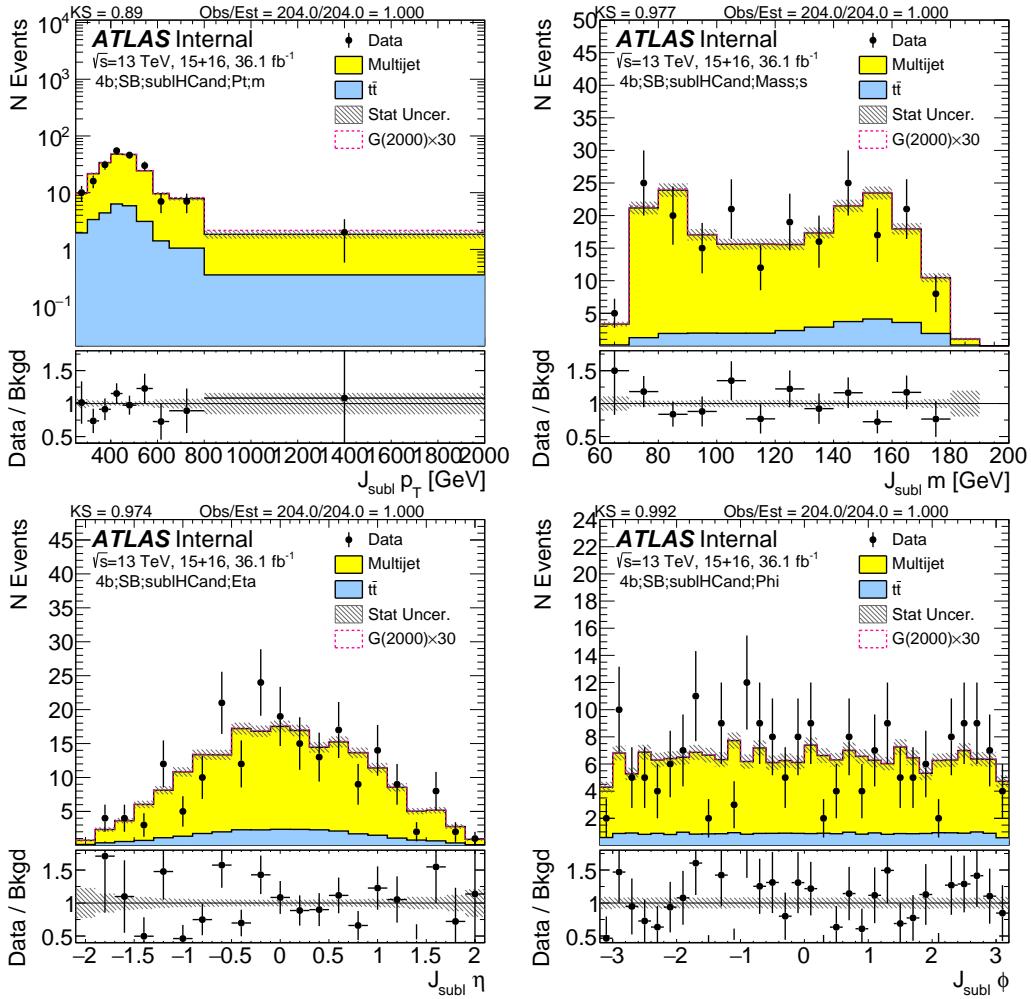
Figures 7.17, 7.18, 7.19, and 7.20 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $4b$  selection. The predicted normalization agrees perfectly by construction, but the shapes are a feature of the prediction. The quality of the prediction is generally good, and no clear systematic biases are observed.

Figures 7.21, 7.22, 7.23, and 7.24 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $3b$  selection. The predicted normalization agrees perfectly by construction, but the shapes are a feature of the prediction. The quality of the prediction is generally good, and no clear systematic biases are observed.

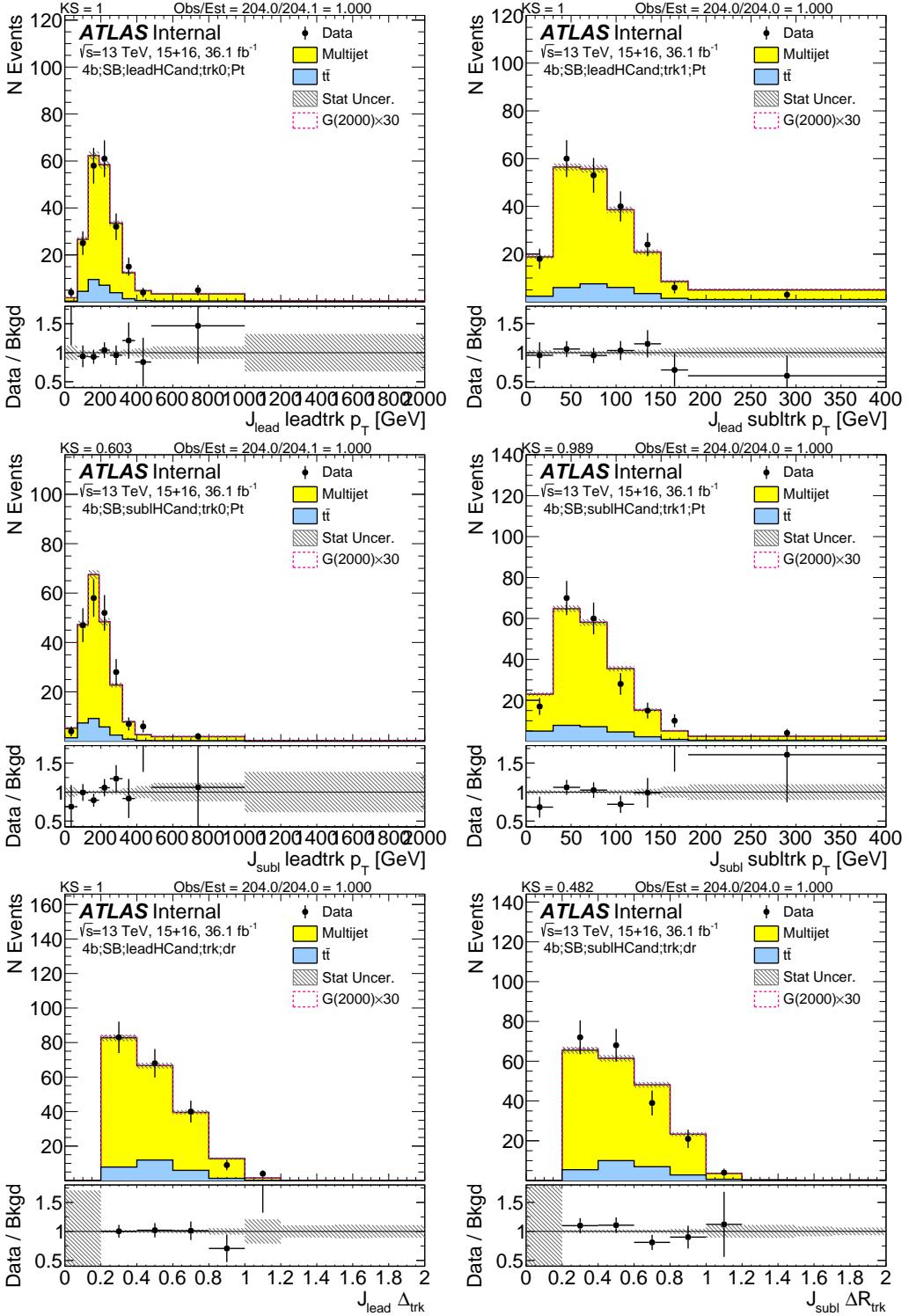
Figures 7.25, 7.26, 7.27, and 7.28 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $2bs$  selection. The predicted normalization agrees perfectly by construction, but the shapes are a feature of the prediction. The quality of the prediction is generally good, and no clear systematic biases are observed.



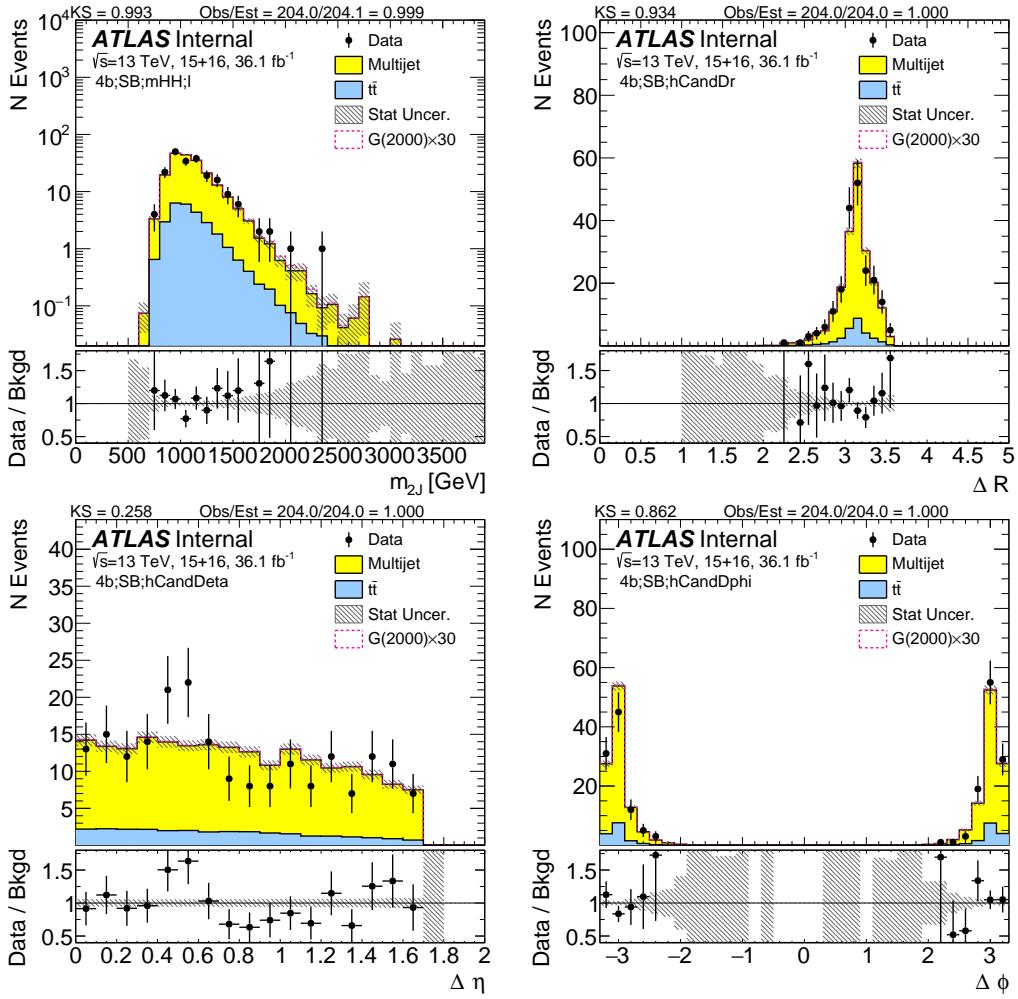
**Figure 6.19:** Kinematics of the lead large- $R$  jet in data and prediction in the sideband region after requiring 4  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



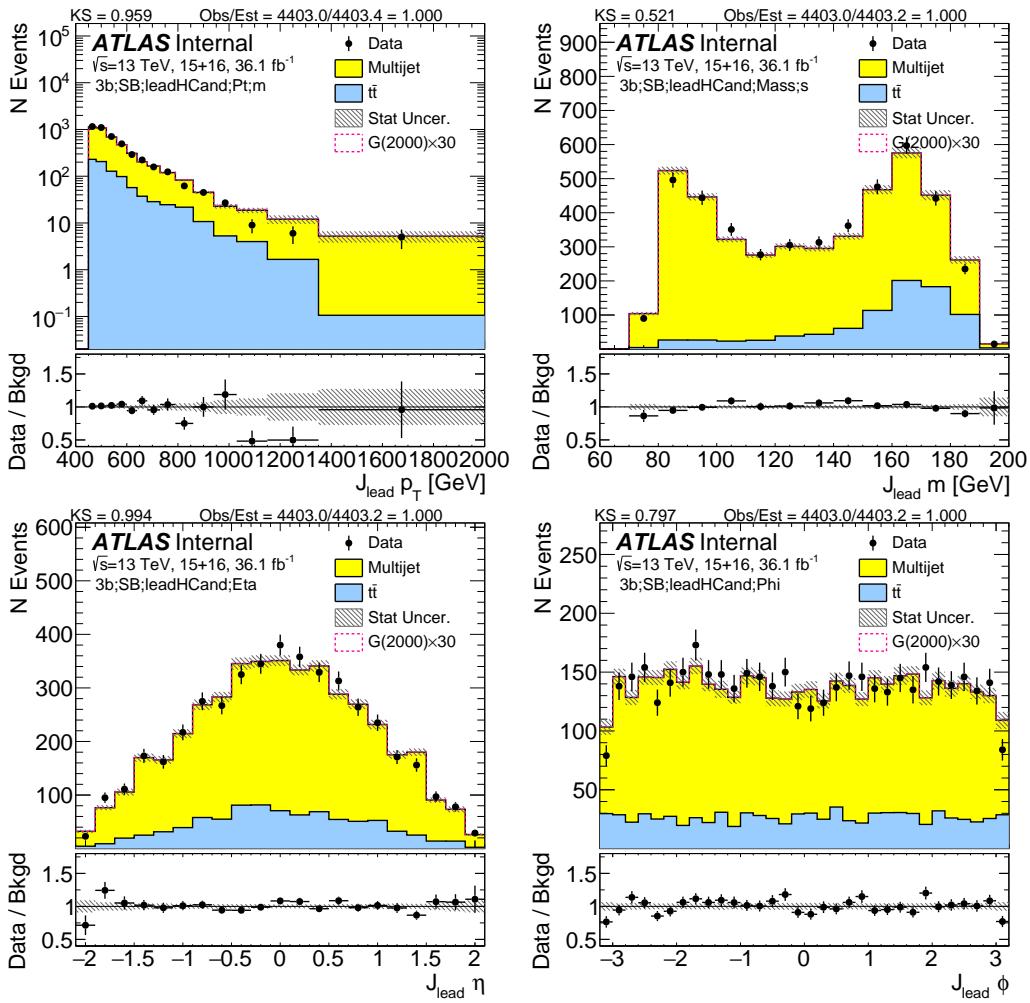
**Figure 6.20:** Kinematics of the sub-lead large- $R$  jet in data and prediction in the sideband region after requiring 4  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



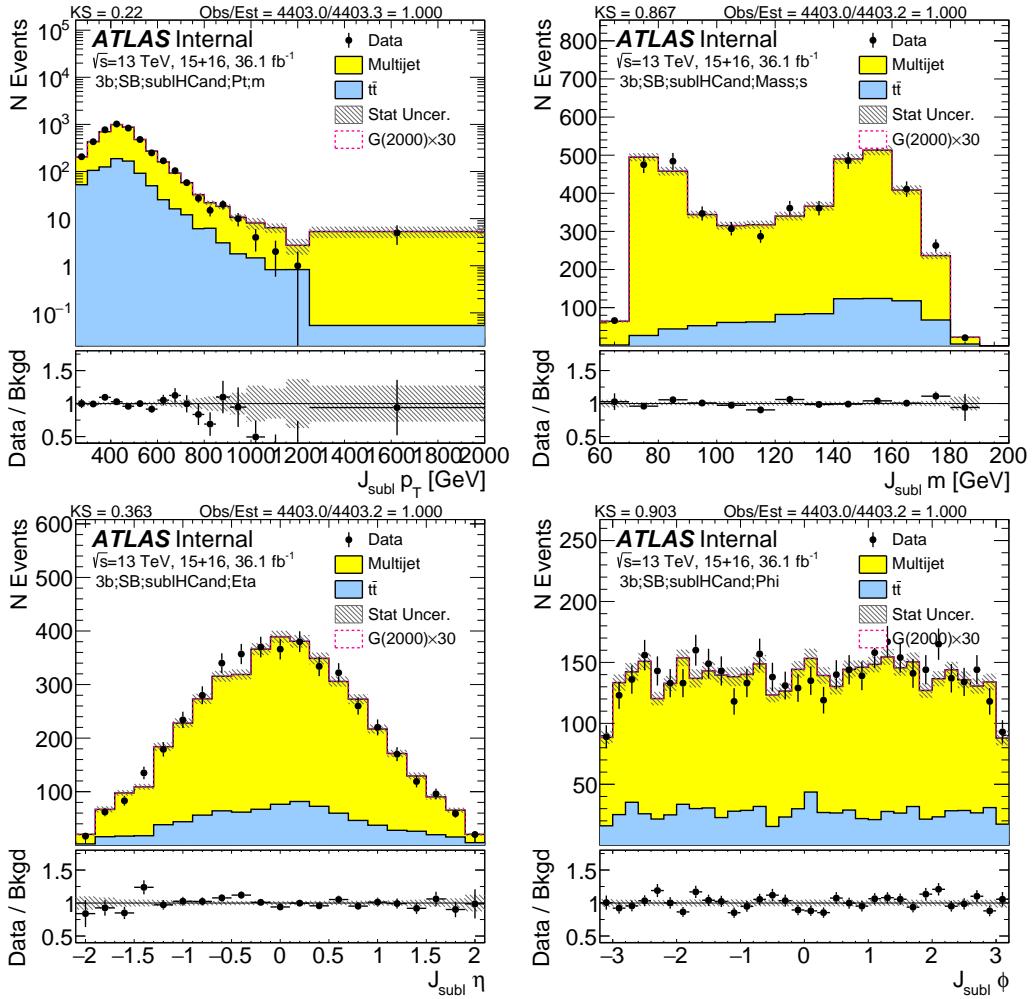
**Figure 6.21:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the sideband region after requiring 4  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. The normalization agrees by construction, and the shapes are a feature of the prediction.



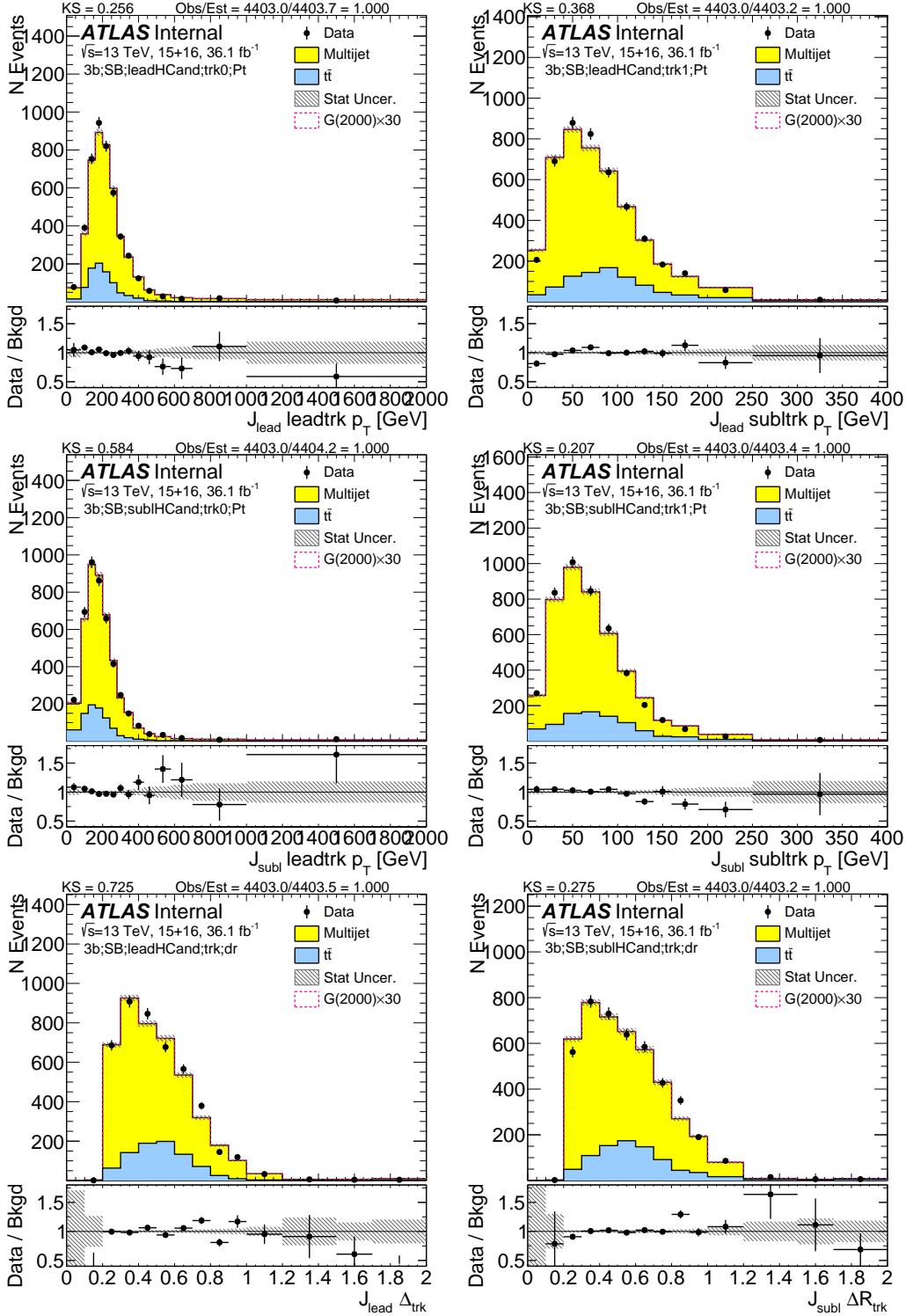
**Figure 6.22:** Kinematics of the large- $R$  jet system in data and prediction in the sideband region after requiring 4  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



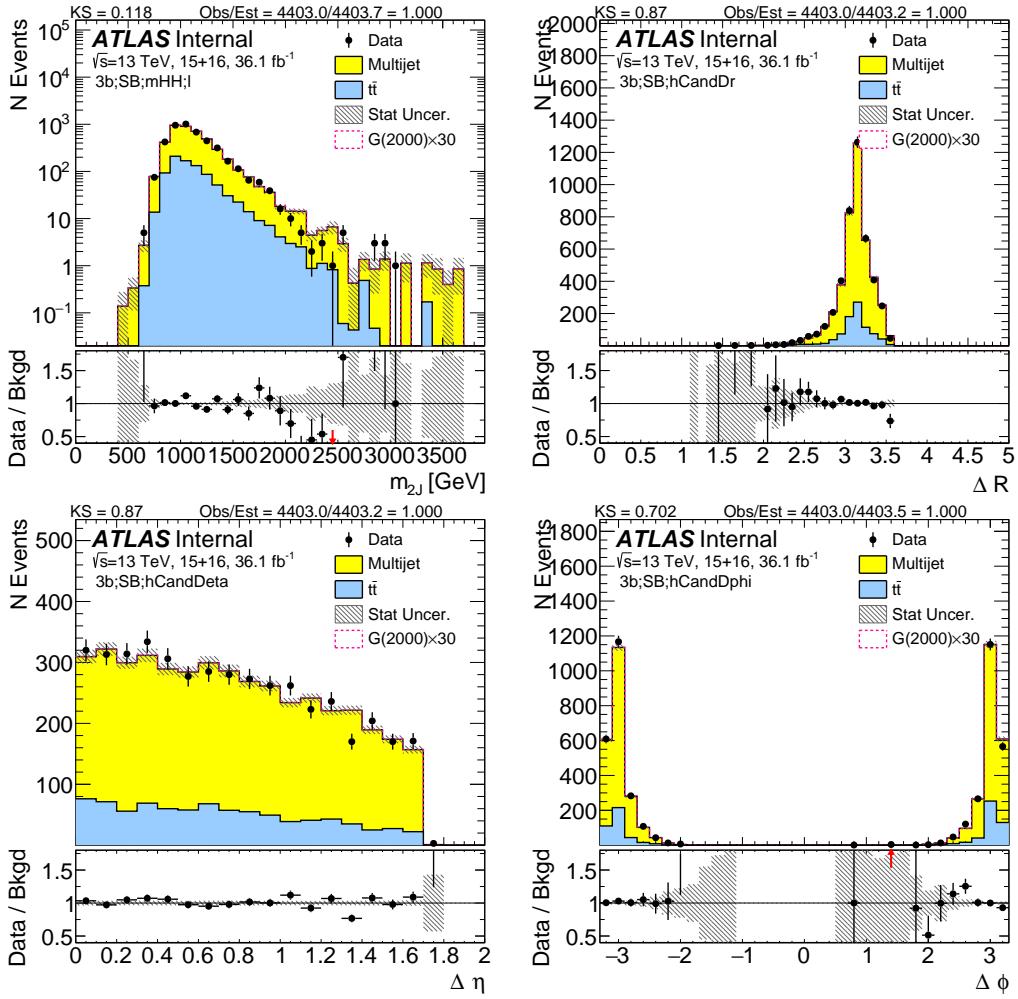
**Figure 6.23:** Kinematics of the lead large- $R$  jet in data and prediction in the sideband region after requiring 3  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



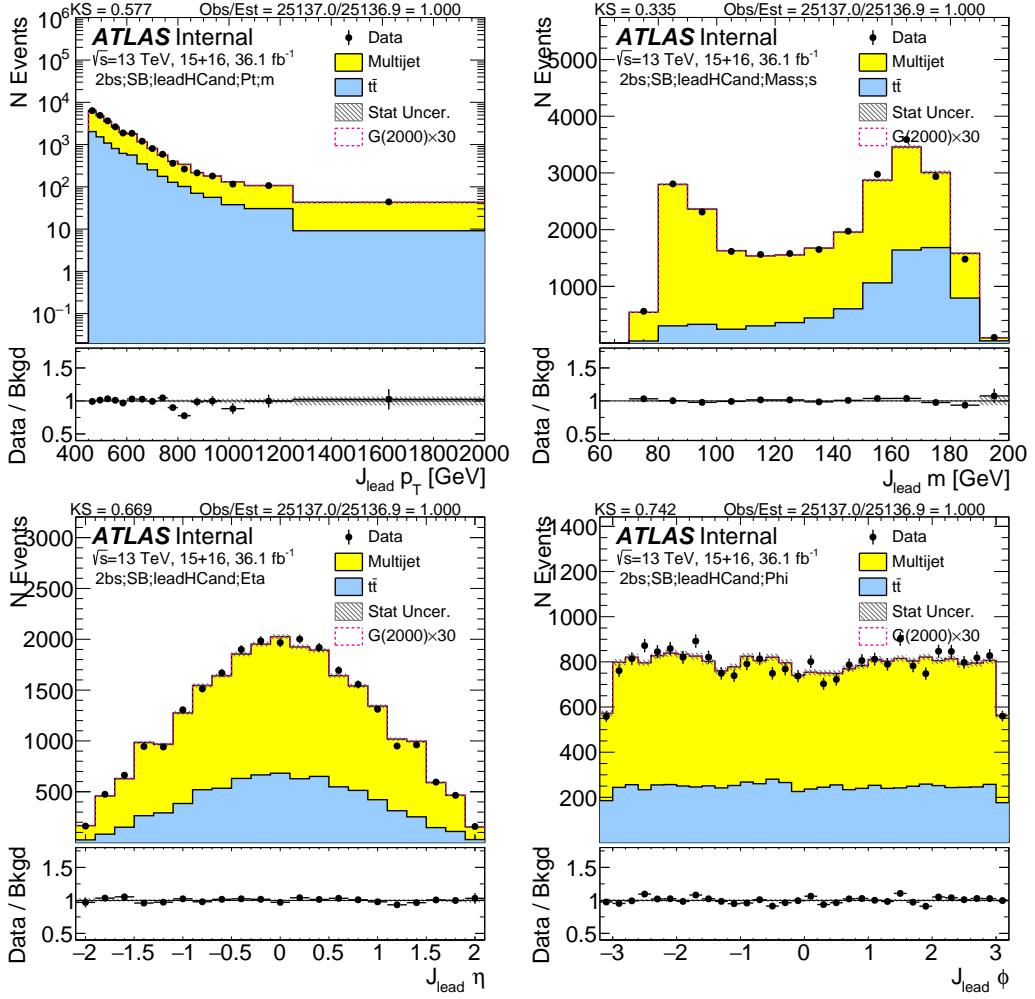
**Figure 6.24:** Kinematics of the sub-lead large- $R$  jet in data and prediction in the sideband region after requiring 3  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



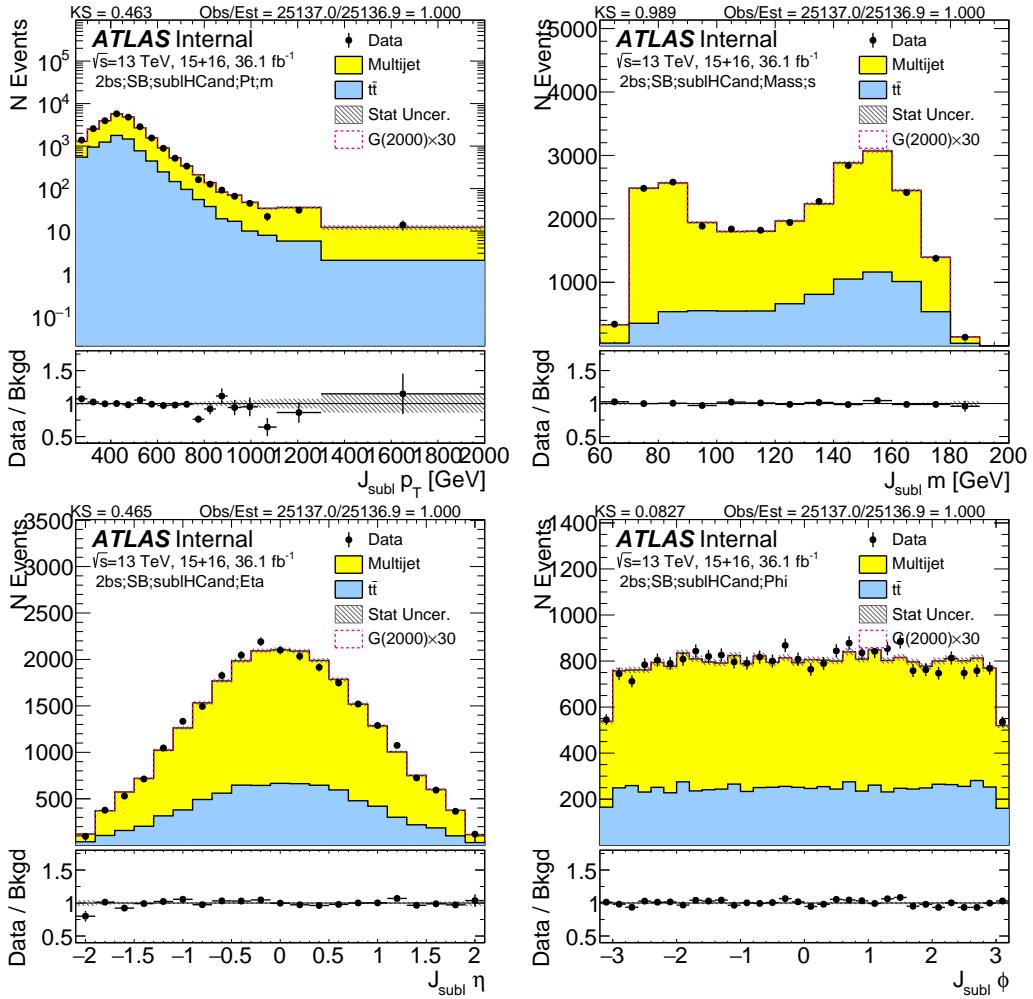
**Figure 6.25:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the sideband region after requiring 3  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. The normalization agrees by construction, and the shapes are a feature of the prediction.



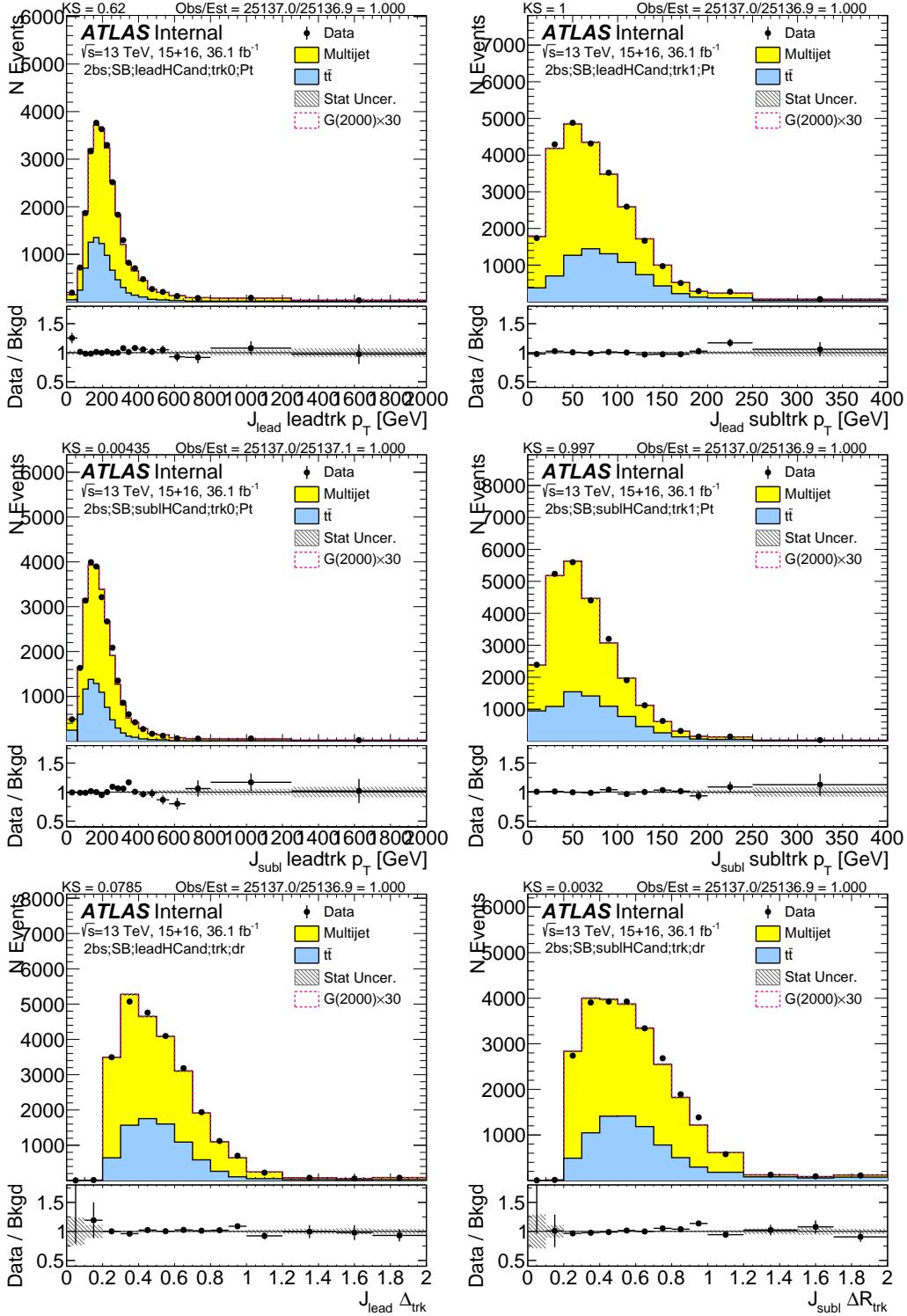
**Figure 6.26:** Kinematics of the large- $R$  jet system in data and prediction in the sideband region after requiring 3  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



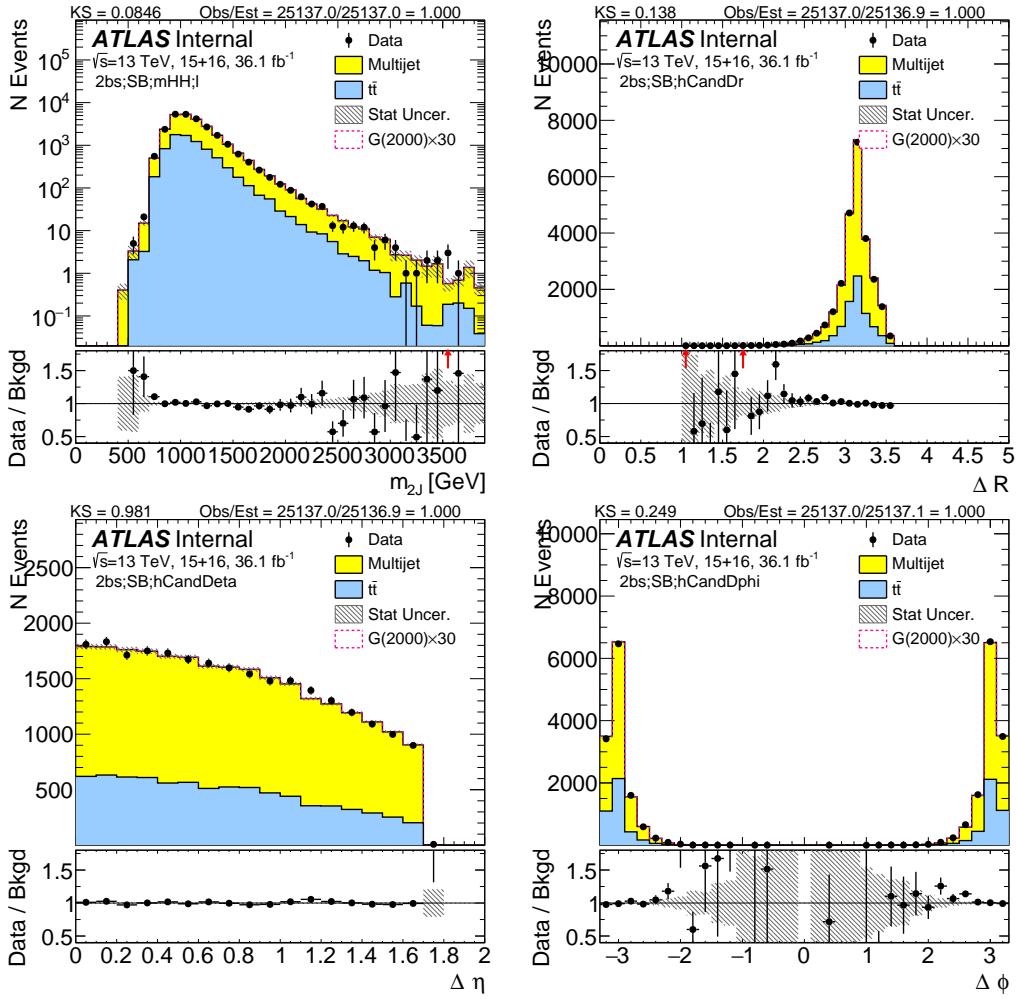
**Figure 6.27:** Kinematics of the lead large- $R$  jet in data and prediction in the sideband region after requiring 2  $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction.



**Figure 6.28:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the sideband region after requiring 2  $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction.



**Figure 6.29:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the sideband region after requiring 2  $b$ -tags split. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. The normalization agrees by construction, and the shapes are a feature of the prediction.



**Figure 6.30:** Kinematics of the large- $R$  jet system in data and prediction in the sideband region after requiring 2  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.

### 6.1.8 PREDICTIONS IN THE CONTROL REGION (CR)

This section shows comparisons of data with the prediction of QCD multi-jets and  $t\bar{t}$  in the control region (CR), which is identical to the signal region (SR) except the large- $R$  jets are required to have masses close but not too close to the Higgs mass. The definition can be seen in Section 7.1.2. The predicted and observed event yields are summarized in Tables ?? and ??.

Figures 7.29, 7.30, 7.31, and 7.32 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $4b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB. The quality of the prediction is generally good, and no clear systematic biases are observed.

Figures 7.33, 7.34, 7.35, and 7.36 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $3b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB. The quality of the prediction is generally good, and no clear systematic biases are observed.

Figures 7.37, 7.38, 7.39, and 7.40 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $2bs$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB. The quality of the prediction is generally good, and no clear systematic biases are observed.

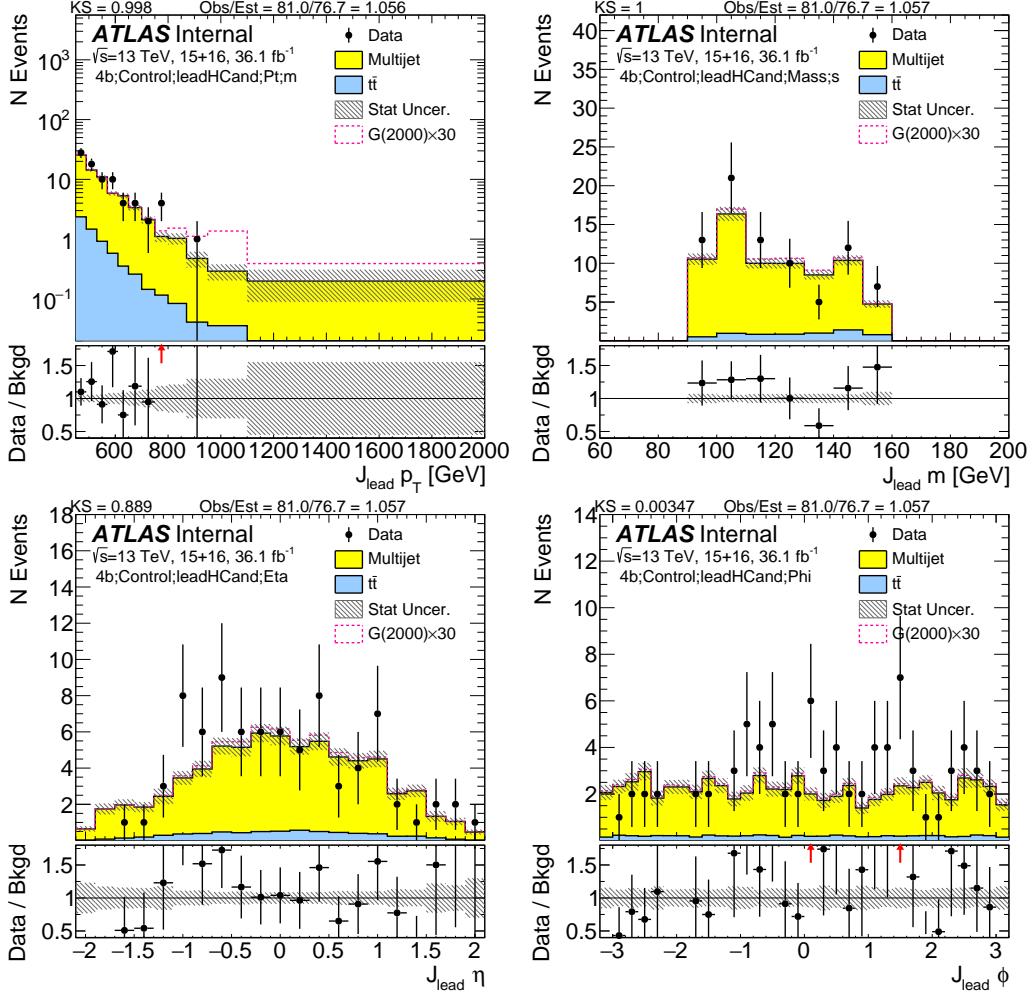


Figure 6.31: Kinematics of the lead large- $R$  jet in data and prediction in the control region after requiring 4  $b$ -tags.

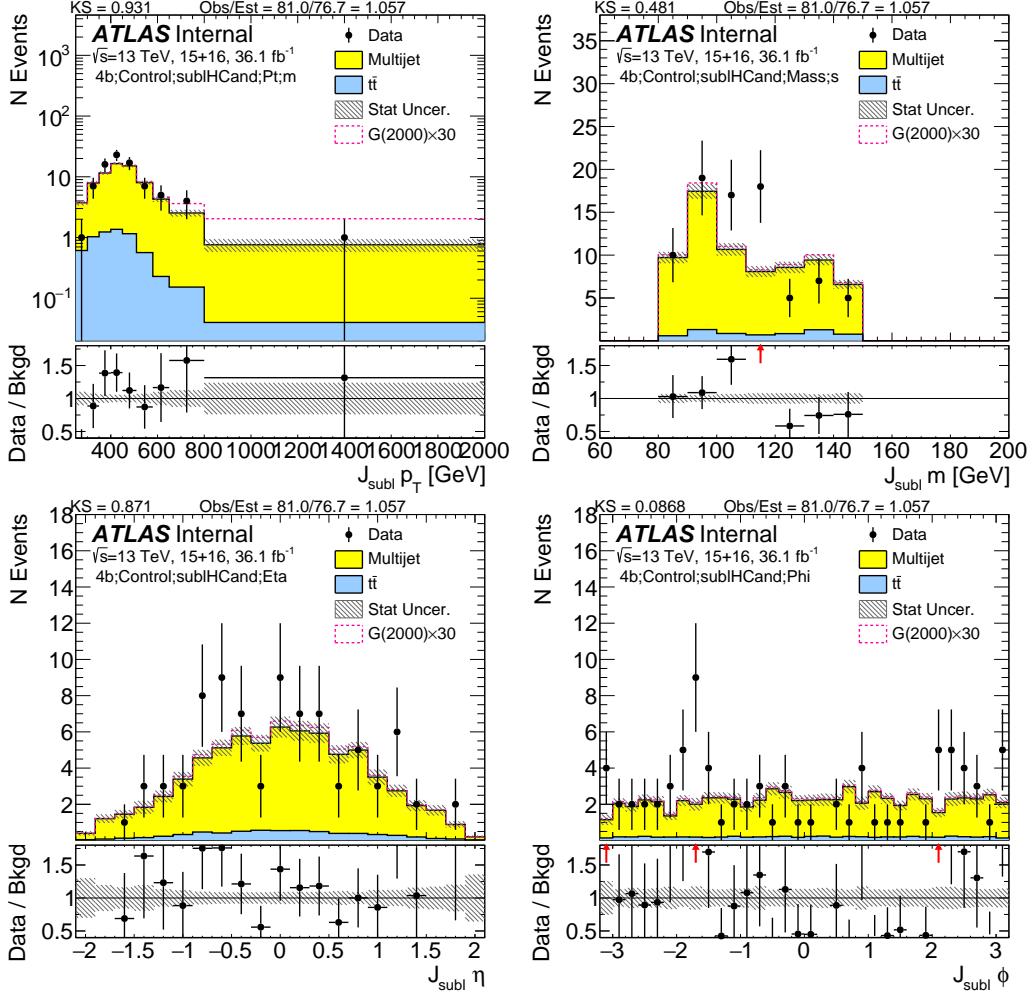
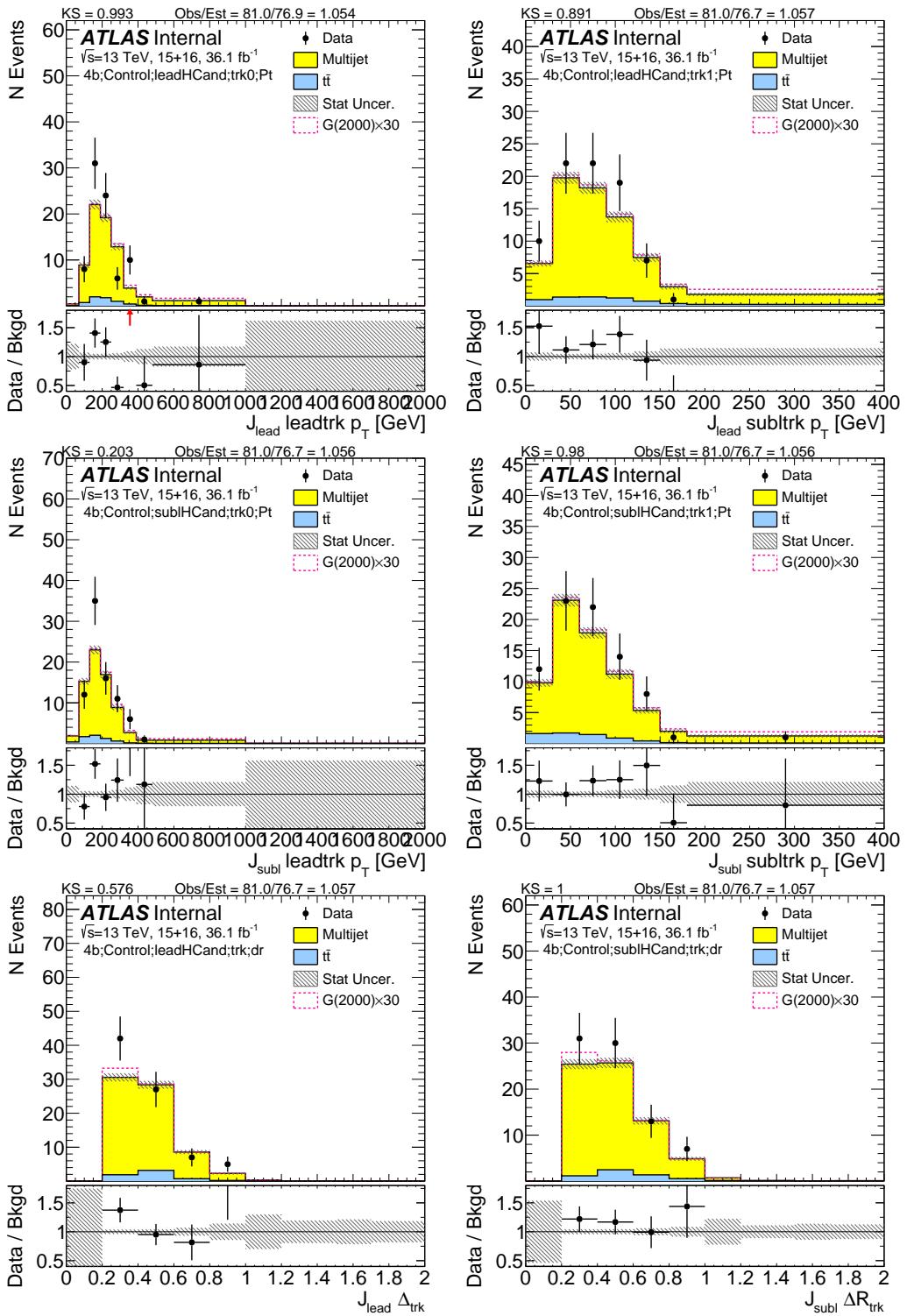


Figure 6.32: Kinematics of the sub-lead large- $R$  jet in data and prediction in the control region after requiring 4  $b$ -tags.



**Figure 6.33:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the control region after requiring 4  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.

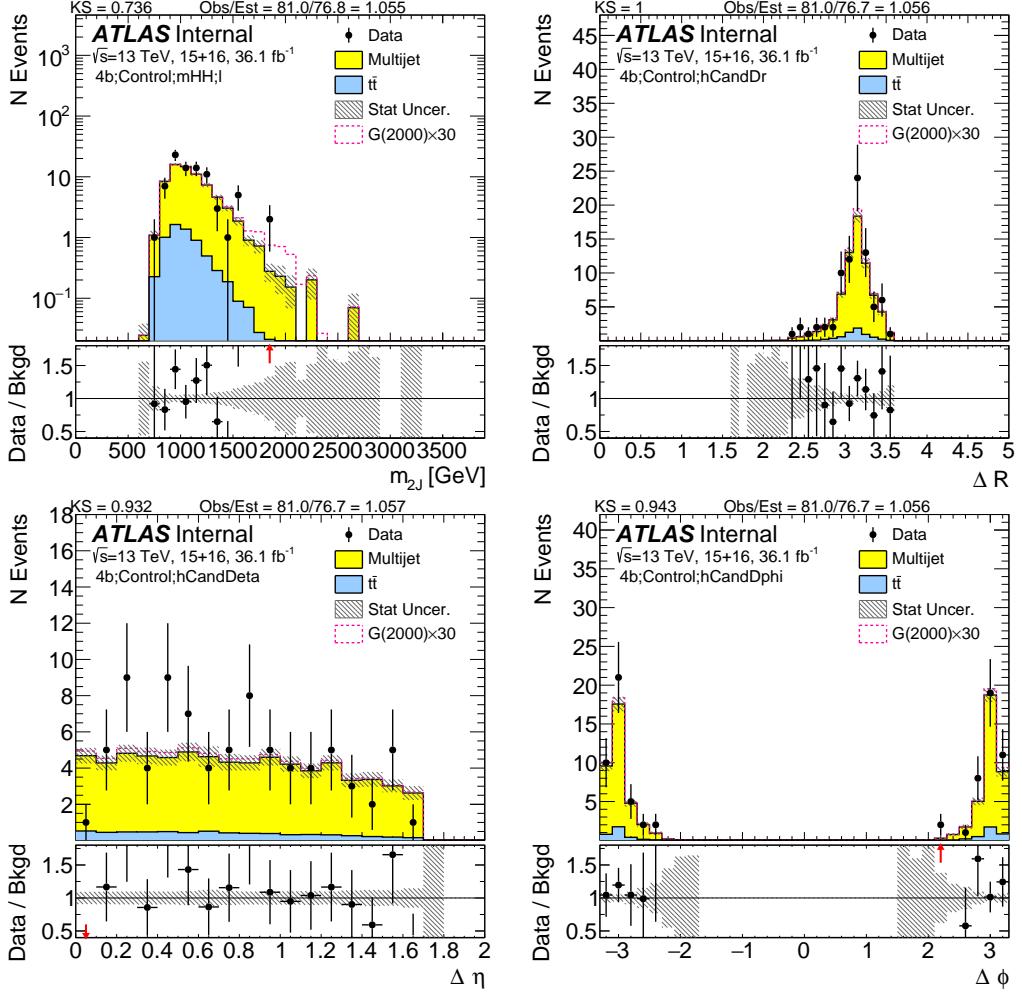
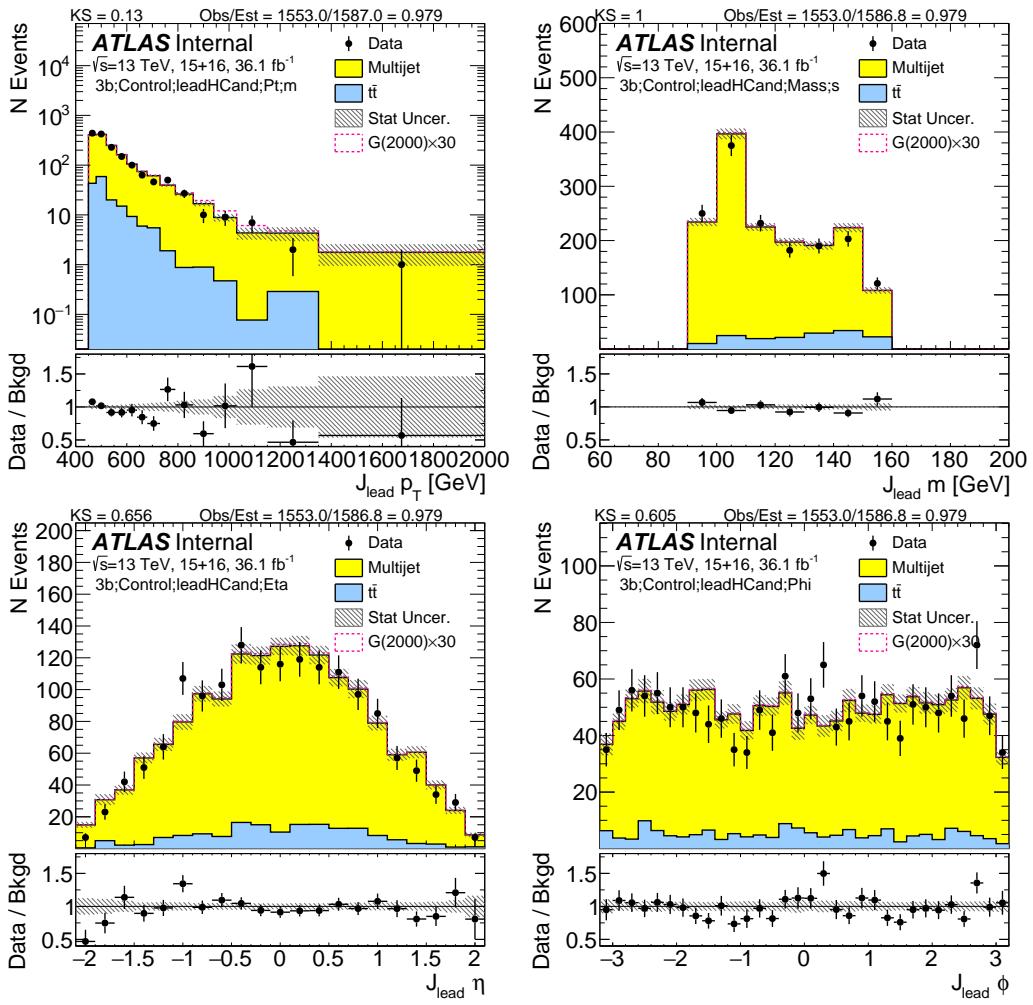


Figure 6.34: Kinematics of the large- $R$  jet system in data and prediction in the control region after requiring 4  $b$ -tags.



**Figure 6.35:** Kinematics of the lead large- $R$  jet in data and prediction in the control region after requiring 3  $b$ -tags.

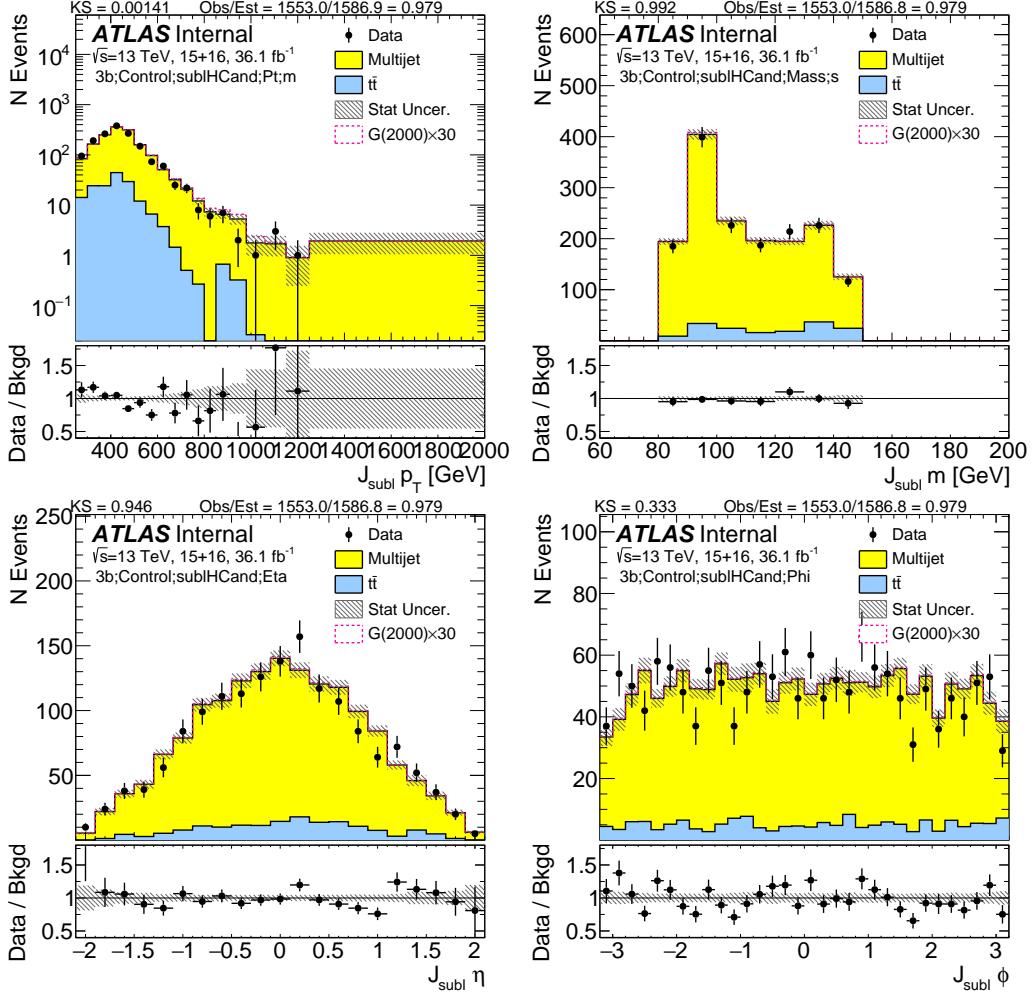
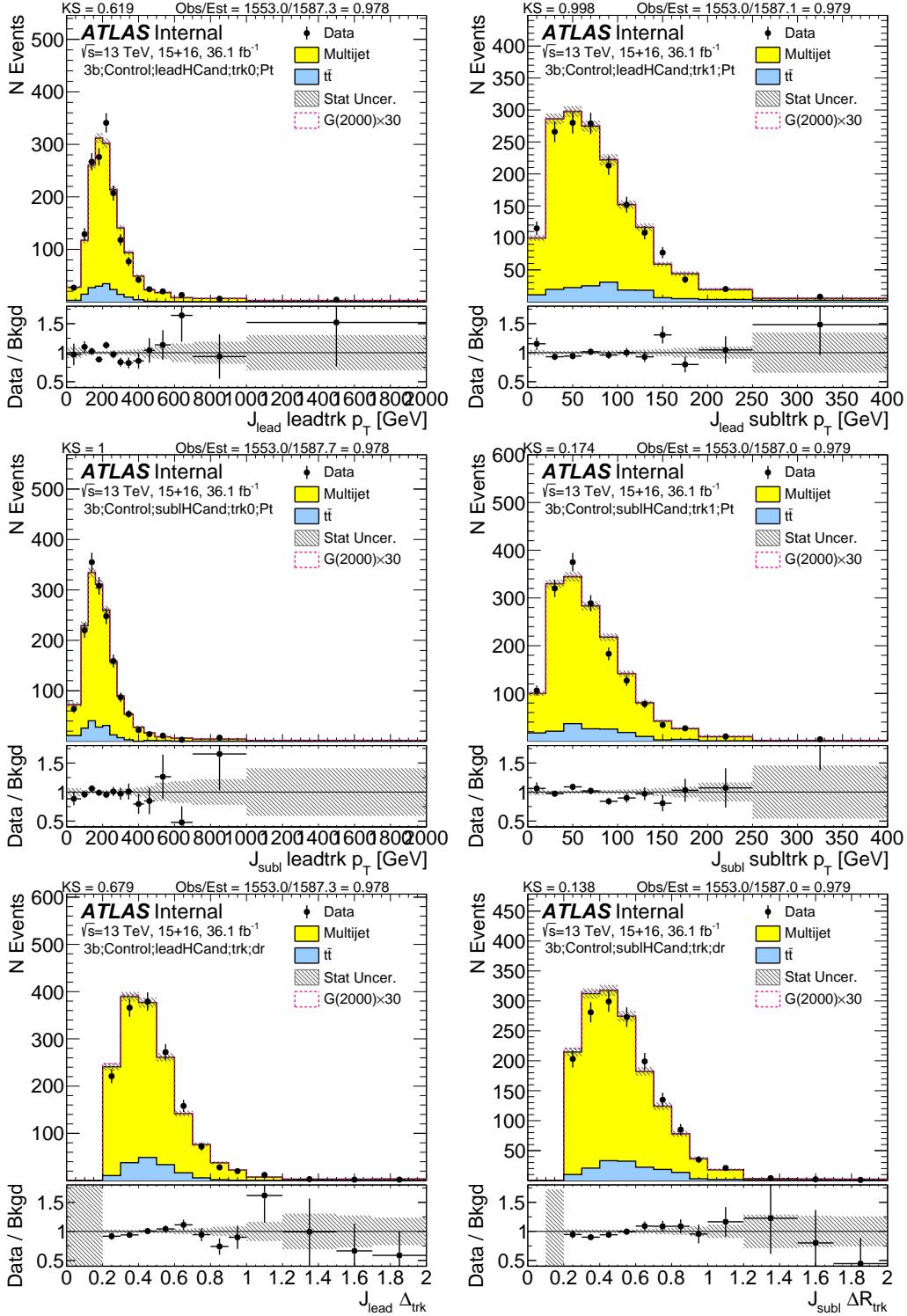


Figure 6.36: Kinematics of the sub-lead large- $R$  jet in data and prediction in the control region after requiring 3  $b$ -tags.



**Figure 6.37:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the control region after requiring 3  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.

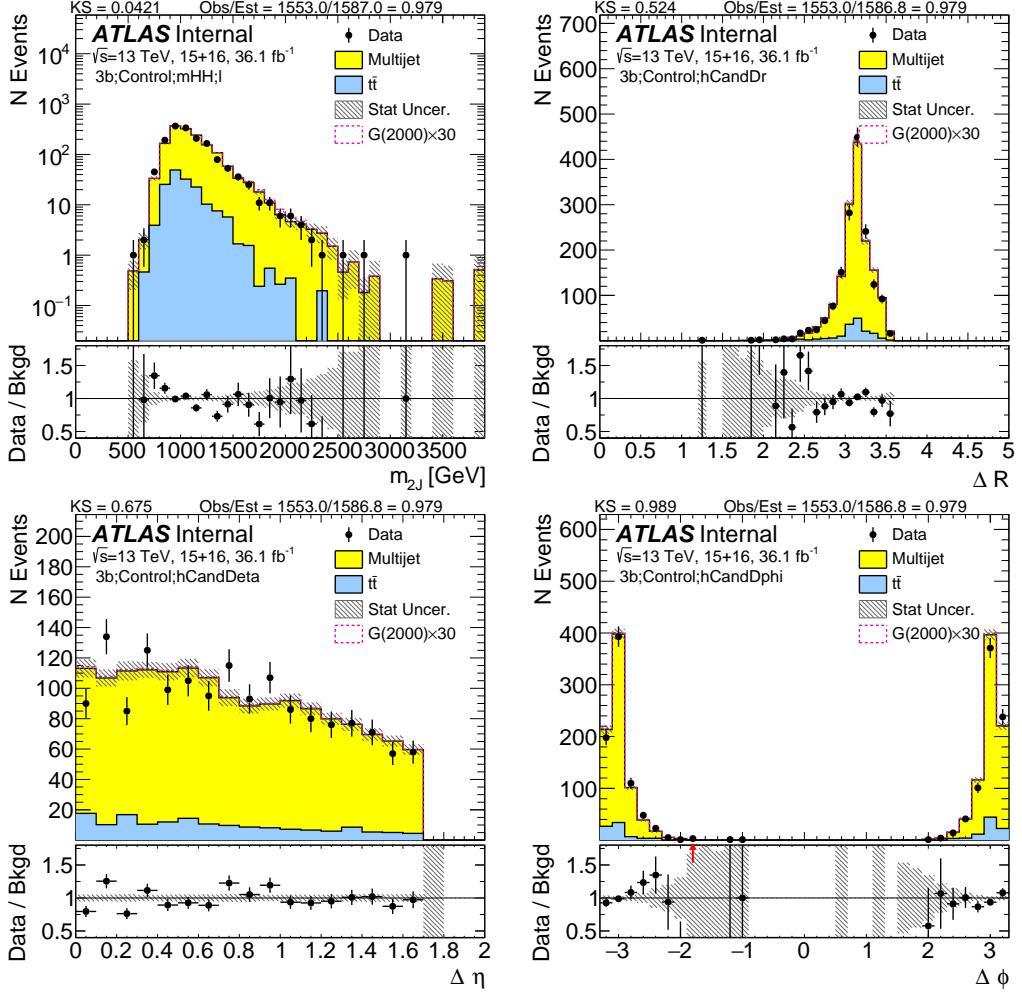
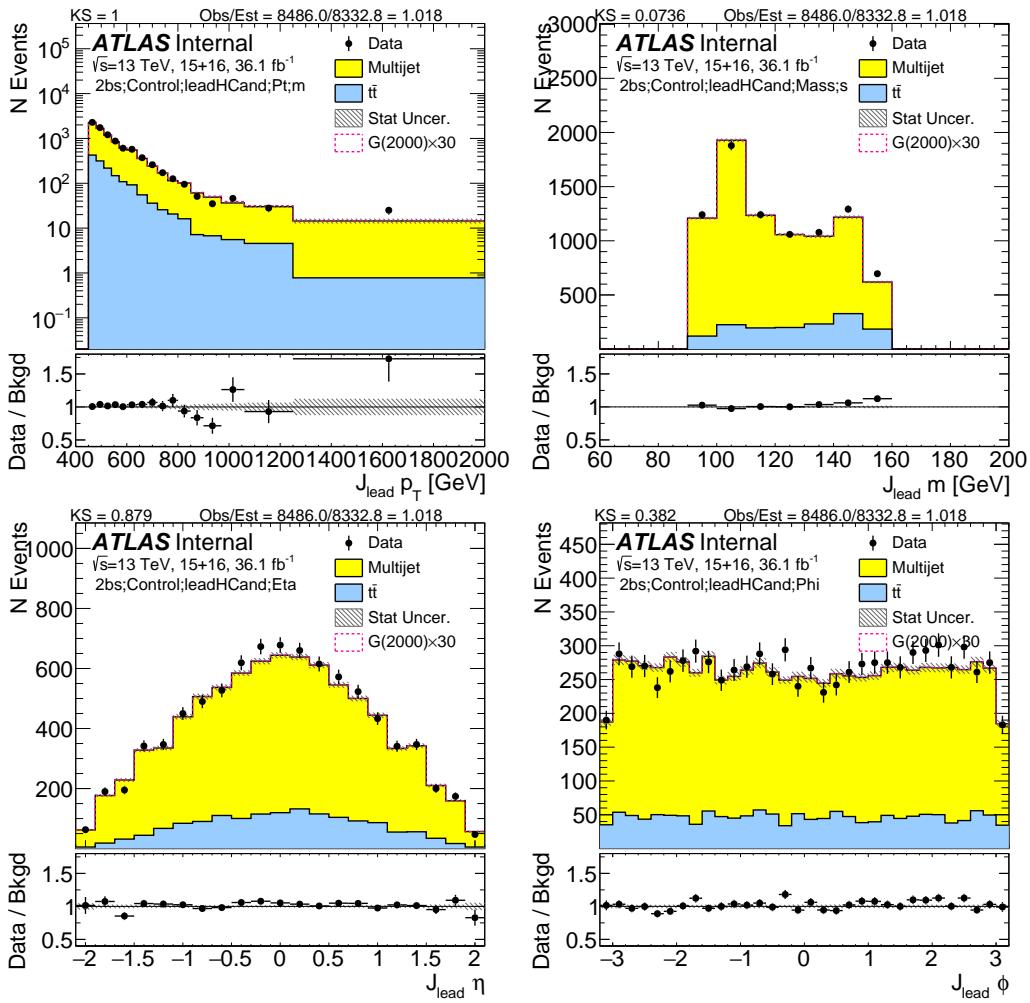
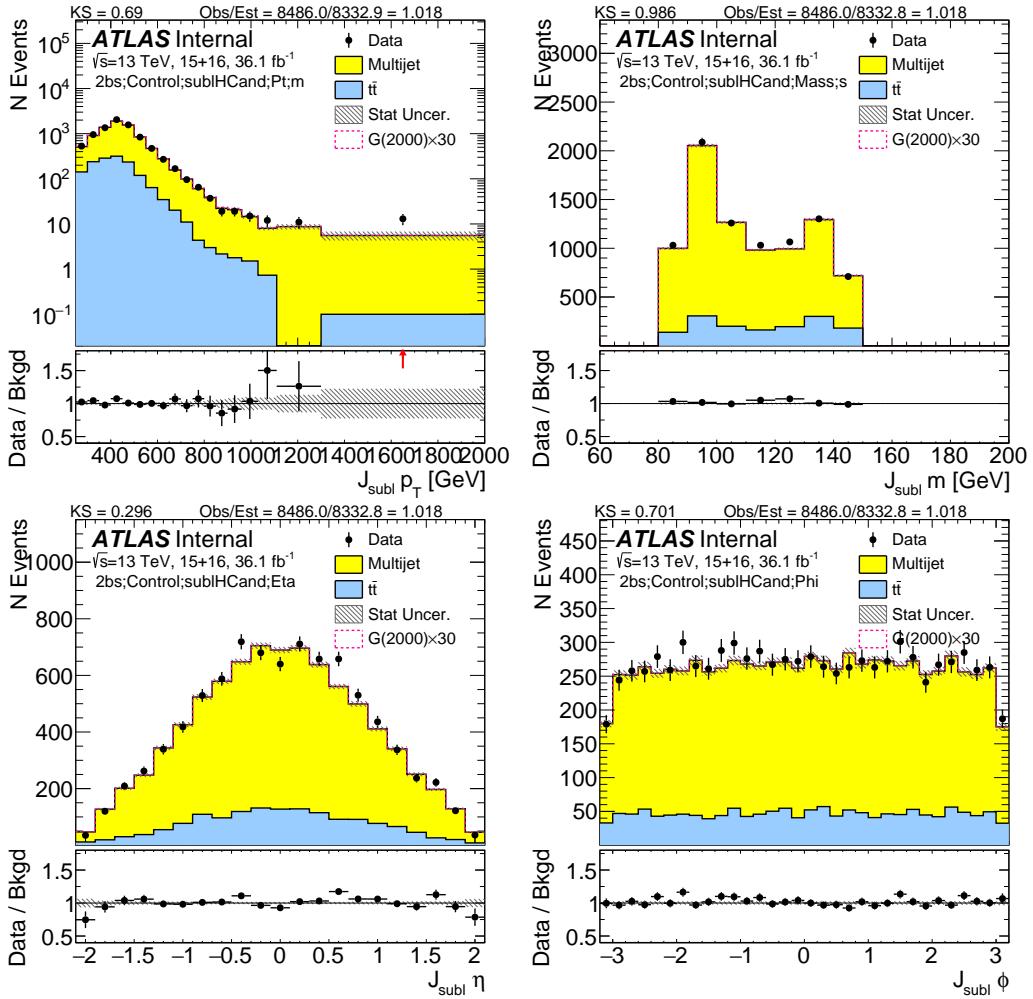


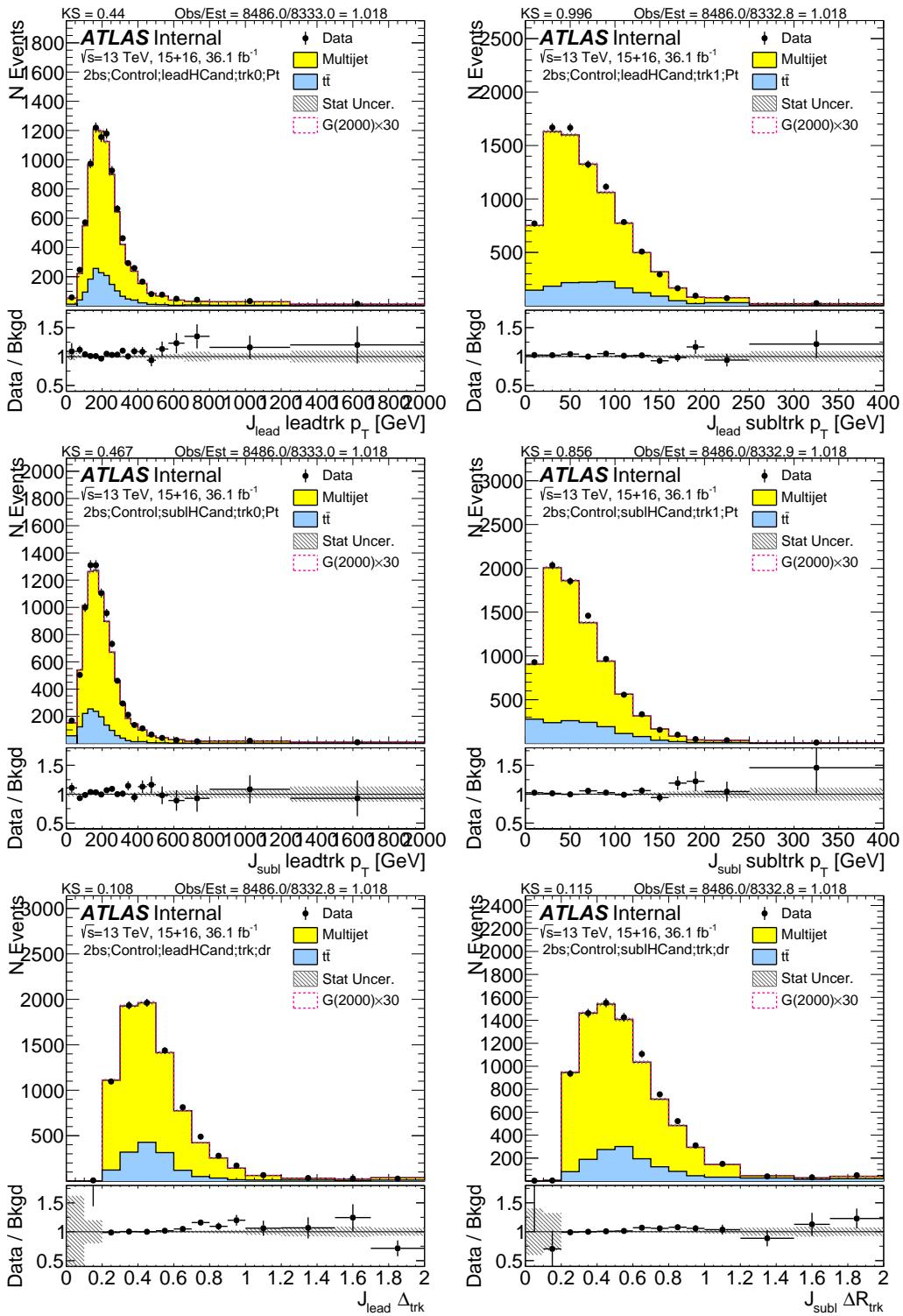
Figure 6.38: Kinematics of the large- $R$  jet system in data and prediction in the control region after requiring 3  $b$ -tags.



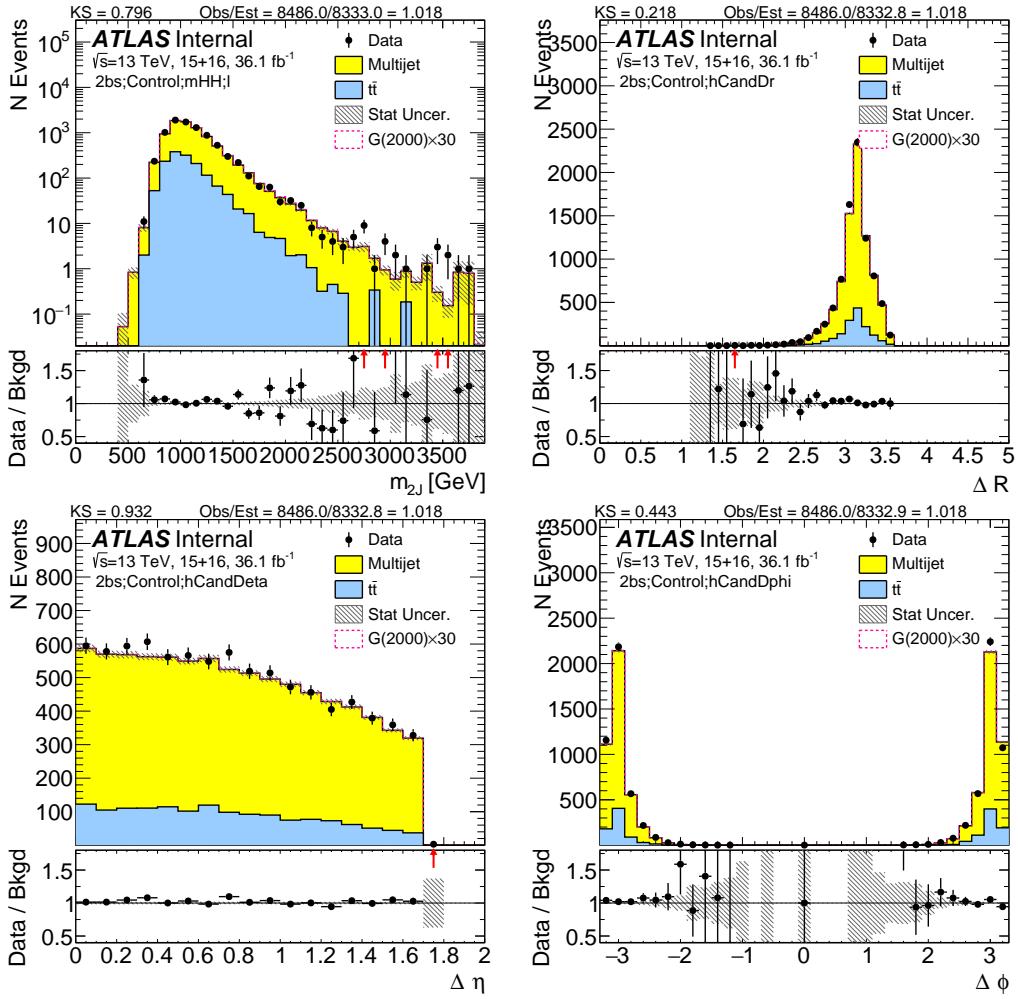
**Figure 6.39:** Kinematics of the lead large- $R$  jet in data and prediction in the control region after requiring 2  $b$ -tags split.



**Figure 6.40:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the control region after requiring 2  $b$ -tags split.



**Figure 6.41:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the control region after requiring 2  $b$ -tags split. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.



**Figure 6.42:** Kinematics of the large- $R$  jet system in data and prediction in the control region after requiring 2  $b$ -tags split.

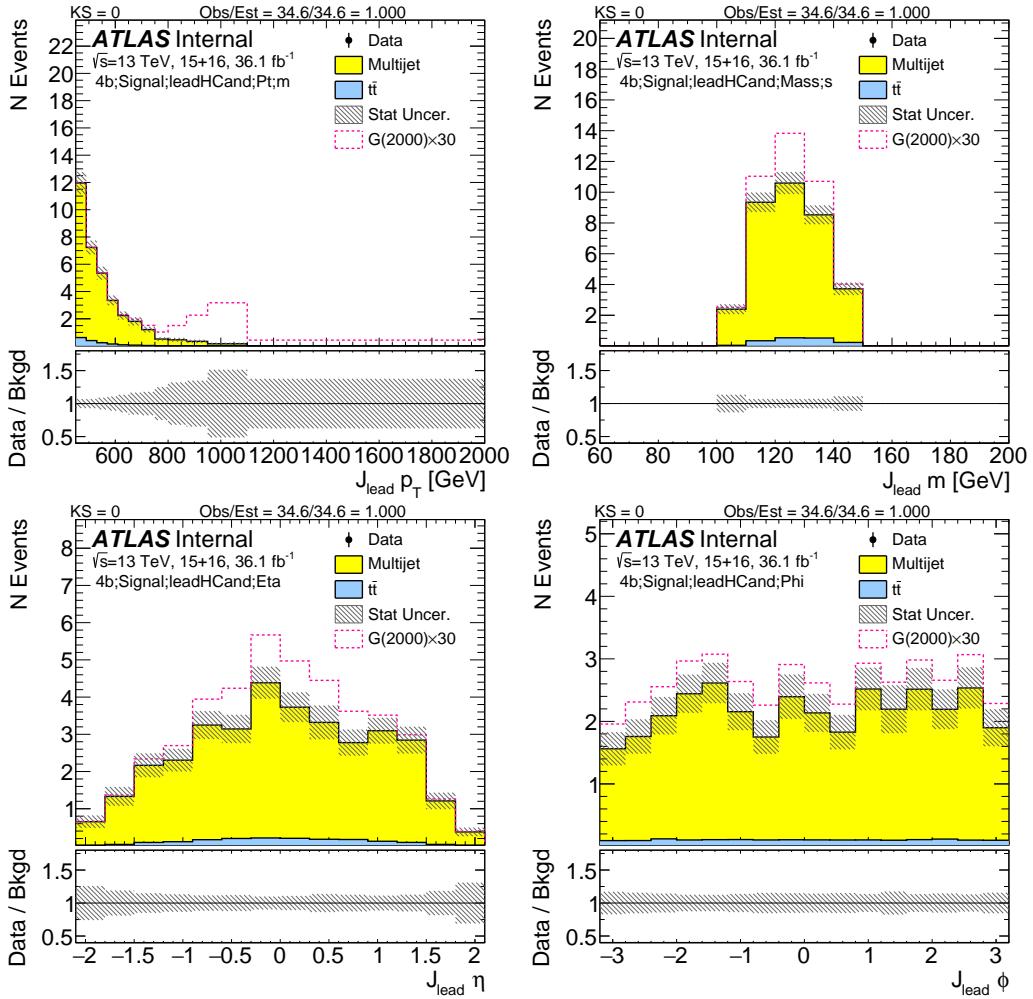
### 6.1.9 SIGNAL REGION PREDICTIONS

This section shows comparisons of data with the prediction of QCD multi-jets and  $t\bar{t}$  in the signal region (SR). Plots shown are blinded. The unblinded data is shown in the Result Section, section ??.

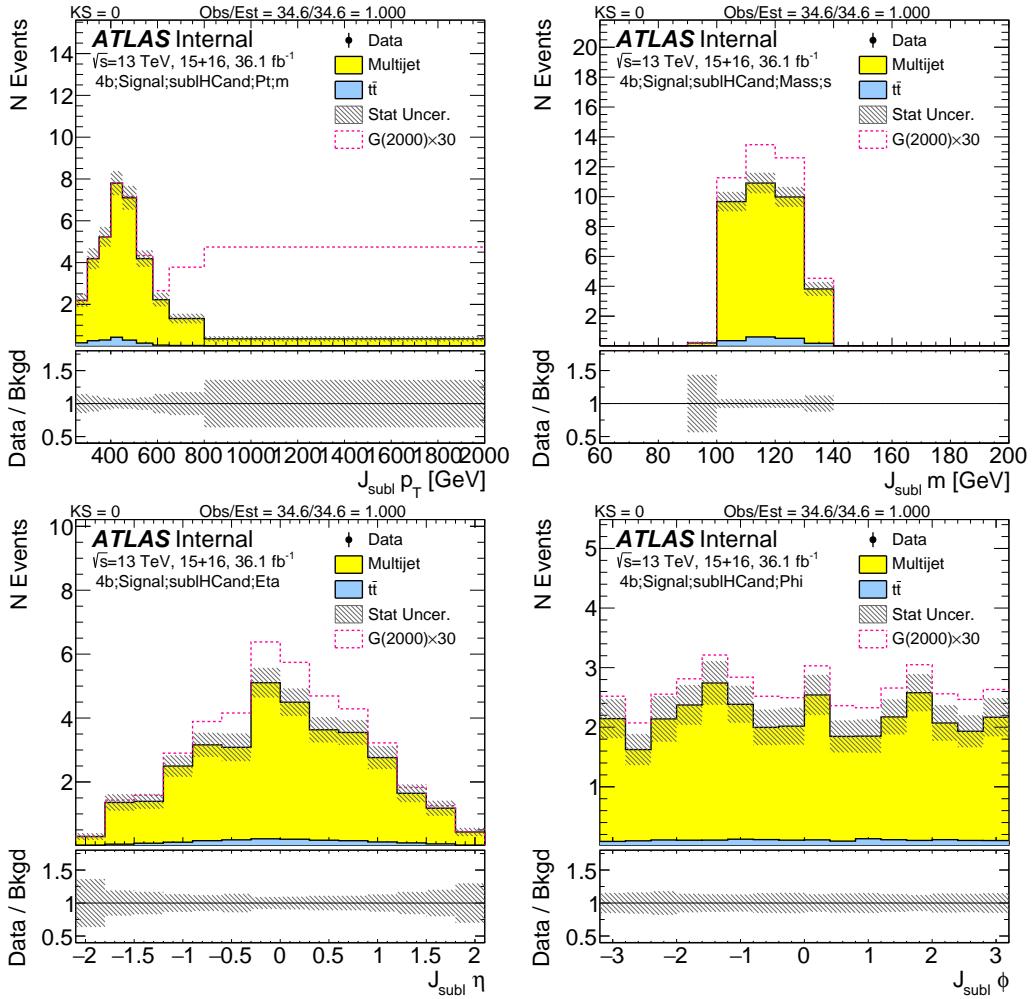
Figures 7.41, 7.42, 7.43, and 7.44 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $4b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB.

Figures 7.45, 7.46, 7.47, and 7.48 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $3b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB.

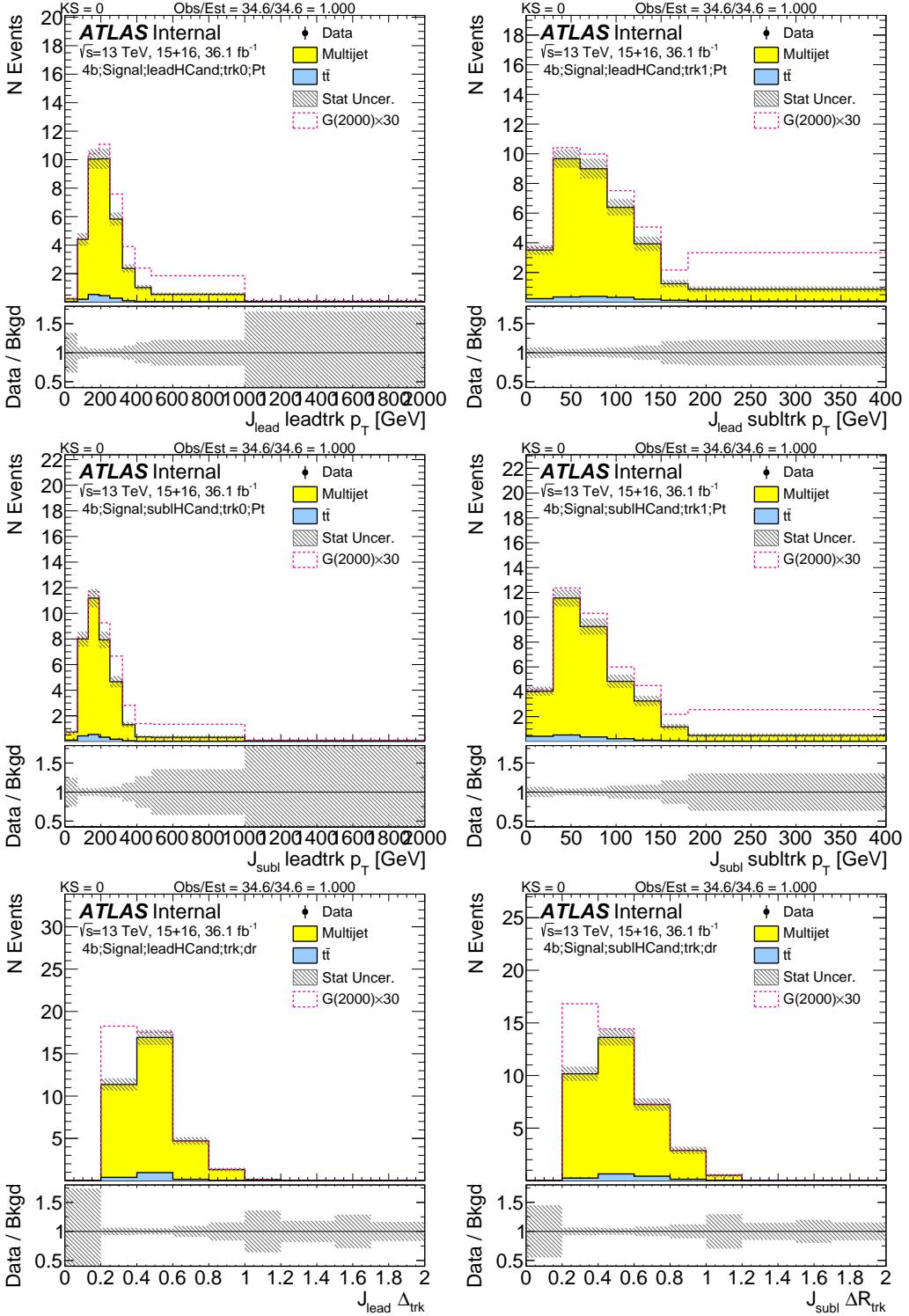
Figures 7.49, 7.50, 7.51, and 7.52 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $2b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB.



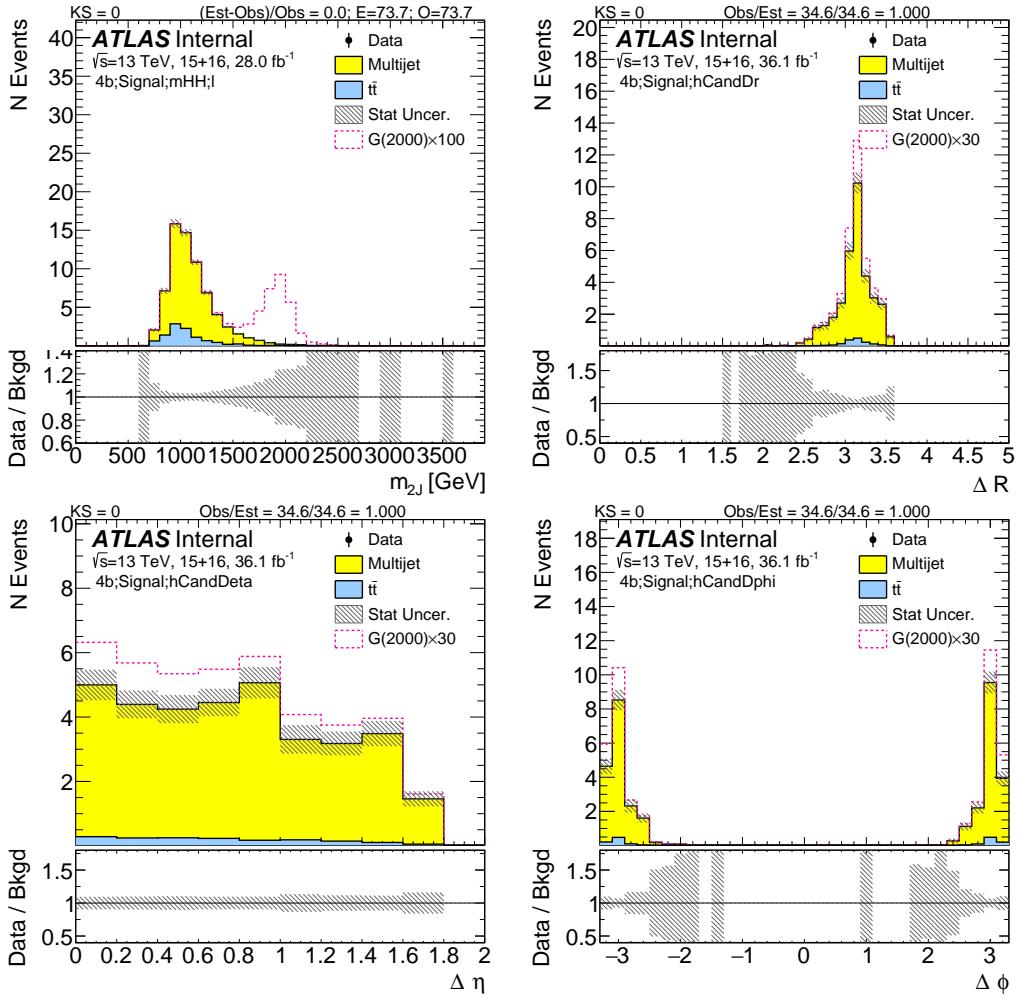
**Figure 6.43:** Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring 4  $b$ -tags. Data is blinded, and will be added after unblinding.



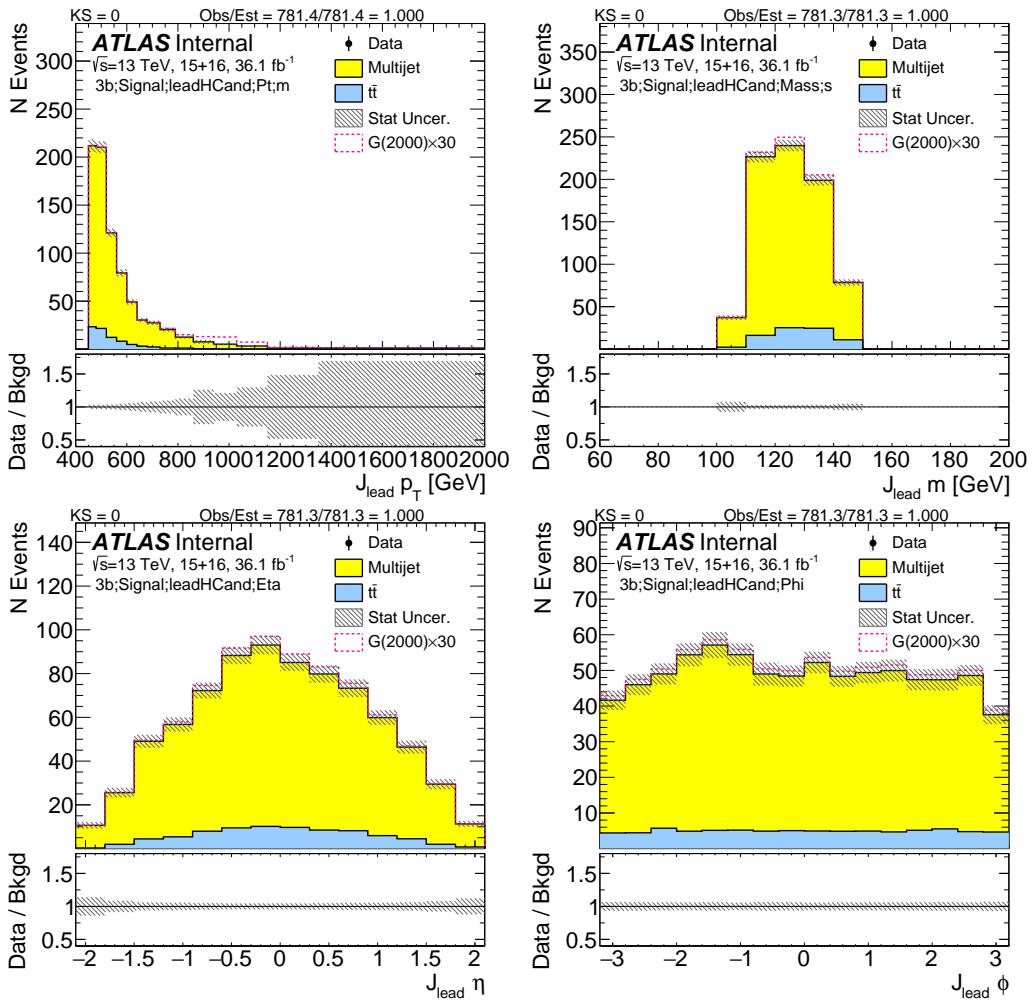
**Figure 6.44:** Kinematics of the sub-lead large- $R$  jet in data and prediction in the signal region after requiring 4  $b$ -tags. Data is blinded, and will be added after unblinding.



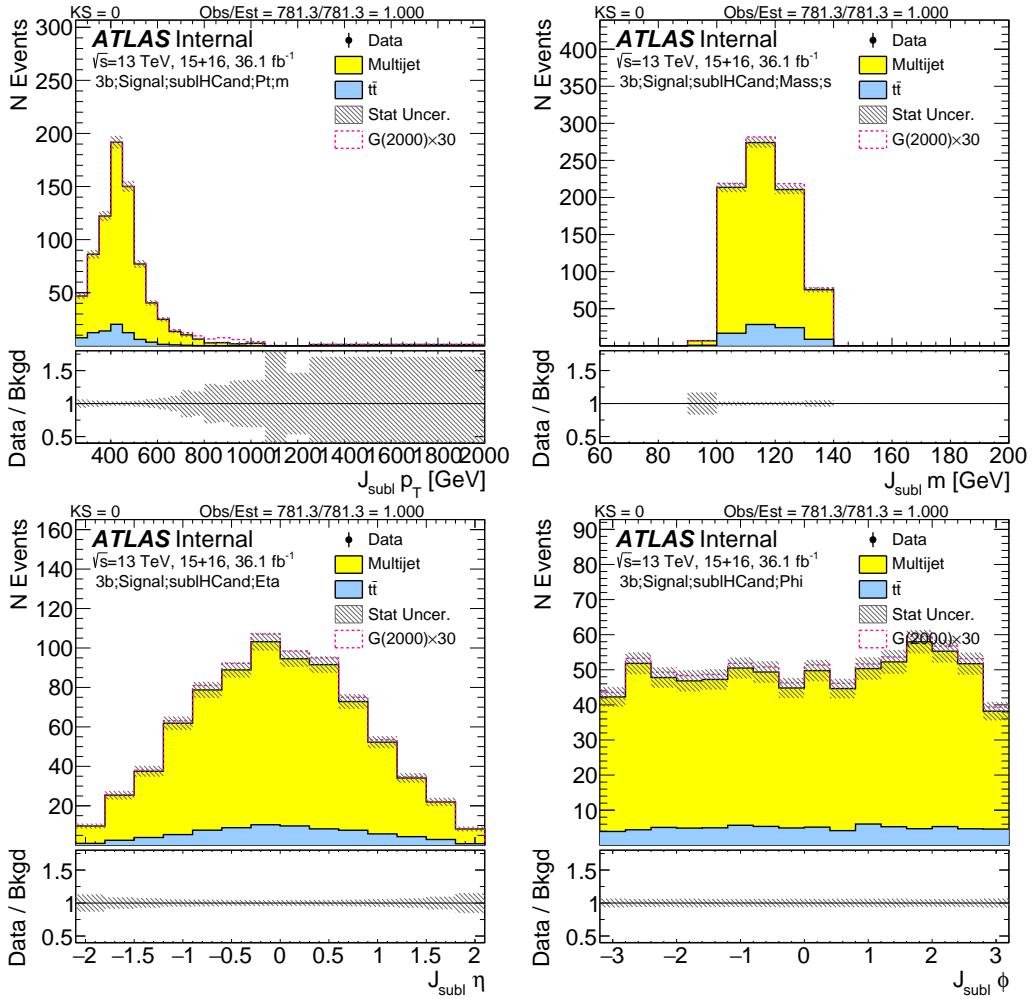
**Figure 6.45:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 4  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. Data is blinded, and will be added after unblinding.



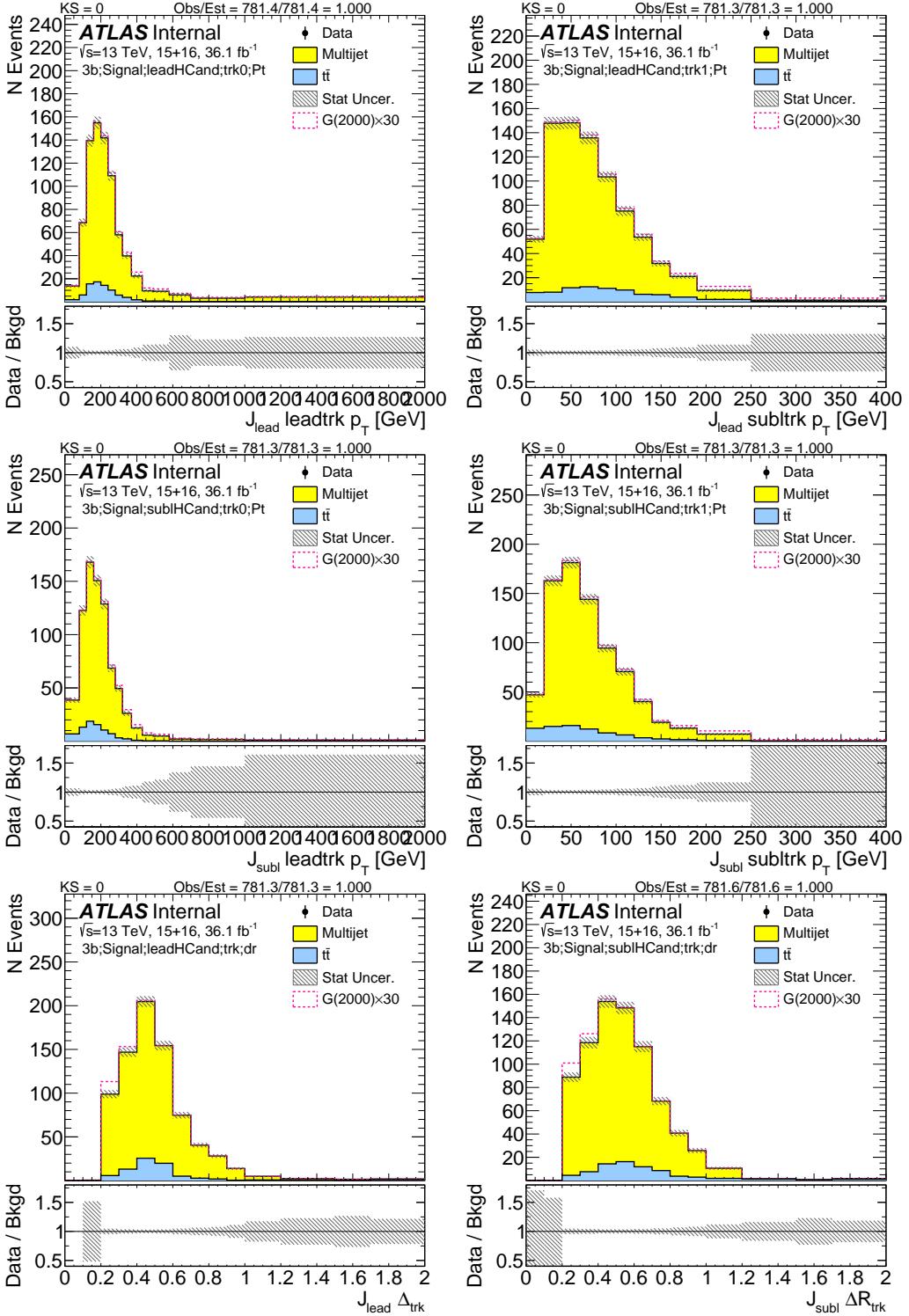
**Figure 6.46:** Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 4  $b$ -tags. Data is blinded, and will be added after unblinding.



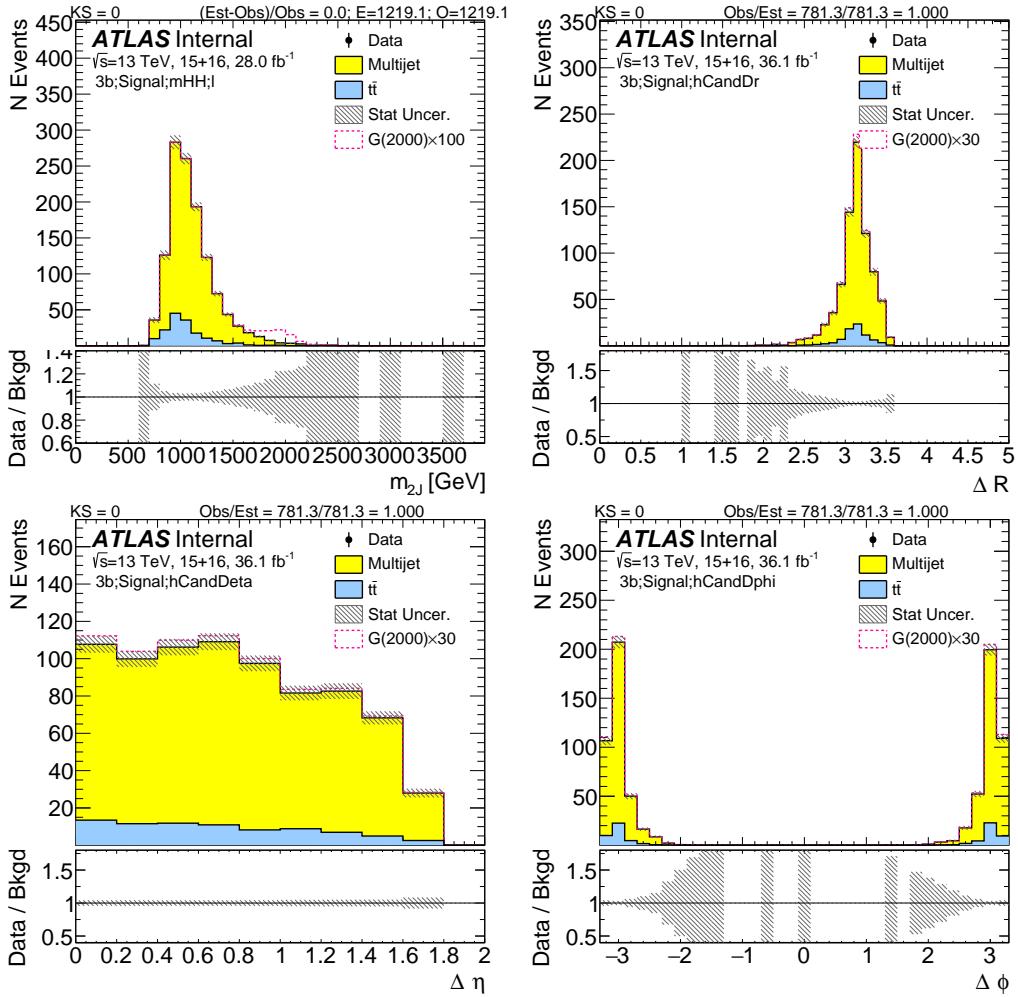
**Figure 6.47:** Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags. Data is blinded, and will be added after unblinding.



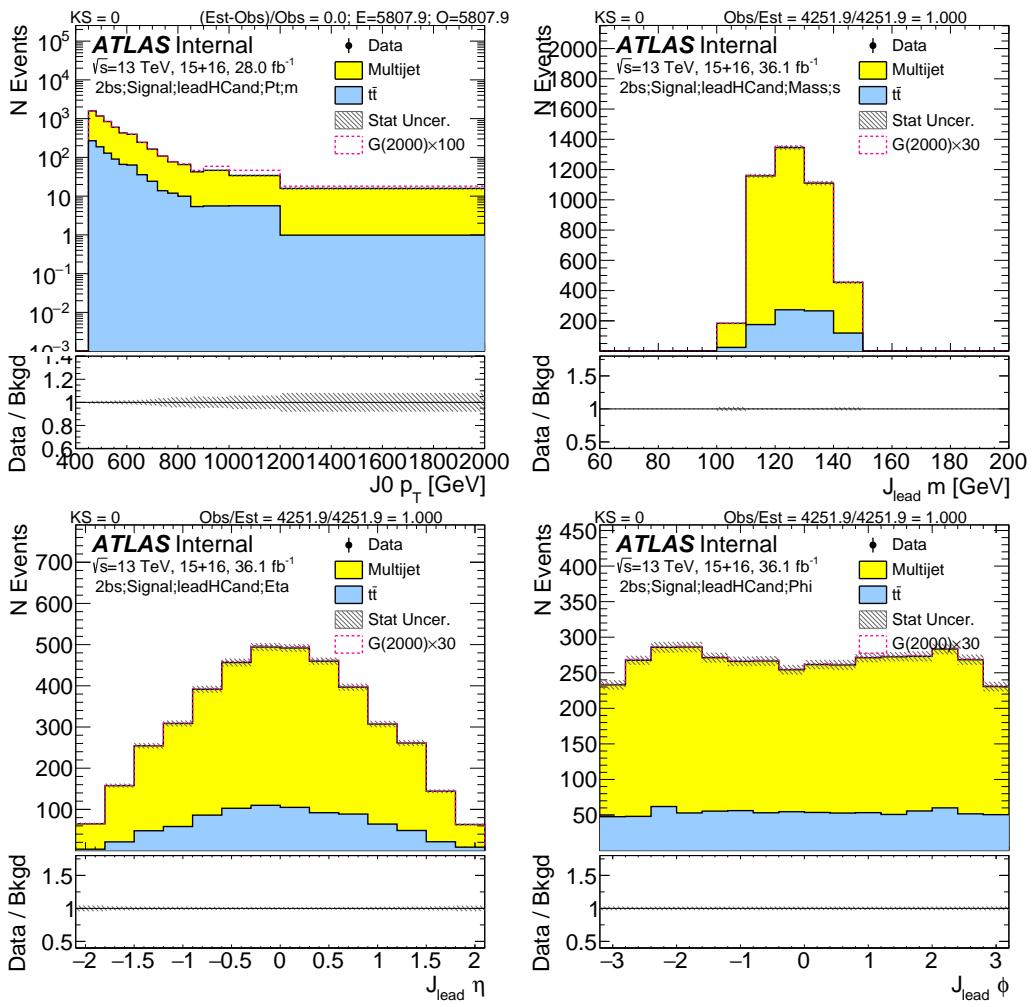
**Figure 6.48:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags. Data is blinded, and will be added after unblinding.



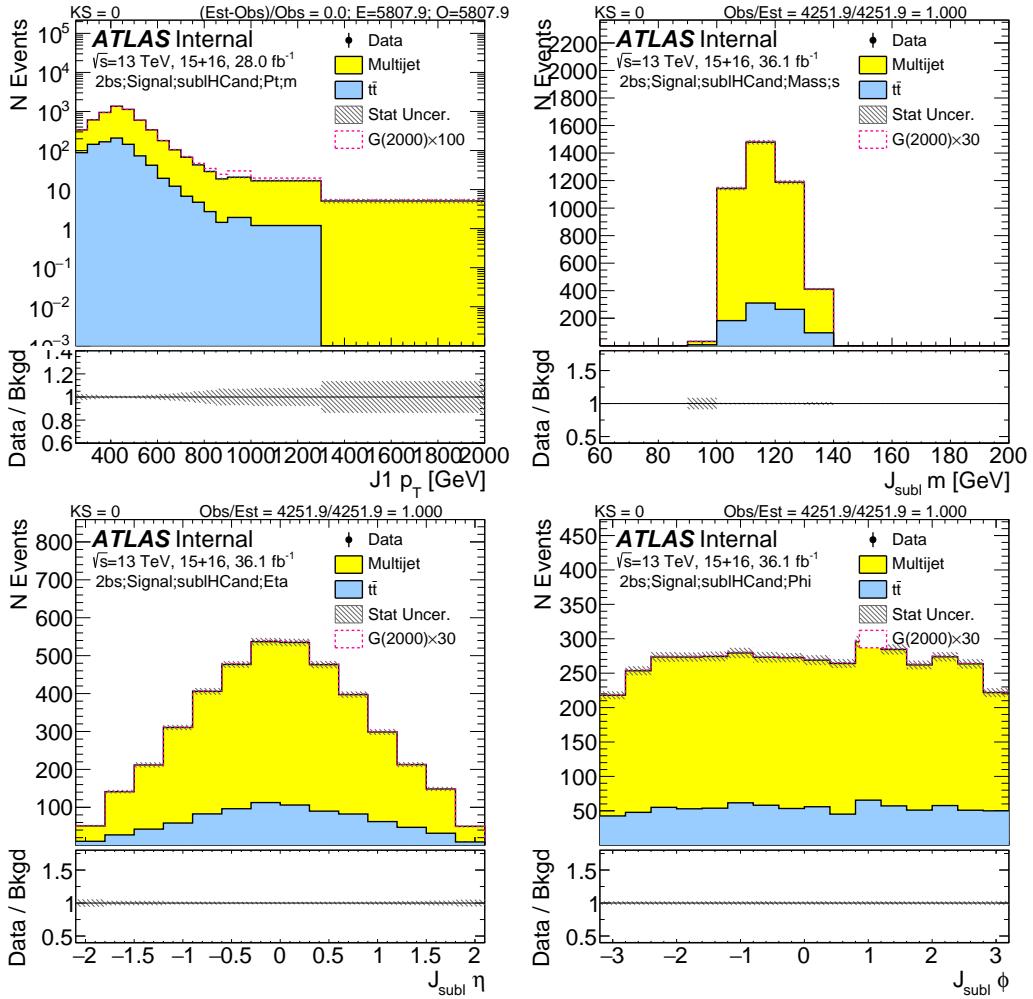
**Figure 6.49:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. Data is blinded, and will be added after unblinding.



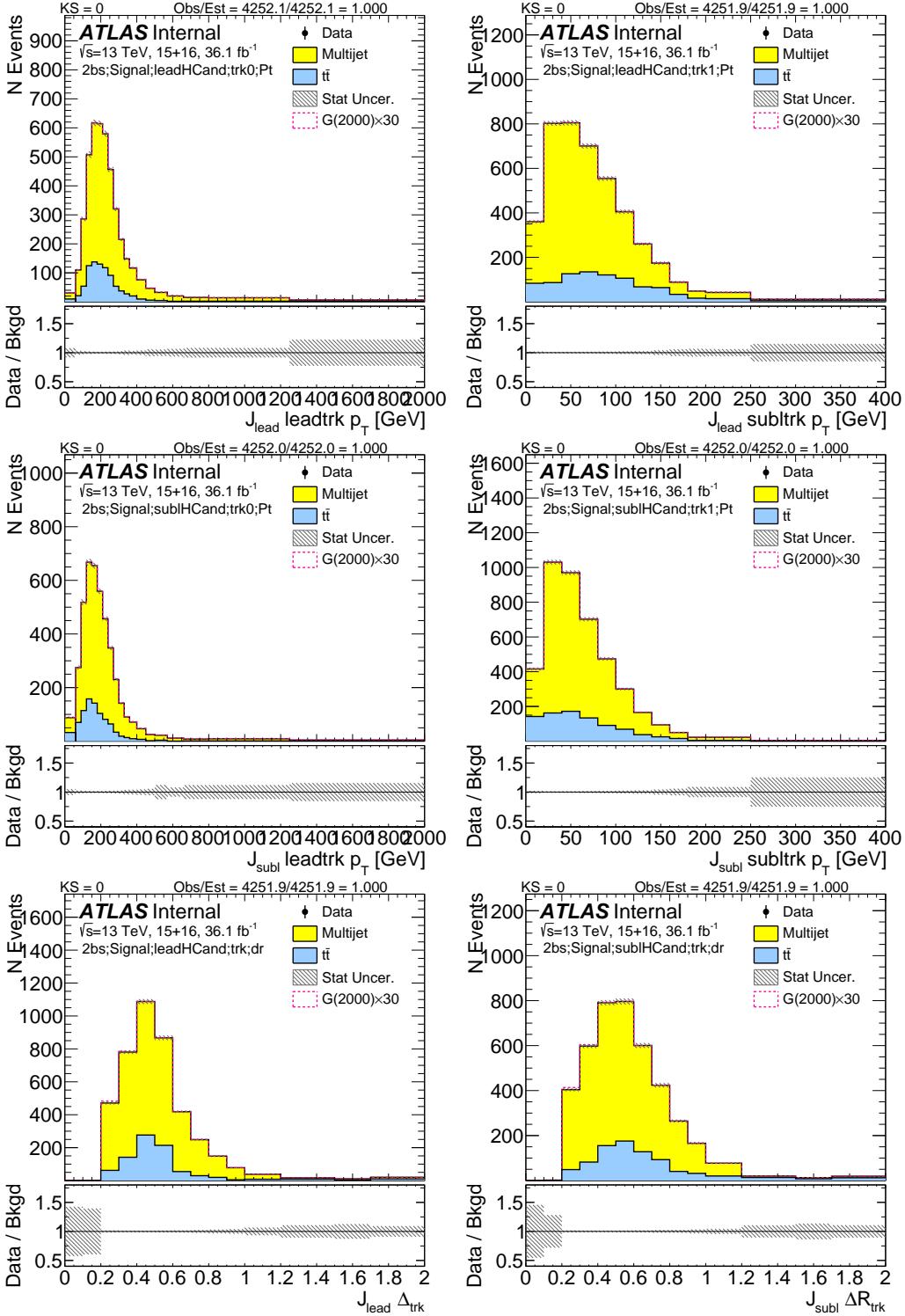
**Figure 6.50:** Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 3  $b$ -tags. Data is blinded, and will be added after unblinding.



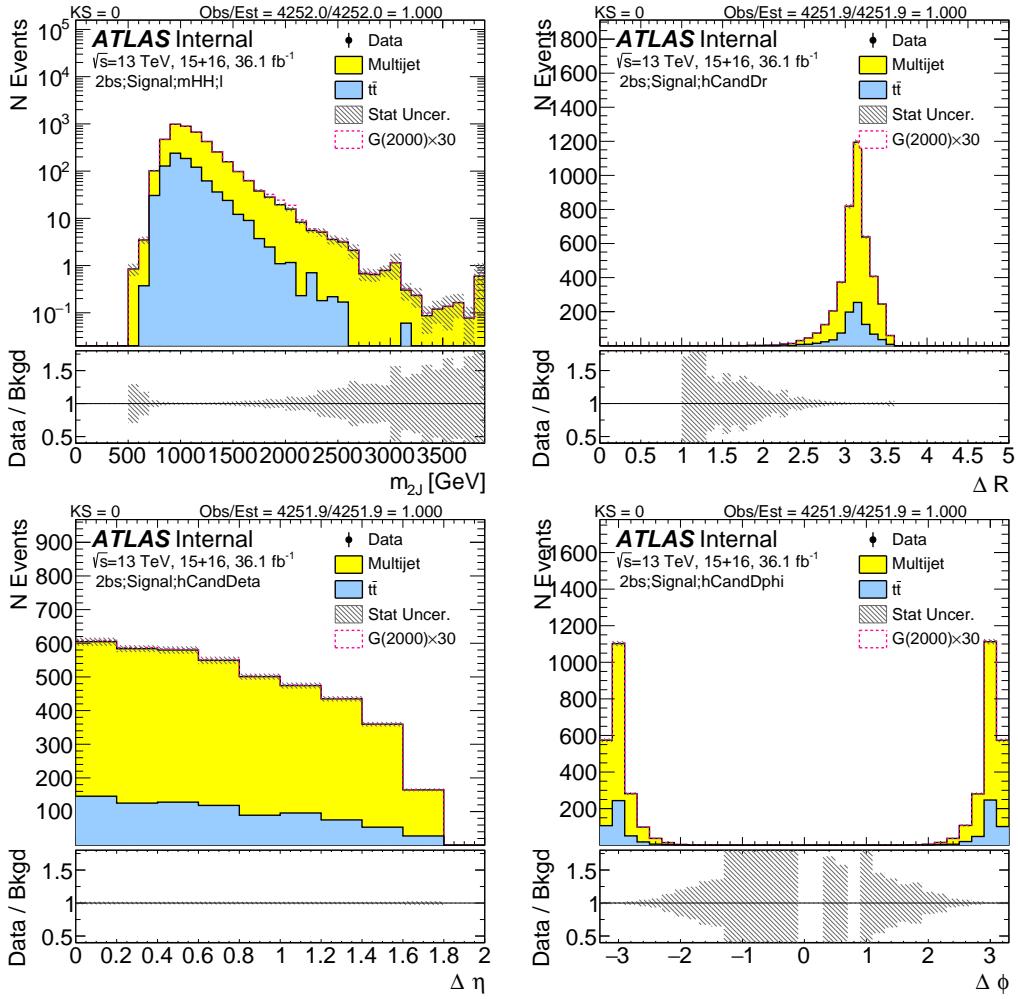
**Figure 6.51:** Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split. Data is blinded, and will be added after unblinding.



**Figure 6.52:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split. Data is blinded, and will be added after unblinding.



**Figure 6.53:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. Data is blinded, and will be added after unblinding.



**Figure 6.54:** Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 2  $b$ -tags split. Data is blinded, and will be added after unblinding.

## SIGNAL REGION SMOOTHING

Due to the improved  $1/2b$  statistics at high di-large  $-R$  jet invariant mass above 1500 GeV and the limited  $t\bar{t}$  statistics above 1100 GeV, different fits are performed to smooth the di-large  $-R$  jet mass distribution in the signal region. The  $1/2b$  QCD background is fit with the MJ8 functional form:

$$y = \frac{a}{\frac{x^2}{\sqrt{s}}} \left(1 - \frac{x}{\sqrt{s}}\right)^{b-c} \log\left(\frac{x}{\sqrt{s}}\right) \quad (6.5)$$

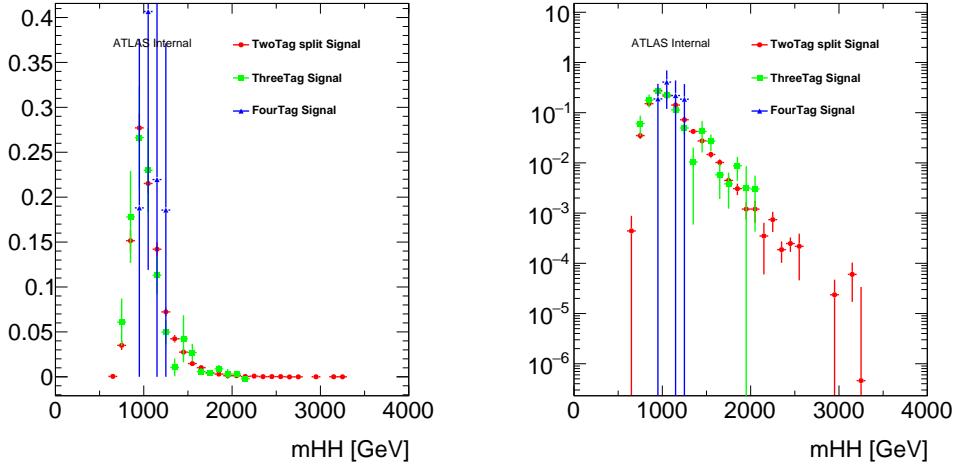
where  $\sqrt{s} = 13000$  GeV, in the range  $1200 < M_{JJ} < 3000$  GeV, and the three free parameters are  $a$ ,  $b$  and  $c$ . This form is used in fitting, as it was seen to be easier for the fits to converge. The signal region  $t\bar{t}$  distribution is fitted also with the dijet functional form, also in the range  $1200 < M_{JJ} < 3000$  GeV, without parameter constraints. The values of the estimated fit parameters in the  $4b$  and  $3b$  and  $2bs$  signal regions can be found in Table 7.3.

Given that the similar  $1/2b$  sample is used for deriving the QCD shape for the  $4/3/2bs$  signal regions, it is not surprising that the slope parameter ( $a$ ) is similar in the  $4/3/2bs$  signal regions for each the QCD backgrounds.

Due to the very limited statistics of the  $4b$   $t\bar{t}$  sample, the  $4b$   $t\bar{t}$  dijet mass shape is used from the  $3b$   $t\bar{t}$  dijet mass shape normalized to the number of  $4b$   $t\bar{t}$  events. A comparison of the shape is shown in Figure 7.53. Good agreement between the  $4b$  and  $3b$  signal region plot is shown.

Figure 7.54 shows the smoothing fits for the QCD background and the  $t\bar{t}$  background in the  $4b$  signal region. Figure 7.55 shows the same for the  $3b$  signal region. Figure 7.56 shows the same for the  $2bs$  signal region. The smoothing statistical uncertainties are also shown on these two plots. More additional uncertainties, such as uncertainty from choice of smoothing function, will be discussed in the Section 8.0.3.

The final smoothed background predictions for the  $4b$  and  $3b$  and  $2bs$  signal regions can be found in Figure 7.57. This includes smoothing statistical uncertainties only. More details on other system-



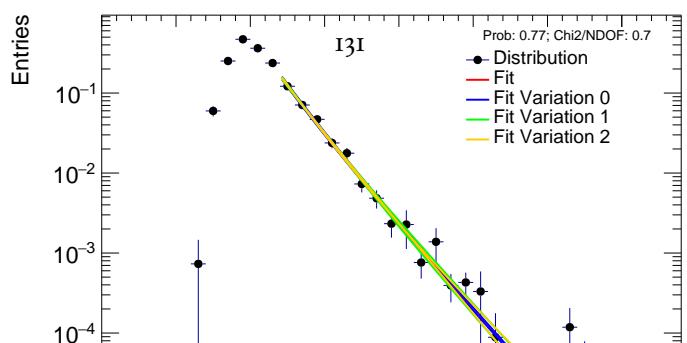
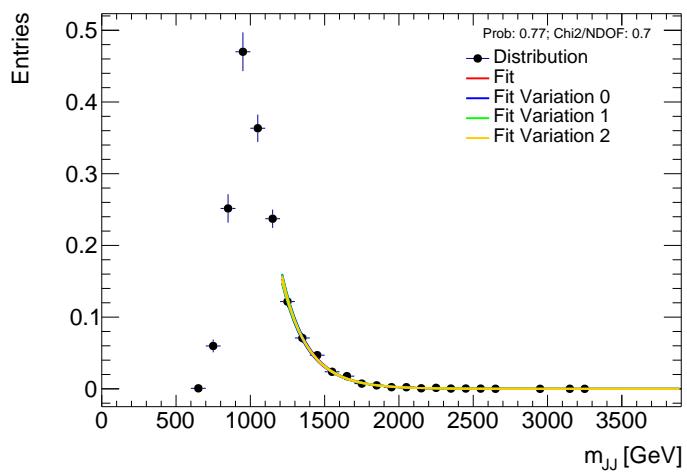
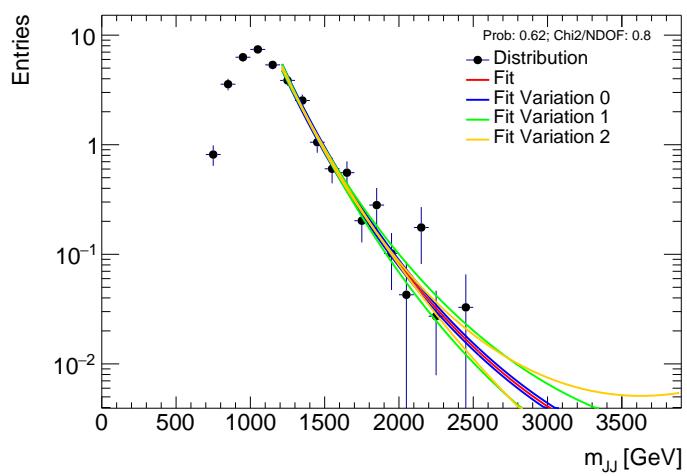
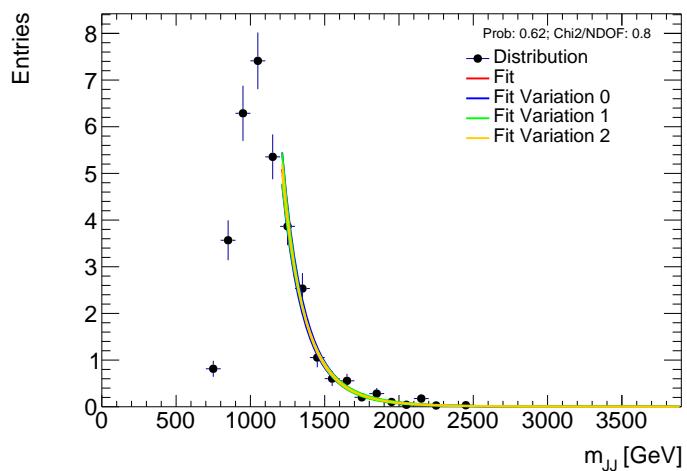
**Figure 6.55:** Comparison of the  $4b$ ,  $3b$  and  $2bs$  signal region  $t\bar{t}$  dijet mass shape. On the left is the linear scale, and on the right is the log scale. Both distributions are normalized to 1 for comparison.

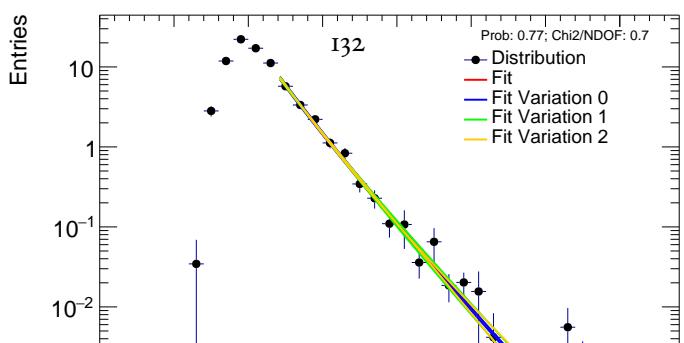
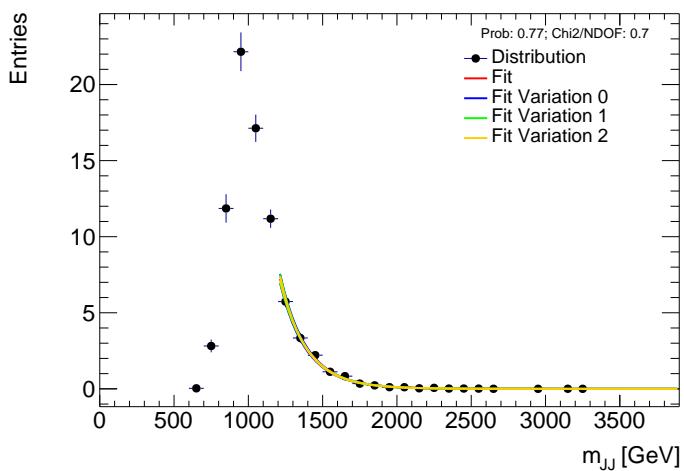
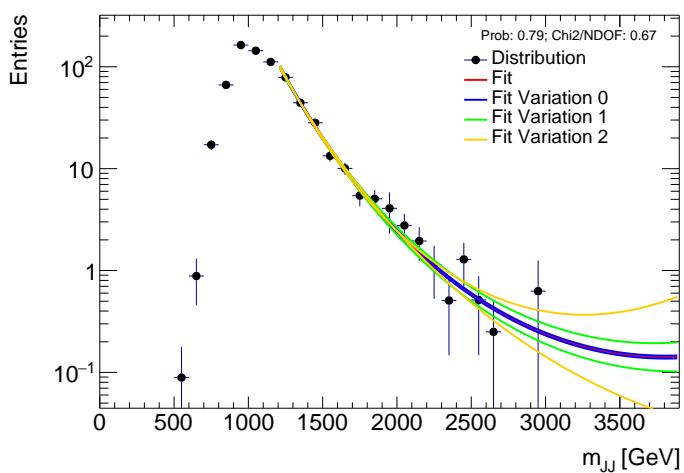
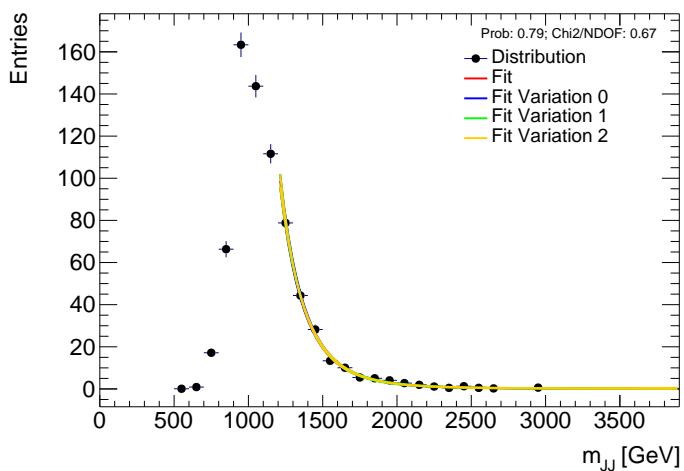
atics, including smoothing systematics, shape uncertainties and other sources of uncertainties would be discussed in Section 8.0.3.

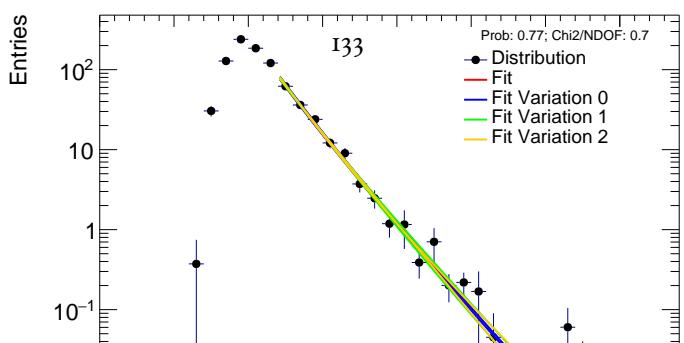
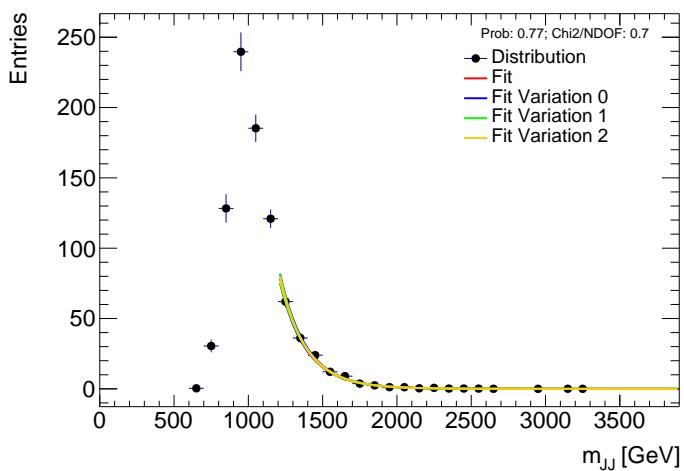
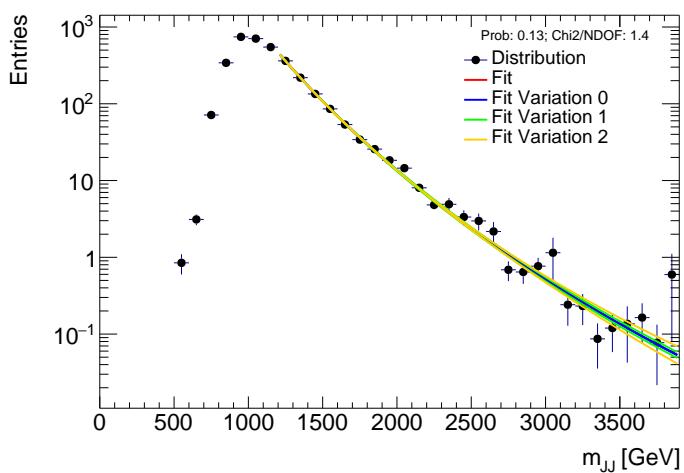
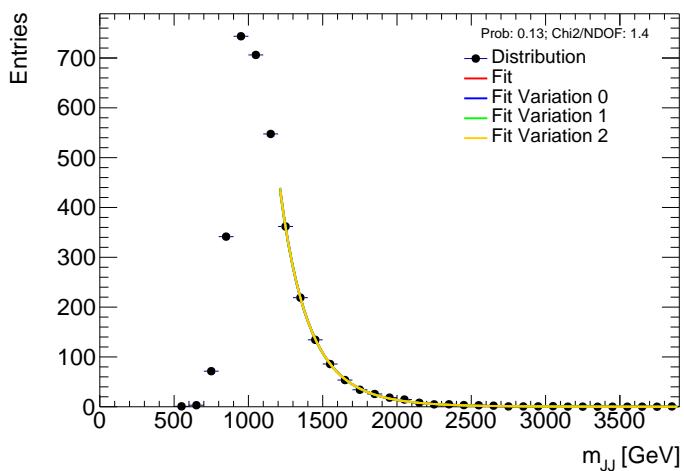
Uncertainties on the fit parameters are propagated as systematic uncertainties, though they are essentially replacing the bin-by-bin statistical uncertainties of the background estimates (which are not used once smoothing is applied). Correlations in the fit parameters of the backgrounds are taken into account when propagating the uncertainties, as described in Appendix ??.

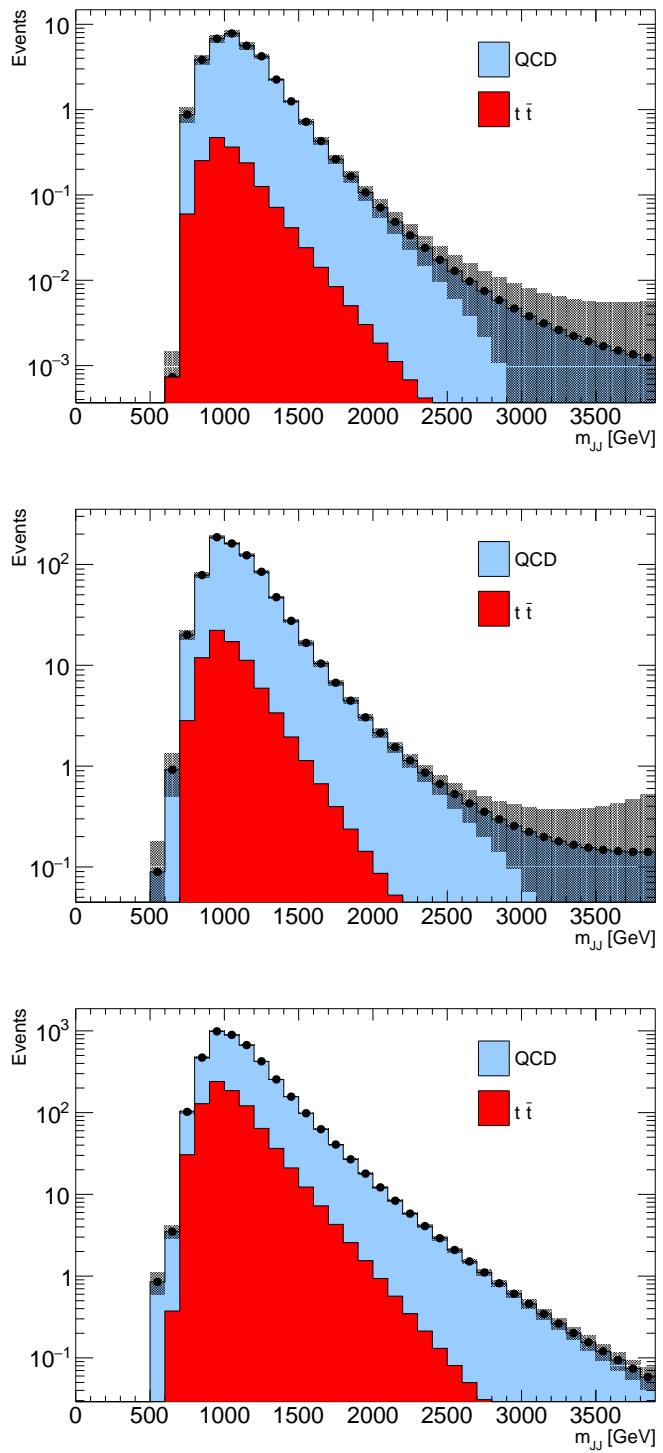
Region	$a_{t\bar{t}}$	$b_{t\bar{t}}$	$c_{t\bar{t}}$	$a_{qcd}$	$b_{qcd}$	$c_{qcd}$
FourTag	-1.02 $\pm$ 1.22	35.62 $\pm$ 10.83	9.05 $\pm$ 9.59	7.75 $\pm$ 1.8	-18.0 $\pm$ 16.75	54.36 $\pm$ 14.28
ThreeTag	2.83 $\pm$ 1.22	35.62 $\pm$ 10.83	9.05 $\pm$ 9.59	10.42 $\pm$ 1.73	-27.1 $\pm$ 15.78	56.91 $\pm$ 13.89
TwoTag split	5.21 $\pm$ 1.22	35.62 $\pm$ 10.83	9.05 $\pm$ 9.59	7.74 $\pm$ 0.36	7.22 $\pm$ 3.11	24.54 $\pm$ 2.78

**Table 6.3:** Smoothing parameters in  $4b$  and  $3b$  and  $2bs$  signal regions, the correlation between parameters is almost always 0.99.









**Figure 6.59:** Smoothed background estimations the  $4b$  (top),  $3b$  (middle), and  $2bs$  (bottom) signal regions. Only smoothing statistical uncertainties are shown here.  
134

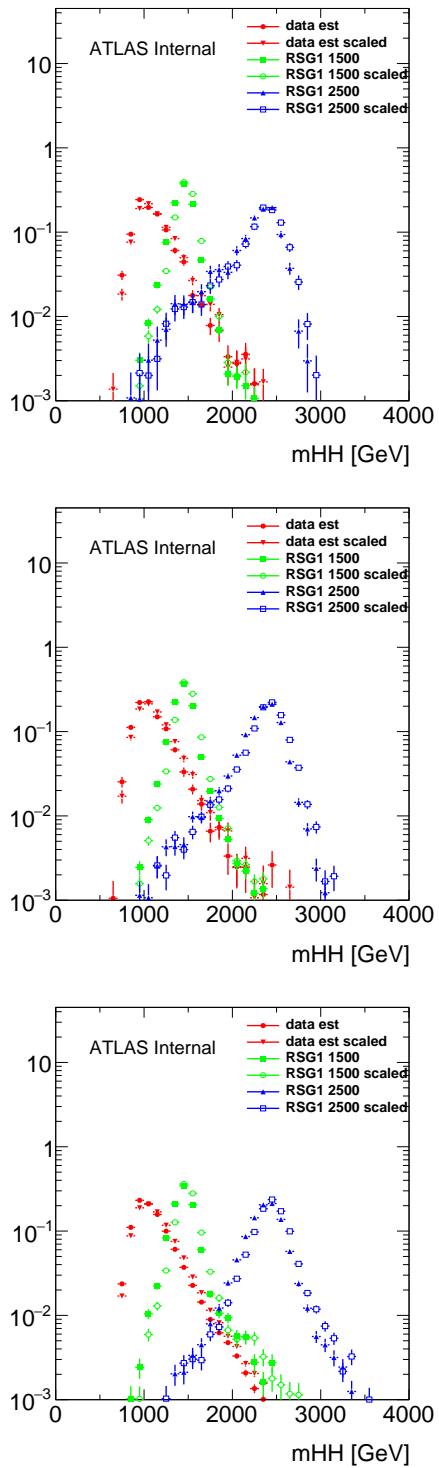
## SCALED DIJET MASS DISTRIBUTION IN SIGNAL REGION

As is done in the resolved analysis, we also consider the scaled  $M_{jj}$  distribution. In this case, the two higgs candidate 4-vectors are scaled by  $m_h/m_j$ , where  $m_h = 125$  GeV, and  $m_j$  is the *large – Rjet* mass of the Higgs candidate. While this distribution is expected to have less impact on the boosted analysis, because the mass correction is small relative to the signal masses being considered, we investigate this variable for possible improvements and for consistency with the resolved analysis. The scaled dijet mass distribution can be found in Figure 7.58. Its impact of the boosted analysis limit can be found in Appendix ??.

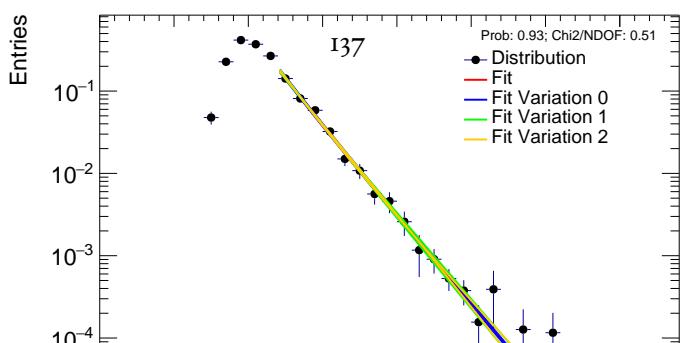
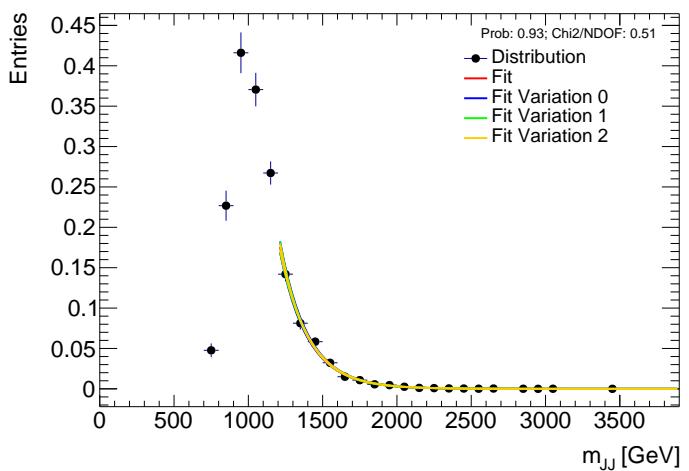
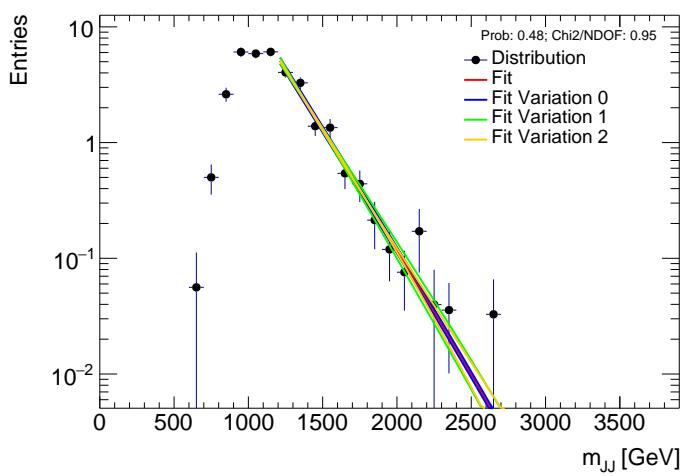
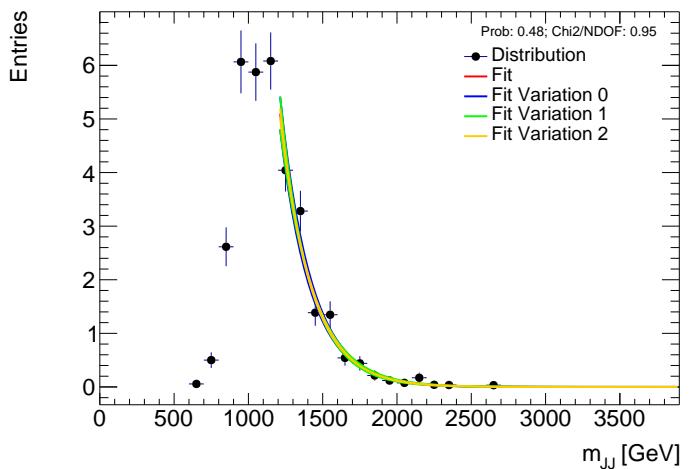
For determining the choice of the final discriminant, both the expected limits on the nominal and scaled dijet mass distribution have been computed. Since the scaled dijet mass distribution based limits are consistent (slightly better at low mass and slightly worse at high mass, with differences of the order of 10%) than the nominal dijet mass limits, we proceed to use the scaled dijet mass distribution for consistency with the resolved analysis.

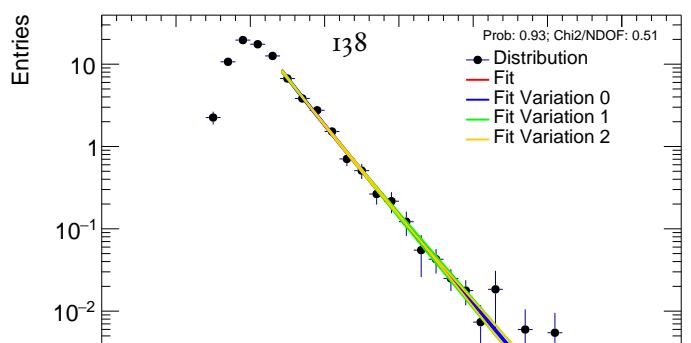
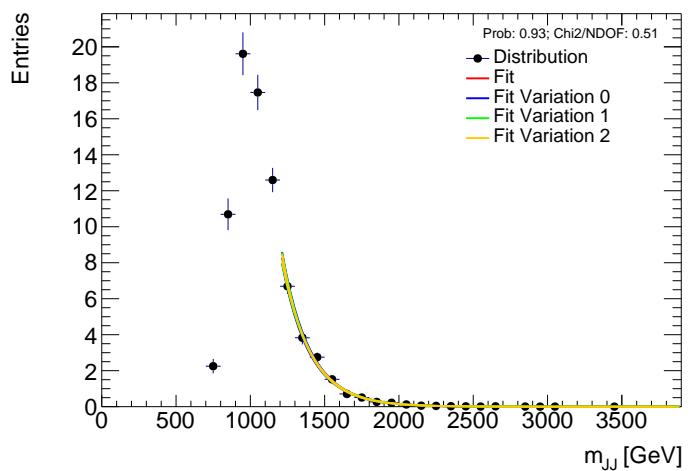
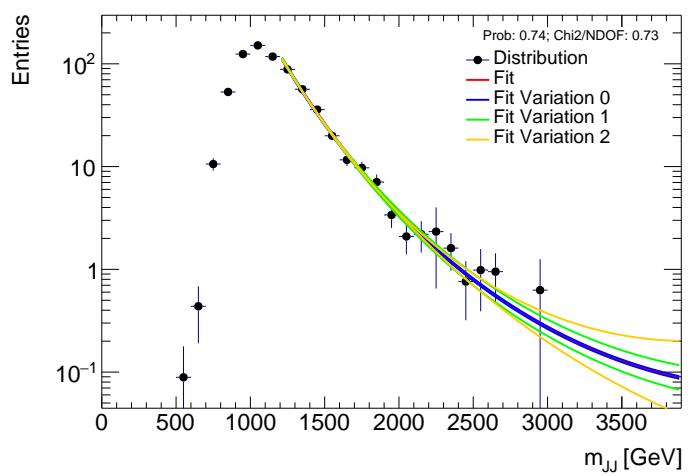
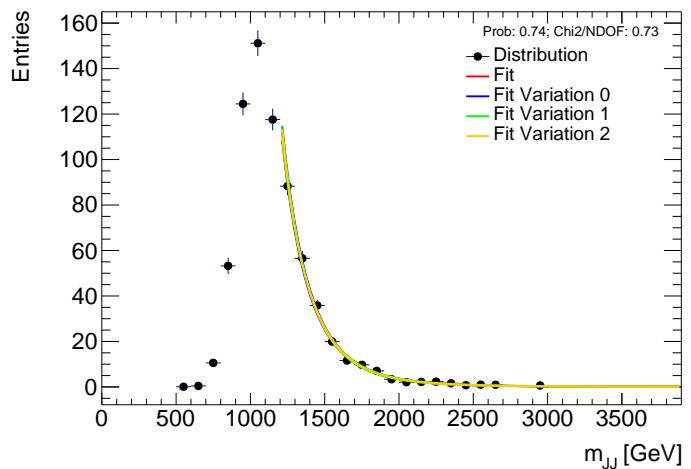
As is done for the dijet mass distribution, the scaled dijet mass distribution is smoothed. The smoothing is performed between 1200 GeV and 3000 GeV for the QCD and  $t\bar{t}$ . The smoothed distributions can be seen in Figures 7.59, 7.60, and 7.61. The values of the estimated fit parameters in the  $4b$  and  $3b$  and  $2bs$  signal regions can be found in Table 7.4.

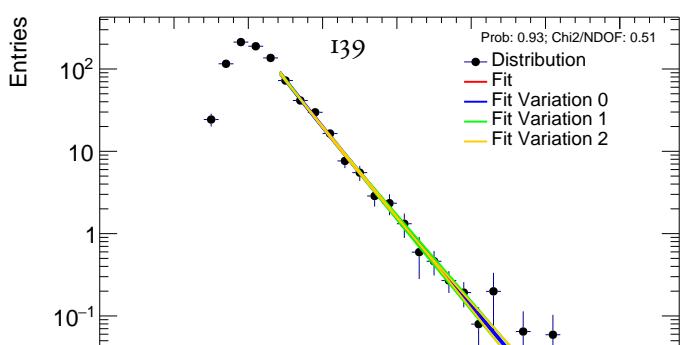
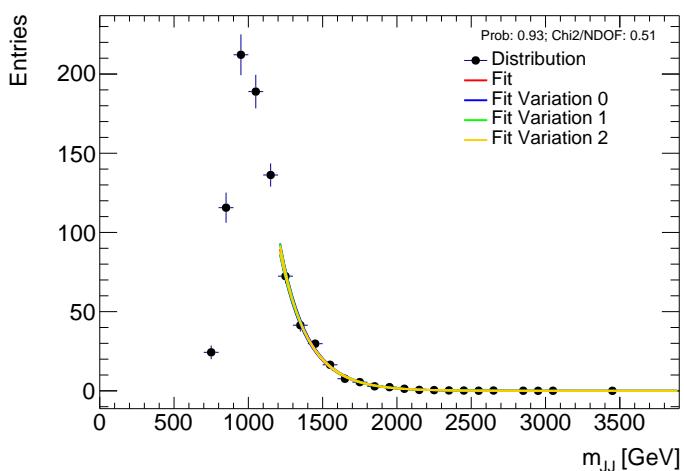
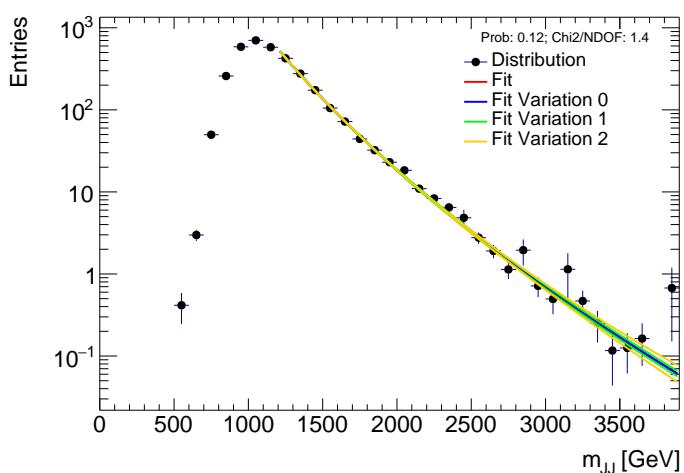
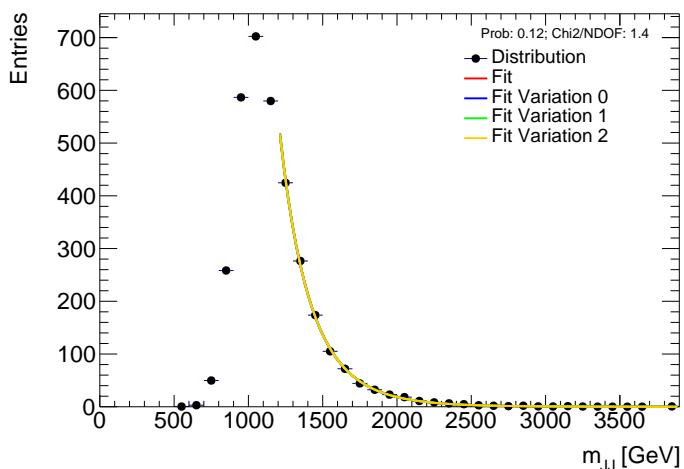
The final signal region prediction, using scaled di-jet mass distribution, with only statistical uncertainties, are shown in Figure 7.63(Figure 7.64) as before(after) smoothing.

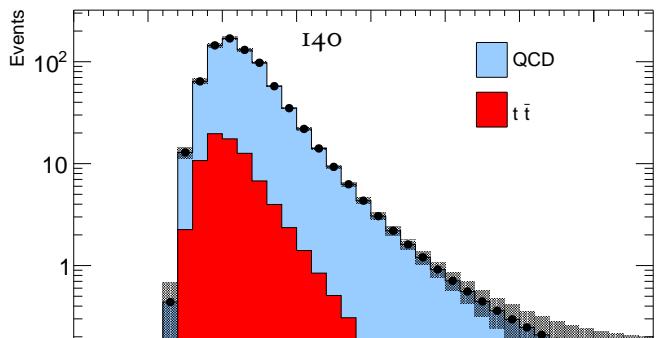
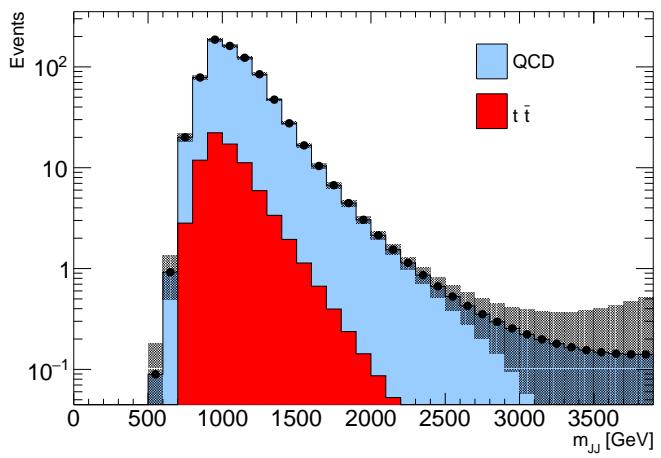
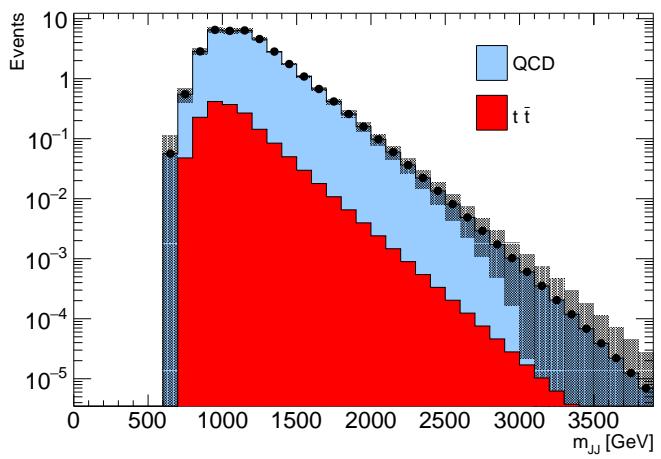
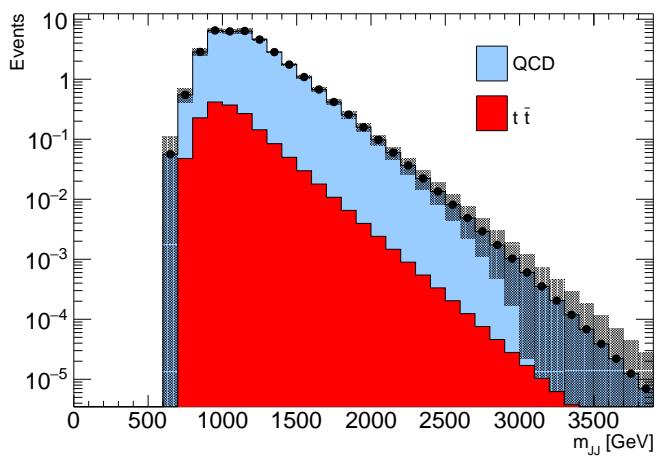


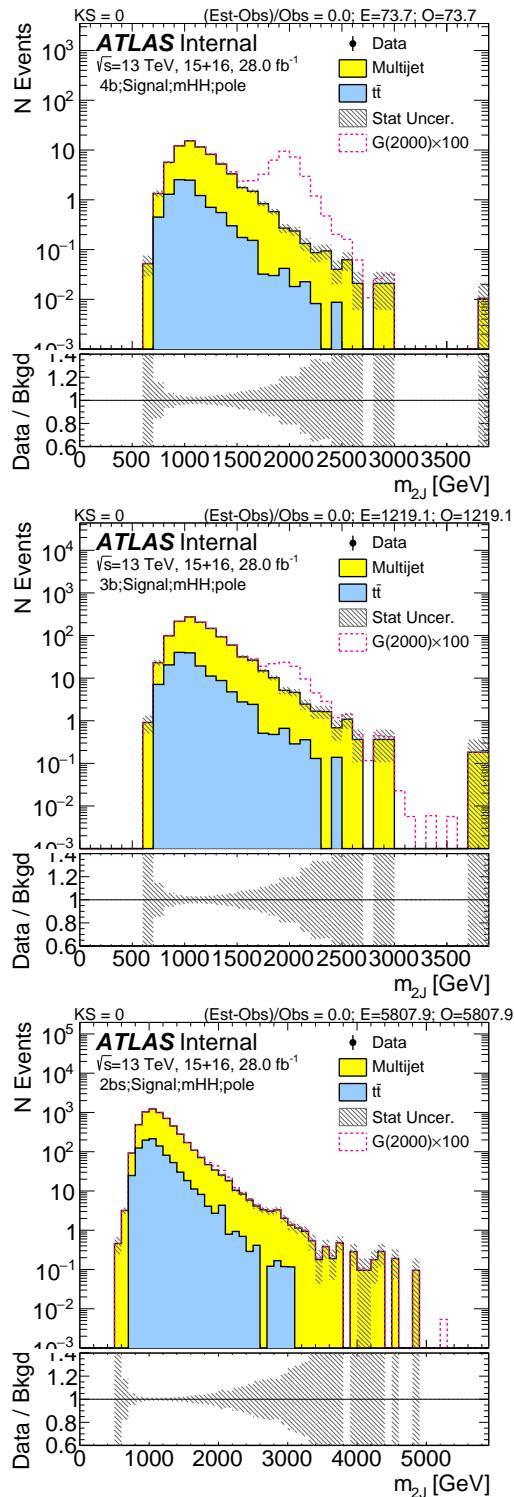
**Figure 6.60:** Normalized Scaled dijet mass distributions for the  $4b$  (top),  $3b$  (middle), and  $2bs$  (bottom) signal regions. For comparison, the unscaled distributions are shown on the same plot.





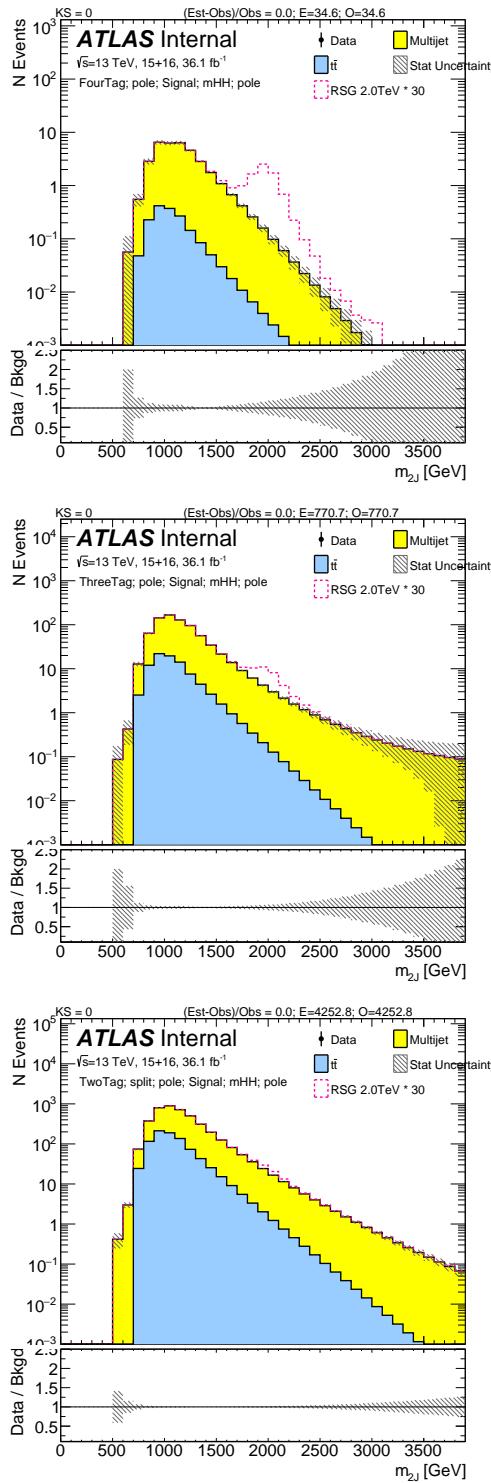






I41

**Figure 6.65:** Background prediction for  $4b$  (top),  $3b$  (middle), and  $2bs$  (bottom) signal region using scaled di-jet mass before smoothing. The uncertainty band includes only statistical uncertainties.



<sup>142</sup>  
**Figure 6.66:** Background prediction for  $4b$  (top),  $3b$  (middle), and  $2bs$  (bottom) signal region using scaled di-jet mass after smoothing. The uncertainty band includes only statistical uncertainties.

Region	$a_{t\bar{t}}$	$b_{t\bar{t}}$	$c_{t\bar{t}}$	$a_{qcd}$	$b_{qcd}$	$c_{qcd}$
FourTag	-2.02 ± 1.17	42.46 ± 9.87	1.31 ± 8.98	-0.49 ± 1.59	53.06 ± 15.2	-11.1 ± 12.8
ThreeTag	1.84 ± 1.17	42.45 ± 9.88	1.32 ± 8.98	8.51 ± 0.98	-13.8 ± 9.14	42.58 ± 7.87
TwoTag split	4.22 ± 1.17	42.45 ± 9.88	1.32 ± 8.98	7.06 ± 0.32	11.54 ± 2.77	19.05 ± 2.48

**Table 6.4:** Smoothing parameters in  $4b$  and  $3b$  and  $2bs$  signal regions for scaled mass distributions, the correlation between parameters is almost always 0.99.

*Every question has a proper answer. Every soul has a proper place.*

Tony.

# 7

## Background Estimation

### 7.1 BACKGROUND ESTIMATION

#### 7.1.1 OVERVIEW

The primary backgrounds to this analysis in the four  $b$ -jet signal region, in order of size, are QCD multi-jet production ( $\sim 95\%$ ),  $t\bar{t}$  ( $\sim 5\%$ ), and  $Z+jets$  ( $< 1\%$ ), where the percentages are the expected fraction of the background coming from each source. In the three  $b$ -jet signal region, the fractions are QCD ( $\sim 90\%$ ),  $t\bar{t}$  ( $\sim 10\%$ ), and  $Z+jets$  ( $< 1\%$ ). In the two  $b$ -jet split signal region, the fractions are QCD ( $\sim 80\%$ ),  $t\bar{t}$  ( $\sim 20\%$ ), and  $Z+jets$  ( $< 1\%$ ).

QCD is by far the dominant background. However, there is no reliable, high-statistics Monte Carlo simulation sample in this region of phase space (i.e with three or four  $b$ -jets collected into two

high- $p_T$  large radius jets) and thus a data-driven background estimation is needed. (See Appendix ??.) For the  $t\bar{t}$ background, Monte Carlo simulation samples of reasonable size are available, and thus can be used to guide an estimation of this background. The  $Z$ +jets background is small enough that we will rely on the Monte Carlo simulation of  $Z$ +heavy flavor jets.  $ZZ \rightarrow b\bar{b}b\bar{b}$  has been estimated to be completely negligible using a particle-level analysis, with less than one event expected after three  $b$ -tags are required, which will be further heavily suppressed by the  $X_{hh}$  requirement.

The QCD background prediction relies on finding a region which is similar enough in event properties that it can be used to estimate the shapes of the expected QCD background. This region is identical to the signal region defined by the full selection, with the exception that the events must have less  $b$ -tagged track jets:

- For the  $2bs$  category, the  $1b$  sample is used for modeling.
- For the  $3b$  and  $4b$  categories, the  $2b$  sample - where the two  $b$ -tagged trackjet are in the same large- $R$  jet - is used for modeling.

To prevent differences in the number of track jets from biasing the dijet mass distribution, the  $1b$ -tagged region requires that each large- $R$  jet has at least one track jet (to model  $2bs$ , ie.  $2b$  tag split). Similarly, the  $2b$ -tagged region requires that one large- $R$  jet has at least one track jet and the other one has at least two track jets (to model  $3b$ ), and each large- $R$  jet has at least two track jets (to model  $4b$ ).

However, this less  $b$ -tagged region only supplies the shapes of the expected background and not the total yield, and a second control sample, which we denote the *Sideband* region, is used to estimate the yield. The Sideband is obtained by doing the full analysis selection, except instead of the  $X_{hh}$  cut an alternative criteria on the large radius jet masses is used, that  $33 < R_{hh}$  and  $R_{hh}^{\text{high}} < 58 \text{ GeV}$ . To validate this approach, a third region, which we denote the *Control* region, is centered on the signal region in the plane of the two large radius jet masses but does not include the signal region, such that  $R_{hh} < 33 \text{ GeV}$ . The control region is used to validate the background estimations

before unblinding. The control and sideband regions are optimized, as shown in the following sections, to accurately estimate the rate of the QCD background (and thus allow for an extrapolation from the  $1b/2b$  estimate to a prediction in the  $4b/3b/2bs$  signal regions), whilst giving a control region which has kinematic properties similar to that of the signal region.

The  $t\bar{t}$  background shape is taken from MC. A data-based estimation of the  $t\bar{t}$  background yield is performed simultaneously with the QCD background yield estimation, by means of a binned likelihood fit. In the plane of the two leading large radius jet masses, the main contribution of the  $t\bar{t}$  background lies in the Sideband region. The data distribution in the Sideband region of the leading- $p_T$  large radius jet mass is fit simultaneously with the QCD shape estimate (from the less  $b$ -tagged sample) and with the  $t\bar{t}$  Monte Carlo shape. This fit is done separately in the  $4b$ ,  $3b$ , and  $2bs$  Sideband regions. From this fit, two terms are determined simultaneously:  $\mu_{QCD}$  and  $\alpha_{tt}$ .  $\mu_{QCD}$  is the ratio of the QCD event yield in the  $2bs/3b/4b$  regions to the amount in each corresponding less  $b$ -tagged region.  $\alpha_{tt}$  is the ratio of the fitted  $t\bar{t}$  event yield to the yield predicted from  $t\bar{t}$  MC. These two numbers are then used as multiplicative constants in other regions of the mass plane (i.e. the Control or Signal regions) to extrapolate from the rates of the less  $b$ -tagged regions to predictions of rates in the  $2bs/3/4$   $b$ -tagged regions, for estimating the amount of QCD, and to correct the rate of  $t\bar{t}$  production wrt. MC. Hence, the underlying assumption is that these scale factors are roughly constant over the 2D large- $R$  jet mass plane for Sideband/Control/Signal regions, which has been verified by performing these fits in small bins across the 2D mass plane. This is shown in Appendix ???. The correction factors are derived separately for the  $4b$ ,  $3b$ , and  $2bs$  regions.

In this section, we describe this approach in more detail and show its validation in data.

### 7.1.2 DEFINITION OF THE SIDEBAND AND CONTROL REGIONS (SB, CR)

The definitions of the SB, CR, and SR in the leading ( $m_J^{\text{lead}}$ ) and sub-leading ( $m_J^{\text{subl}}$ ) large- $R$  jet mass plane are found in Table 7.1. These regions can be seen in the leading and sub-leading large- $R$  jet

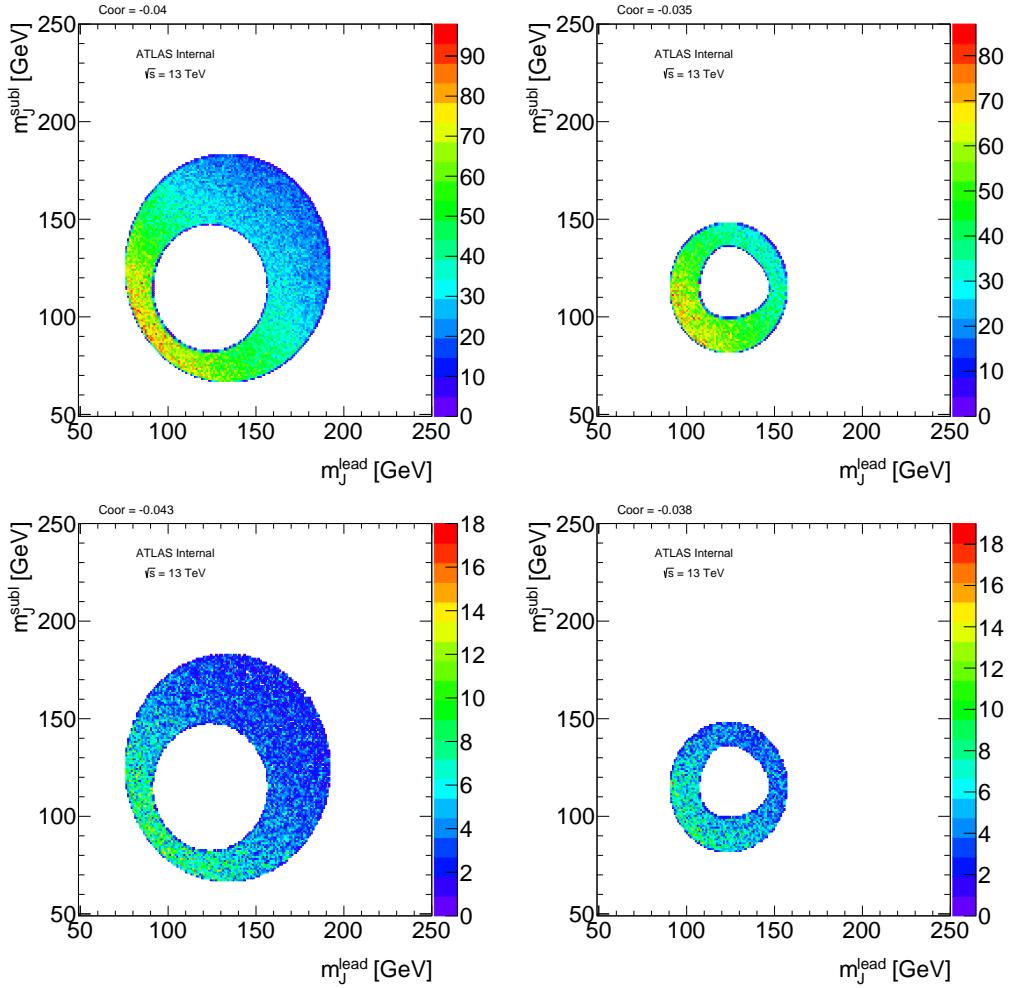
mass plane in Figure 7.1. As a reminder, the definition of  $X_{hh}$ ,  $R_{hh}$  and  $R_{hh}^{\text{high}}$  are:

$$X_{hh} = \sqrt{\left(\frac{m_J^{\text{lead}} - 124 \text{ GeV}}{\sigma(m_J^{\text{lead}})}\right)^2 + \left(\frac{m_J^{\text{subl}} - 115 \text{ GeV}}{\sigma(m_J^{\text{subl}})}\right)^2} R_{hh} = \sqrt{(m_J^{\text{lead}} - 124)^2 + (m_J^{\text{subl}} - 115)^2} R_{hh}^{\text{high}} = \sqrt{(m_J^{\text{lead}} - 134)^2 + (m_J^{\text{subl}} - 125)^2} \quad (7.1)$$

Region	Definition
Signal Region (SR)	$X_{hh} < 1.6$
Control Region (CR)	$R_{hh} < 33 \text{ GeV}$ and $X_{hh} > 1.6$
Sideband Region (SB)	$33 \text{ GeV} < R_{hh} \text{ and } R_{hh}^{\text{high}} < 58 \text{ GeV}$

**Table 7.1:** Definitions of the Signal (SR), Sideband (SB) and Control (CR) regions.

The CR is chosen to be as close as possible to the signal region, thus allows a good test for the background predictions, avoids the top mass peak around 175 GeV, and still gives reasonably statistics. The SB definition is optimized so as to also be a reasonable proxy for the events contained in the CR and SR. Being farther from the SR means that the exact kinematics will not be the same, but one can avoid very large and very small mass jets not present in the SR by appropriate choice of SB. The optimization of the CR and SB definitions can be found in Appendix ???. The choice of SB's impact on the predicted QCD background normalization, which is derived from the SB, can be found in Appendix ??.



**Figure 7.1:**  $m_J^{\text{lead}}$  vs.  $m_J^{\text{subl}}$  in data in the 1 $b$ -tag (top) and 2 $b$ -tag (bottom) selection, the plots show the boundary between the Sideband (left) and Control (right) regions.

### 7.1.3 QCD MULTI-JETS

The QCD multi-jets prediction relies on finding a region which is similar enough in event properties so that it can be used to estimate the shapes of the expected background. This region is defined to be identical to the signal region except requiring both of the large- $R$  jets to pass the  $\geq 2$  (like in  $4b$ ) or  $\geq 1/2$  (like in  $3b$ ) or  $\geq 1$  (like in  $2b$  split) track jet requirement, but have two or one associated  $b$ -tagged track jets only on one of the large- $R$  jets. However, this  $1b/2b$  region only provides shapes of the expected background and not the total yield.

It should be noted that the  $1b/2b$  region is orthogonal to the  $4b/3b/2bs$  signal regions. In addition, the MC predicted  $t\bar{t}$  events in the  $1b/2b$  regions are subtracted from the data to produce the  $1b/2b$  QCD estimation. This procedure follows closely the method used in Run 1 and also used in the resolved analysis, but requiring  $1b$ -tag for the  $2bs$  background estimation.

It should also be noted that the resolved veto will impact the  $4b$  background estimation. Specifically,  $2b$  events are excluded when they have at least two resolved jets that are  $b$  tagged (passing resolved 70% working point) and passing the resolved  $X_{hh} < 1.6$  cut if using two other non  $b$ -tagged resolved jets to make the Higgs candidates. This ensures that a similar sculpting effect is reflected in the background estimation, and a check for this can be found in Appendix ??.

Given the  $1b/2b$  samples, which predict shapes for the QCD background, the normalization of the QCD background is determined in the sideband by fitting the leading jet mass distribution simultaneously with QCD and  $t\bar{t}$  background templates, as described in section 7.1.5. This fit gives a scaling factor for QCD, called  $\mu_{QCD}$  (and for  $t\bar{t}$ , called  $\alpha_{t\bar{t}}$ ) which can be applied to scale the  $1b/2b$  predictions in the CR or SR to the predicted normalizations in those regions.

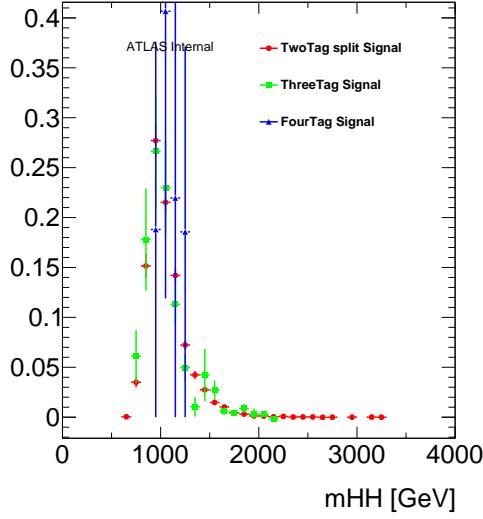
It should be noted that there can be kinematic differences between the  $1b/2b$  samples and the  $4b/3b/2bs$  regions. Thus a kinematic reweighting is applied to correct for such differences, as described in Section 7.1.6.

#### 7.1.4 $t\bar{t}$ BACKGROUND

The number of  $t\bar{t}$  events in the signal region coming mainly from the all-hadronic decay mode (with a smaller contribution from the leptonic + jets decay mode) comprises of around 5-20% of the inclusive total background in the  $4b/3b/2bs$  regions due to the high  $p_T$  threshold imposed on the leading *large-R* calorimeter jet. In addition, the normalization and the shape of  $t\bar{t}$  events in the sideband region can affect the QCD estimate described in the previous section.

For the normalization of the  $t\bar{t}$  background, we start with the MC prediction estimated by scaling the MC sample by the cross section and luminosity, and then applying the boosted event selection. To account for possible differences between data and MC, a normalization scaling factor is derived from a fit to data in the sideband region. Although estimated in the SB, this normalization scaling factor will be applied also in the CR and SR. Separate fits are done in the  $4b/3b/2bs$  SB, thus deriving separate normalization scaling factor for the  $4b/3b/2bs$  samples. For the shape of the  $t\bar{t}$  background, no data driven methods were identified, and thus the MC shape is used.

However, it should be noted that in the  $4b$  and  $3b$  signal region, there were not sufficient MC statistics to get a reasonable shape estimate. As a result, in the  $4b$  and  $3b$  signal region, the  $2bs$  shapes will be used (but the normalization will still be that estimated for the  $4b/3b$  sample). Since the same shape is used for the  $4b/3b/2bs$  SR predictions of  $t\bar{t}$ , the shape systematics are considered correlated in the final results and limit setting. A comparison between the  $4b/3b/2bs$  shapes for the di-large- $R$ -jet mass distributions (the final discriminant) in the SR can be found in Figure 7.2. As we can see, the shapes are compatible, with the  $4b$  having much larger statistical uncertainties. Differences between these distributions will be used as a systematic, as described in Section 8.



**Figure 7.2:** comparison between the  $2b$ ,  $3b$ , and  $4b$  shapes for the di-large- $R$ -jet mass distributions (the final discriminant) in the SR.

### 7.1.5 FITTING PROCEDURE FOR QCD AND $t\bar{t}$ NORMALIZATION

The number of  $4b/3b/2bs$  events in data observed in a given region (SB / CR / SR) can be described as:

$$N_{\text{data}}^b = \mu_{\text{qcd}}^b N_{\text{qcd}}^{xb} + \alpha_{t\bar{t}}^b N_{t\bar{t}}^{b} + N_{Z+jets}^b \quad (7.2)$$

where  $\nu_b$  is the number of  $b$ -tagged track jets required,  $x$  is 1 for  $2bs$  and 2 for  $3b$  and  $4b$ .  $\mu_{\text{multijet}}$  is essentially an estimate of the ratio of the number of QCD events with  $\nu_b$   $b$ -tagged track jets, to the number of  $1b/2b$  QCD events, while the  $t\bar{t}$  normalization parameter  $\alpha_{t\bar{t}}$  applied after the  $t\bar{t}$  is scaled to the total integrated luminosity, is a correction to the MC prediction in this phase space. The same equation can be applied to the  $4b/3b/2bs$  region (replacing  $\nu_b$  by  $4b/3b/2s$   $b$  in Equation 7.2).

In order to constrain the QCD and  $t\bar{t}$  background normalizations using data, a simultaneous fit

is applied to extract both the  $t\bar{t}$  normalization with respect to the yields from simulation and the number of  $1b/2b$  data events for the QCD background. These scaling parameters are determined independently for the  $4b/3b/2s$  signal regions. But as the procedure is the same for those three signal regions, we denote these scaling factors simply  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$  in the following text.

A binned maximum likelihood fit is employed to find the values of  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$ , as well as the correlation between the two parameters. The fit is performed on the leading- $p_T$  jet mass spectrum in the sideband region, as it has the best separation between QCD and  $t\bar{t}$  shapes. Due to the  $p_T > 450$  GeV cut imposed on the leading  $large - R_{\text{jet}}$ , the hadronic top quark is likely to be fully reconstructed inside of the  $large - R_{\text{jet}}$  and the leading jet mass in the  $t\bar{t}$  sample has a clean peak around  $M = 170$  GeV in the sideband region.

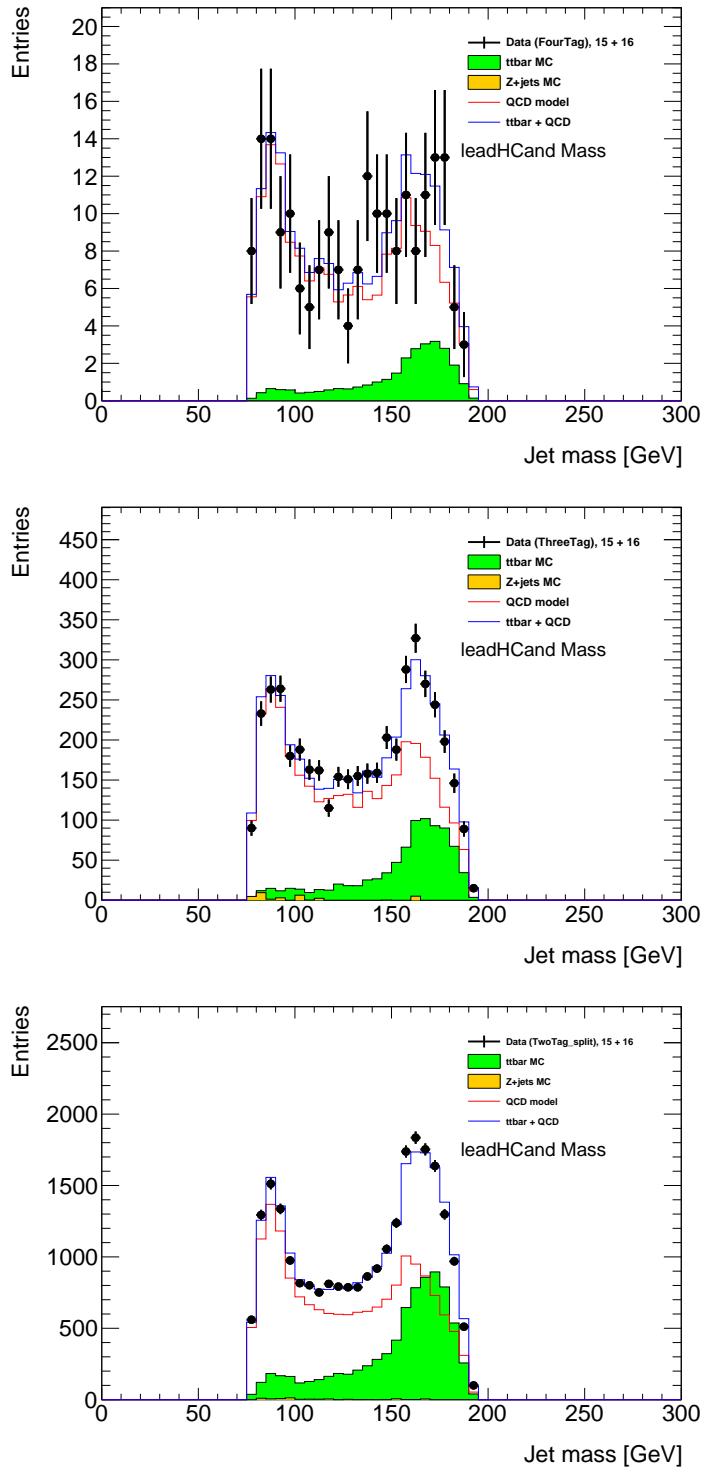
The values of  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$  as estimated by the fits in the  $4b/3b/2bs$  sideband regions can be found in Table 7.2, along with the correlation of the fitted parameters  $\xi(\mu_{qcd}, \alpha_{t\bar{t}}) = \frac{\text{Cov}(\mu_{qcd}, \alpha_{t\bar{t}})}{\sqrt{\text{Cov}(\mu_{qcd}) \text{Cov}(\alpha_{t\bar{t}})}}$ .  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$  are approximately 70% negatively correlated, which is not surprising as they are the only two components fit to the data distribution and their sum needs to predict the SB total event count.

Sample	$\mu_{qcd}$	$\alpha_{t\bar{t}}$	$\xi(\mu_{qcd}, \alpha_{t\bar{t}})$
FourTag	$0.0332 \pm 0.00428$	$0.891 \pm 0.599$	-0.785
ThreeTag	$0.163 \pm 0.00434$	$0.8 \pm 0.0733$	-0.72
TwoTag split	$0.0627 \pm 0.000573$	$0.986 \pm 0.0186$	-0.47

**Table 7.2:** Background scaling parameters ( $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$ ) estimated from fits to the leading jet mass distributions in  $4b/3b/2bs$  sideband regions.  $\xi(\mu_{qcd}, \alpha_{t\bar{t}}) = \frac{\text{Cov}(\mu_{qcd}, \alpha_{t\bar{t}})}{\sqrt{\text{Cov}(\mu_{qcd}) \text{Cov}(\alpha_{t\bar{t}})}}$

Figure 7.3 shows the post-fit spectrum of the leading  $large - R_{\text{calorimeter}}$  jet mass in the  $4b/3b/2bs$  sideband regions. The normalization of  $t\bar{t}$  is constrained by the top quark mass peak around 170 GeV. The shapes of the data is also well modeled by the predicted background. The fitting errors on

$\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$  are applied as systematic uncertainties taking into account their correlation. This will be explained in more detail in the systematics section.



**Figure 7.3:** Simultaneous fit of  $\mu_{\text{multijet}}$  and  $\alpha_{\bar{t}t}$  in  $4b$  (top) and  $3b$  (middle) and  $2b$  (bottom) sideband region using leading large –  $R$  calorimeter jet mass spectrum. 154

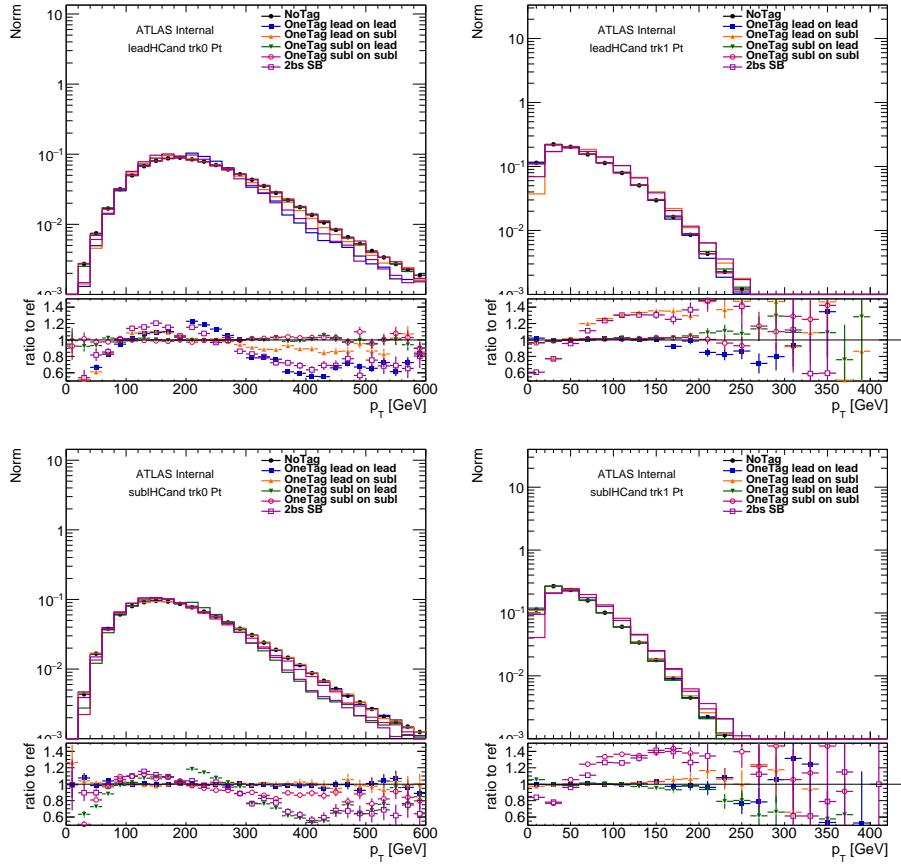
### 7.1.6 KINEMATIC REWEIGHTING

Due to the large contribution from the completely data-driven QCD background, it is important to model this background as good as possible in all regions of the analysis. Using the  $1/2b$  region to model the  $2bs$ ,  $3b$ , and  $4b$  regions can introduce discrepancies in the modeling of the estimated QCD background versus the real  $n b$  data. These discrepancies arise possibly from the non-trivial effect that  $b$ -tagging has on jet kinematics. The natural choice of reweighting variable is the  $p_T$  of the track jets in the event, since these are the objects that we apply the  $b$ -tagging to. Also, large- $R$  jet  $p_T$  is also reweighted to account for the effect from light and charm quark composition difference at different energy scales. The three chosen variables are the leading large- $R$  jet  $p_T$ , leading large- $R$  jet leading trackjet  $p_T$  and subleading large- $R$  jet leading trackjet  $p_T$ .

In order to account for the  $b$ -tagging effect, a reweighting on the  $1/2b$  data is adopted. The basic idea is to reweight the non  $b$ -tagged Higgs candidate to have kinematic distributions just like a  $b$ -tagged Higgs candidate. The idea is demonstrated in Figure 7.4. It shows that  $2bs$  has very similar kinematic distributions on the trackjet  $p_T$  as the  $1b$  sample, when the variable is the  $b$  tagged trackjet.

For  $2bs$ , the  $1b$  non-tagged Higgs candidate is reweighted to be like a  $1b$  tagged Higgs candidate; for  $3b$ , the  $2b$  non-tagged Higgs candidate is reweighted to be like a  $1b$  tagged Higgs candidate; for  $4b$ , the  $2b$  non-tagged Higgs candidate is reweighted to be like a  $2b$  tagged Higgs candidate. For each category, the events are split into two orthogonal subgroups, depending on whether leading/subleading Higgs candidate is  $b$ -tagged, the event is then reweighted such that the untagged Higgs candidate's distribution behaves like the corresponding  $b$ -tagged Higgs candidate's.

To avoid potential biases in the final distributions used for the analysis, a reweighting technique is applied to the  $1/2b$  data only. Since each signal region is modeled by a different  $1/2b$  tag category:  $2bs$  by  $1b$  tag events with at least 1 track jets on both large- $R$  jets,  $3b$  by  $2b$  tag events with at least one track jets on one large- $R$  jet and at least two track jets on the other large- $R$  jet, and  $4b$  by  $2b$  tag events



**Figure 7.4:** Comparison of different trackjet  $p_T$  distributions. Top row is for leading  $p_T$  Higgs candidate, and bottom row is for subleading  $p_T$  Higgs candidate. Left column is for the leading  $p_T$  trackjet of the Higgs candidate, and right column is for the subleading  $p_T$  trackjet of the Higgs candidate. Shown in the plot are just data distributions, inclusive of SB, CR, and SR regions for  $0b$  and  $1b$ , while for  $2bs$  only the SB region is shown.  $1b$  sample is further split into four subcategories, depending on which trackjet gets  $b$  tagged. OneTag lead on lead means the  $b$  tagged trackjet is the leading trackjet of the leading Higgs candidate, OneTag lead on subl means the  $b$  tagged trackjet is the subleading trackjet of the leading Higgs candidate, OneTag subl on lead means the  $b$  tagged trackjet is the leading trackjet of the subleading Higgs candidate, and OneTag subl on subl means the  $b$  tagged trackjet is the subleading trackjet of the subleading Higgs candidate. At the bottom ratio plot, all the ratio are taken with respect to the  $0b$  tagged distribution.

with at least two track jets on both large- $R$  jets, the reweighting procedure is the same but orthogonal for the three different channels. Note the  $2b$  sample is already split into separate parts, as described in paragraph 7.1.3.

The detailed procedure is listed as follows:

- Subtracting  $1/2b$  tag  $t\bar{t}$  and  $Z+jets$  samples in the sideband from the  $1/2b$  tag data in the Sideband + Control + Signal regions to get the  $1/2b$  QCD inclusive estimate.
- Separate the  $1/2b$  tag sample further to sample A. that has the  $b$ -tagged Higgs is the leading  $p_T$ Higgs candidate, and B. that the  $b$ -tagged Higgs is the subleading  $p_T$ Higgs candidate.
- For each variable, i.e. the large- $R$  jet  $p_T$ , normalize sample A to sample B total number of events, take the ratio of sample A distribution over sample B distribution, and fit the ratio with a spline function. (TSpline3)
- Use this functional form to extract reweighting values for each variable that is considered. The reweighting value for each variable is also constrained to be within a  $-30\%$  to  $+40\%$  range compared to one, to avoid over corrections and failed fit situations.
- For each event, all the weights are multiplied together to change the  $1/2b$  tag data event weight. Another constraint is applied, such that each total reweighting value is constrained to be within a  $10\%$  to  $+1000\%$  range compared to one, again to avoid over corrections.
- The reweighting is done on the three variables: large- $R$  jet  $p_T$  and the two track jet  $p_{T\gamma}$ s, which is counted as one iteration of reweighting.
- A total of ten iterations are used to stabilize the reweighting. The reweighting is roughly converging after three iterations.

For reweighting method comparisons and validations in data and Dijet MC, see Appendix ??.

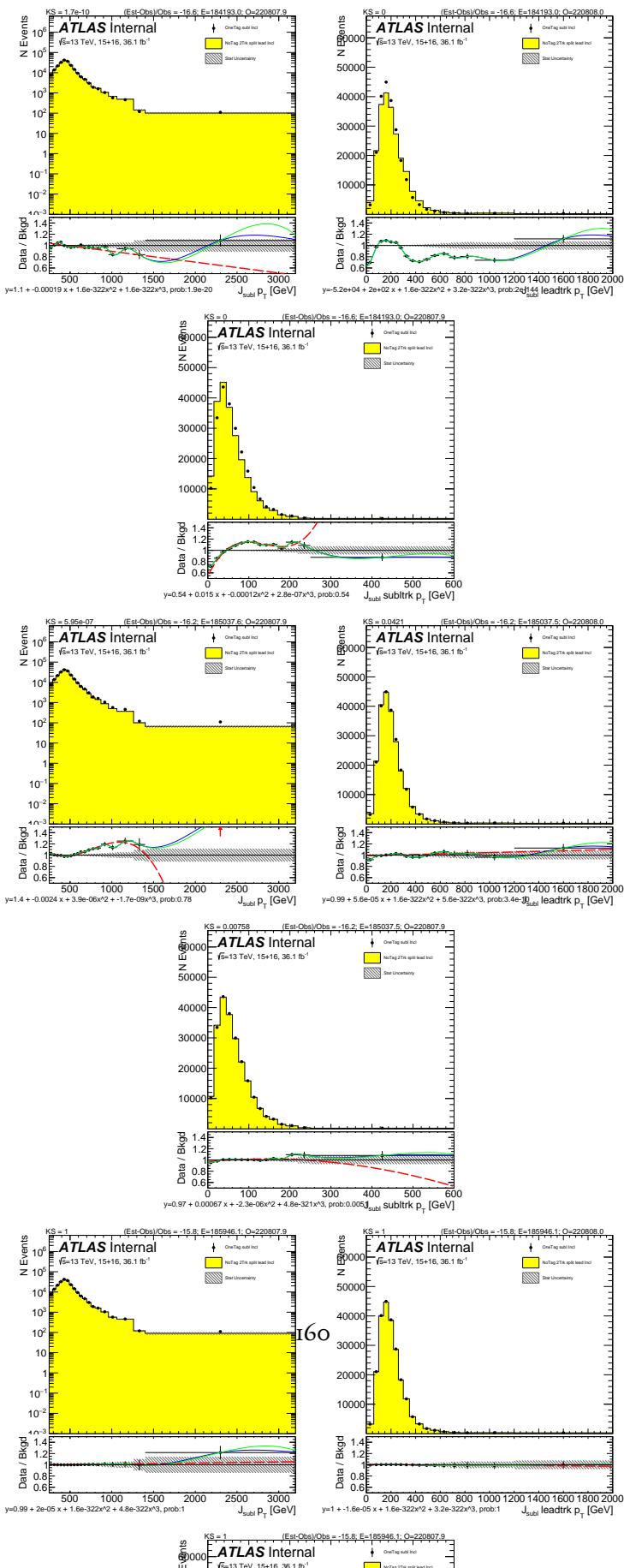
## REWEIGHTING FITS

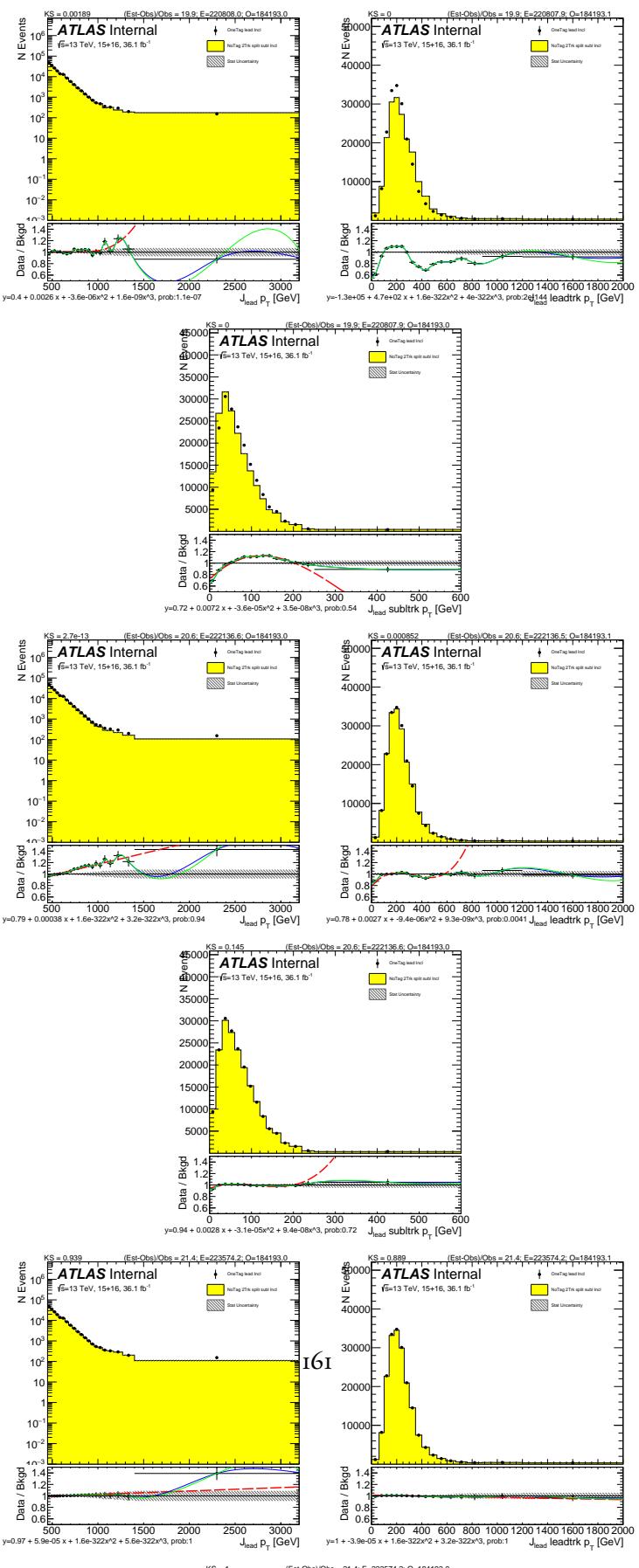
The first iteration, second iteration, and last iteration of fits for  $2b$ s, where in  $1b$  data, the non-tagged Higgs candidate are reweighted to be like a  $1b$  tagged Higgs candidate, can be seen in Figure 7.5 and 7.6. Similar distributions for  $3b$ , where in  $2b$  data, the non-tagged Higgs candidate are reweighted to be like a  $1b$  tagged Higgs candidate, are shown in Figure 7.7 and 7.8. Similar distributions for  $4b$ , where in  $2b$  data, the non-tagged Higgs candidate are reweighted to be like a  $2b$  tagged Higgs

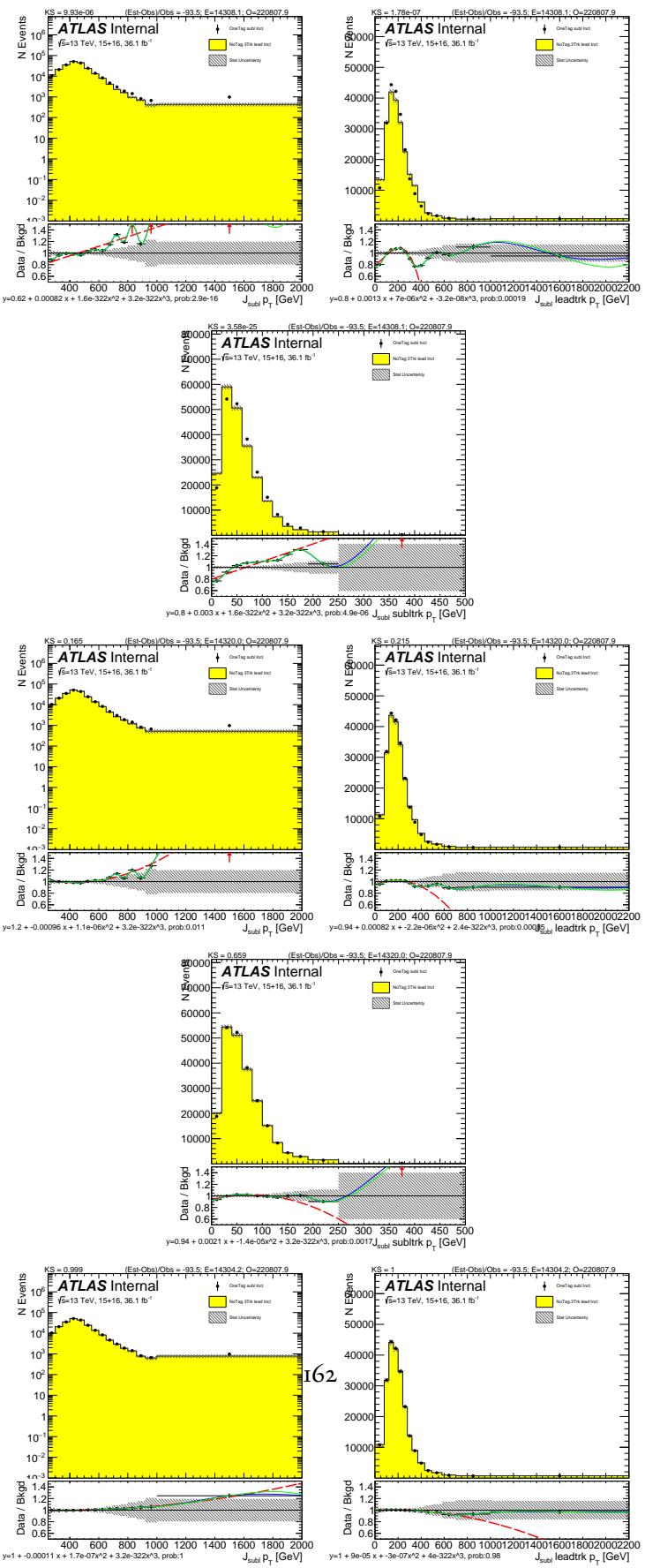
candidate, are shown in Figure 7.9 and 7.10. The before reweighting distribution (first row), the reweighting result after the first iteration (second row), and the final distribution after reweighting (last row) are presented.

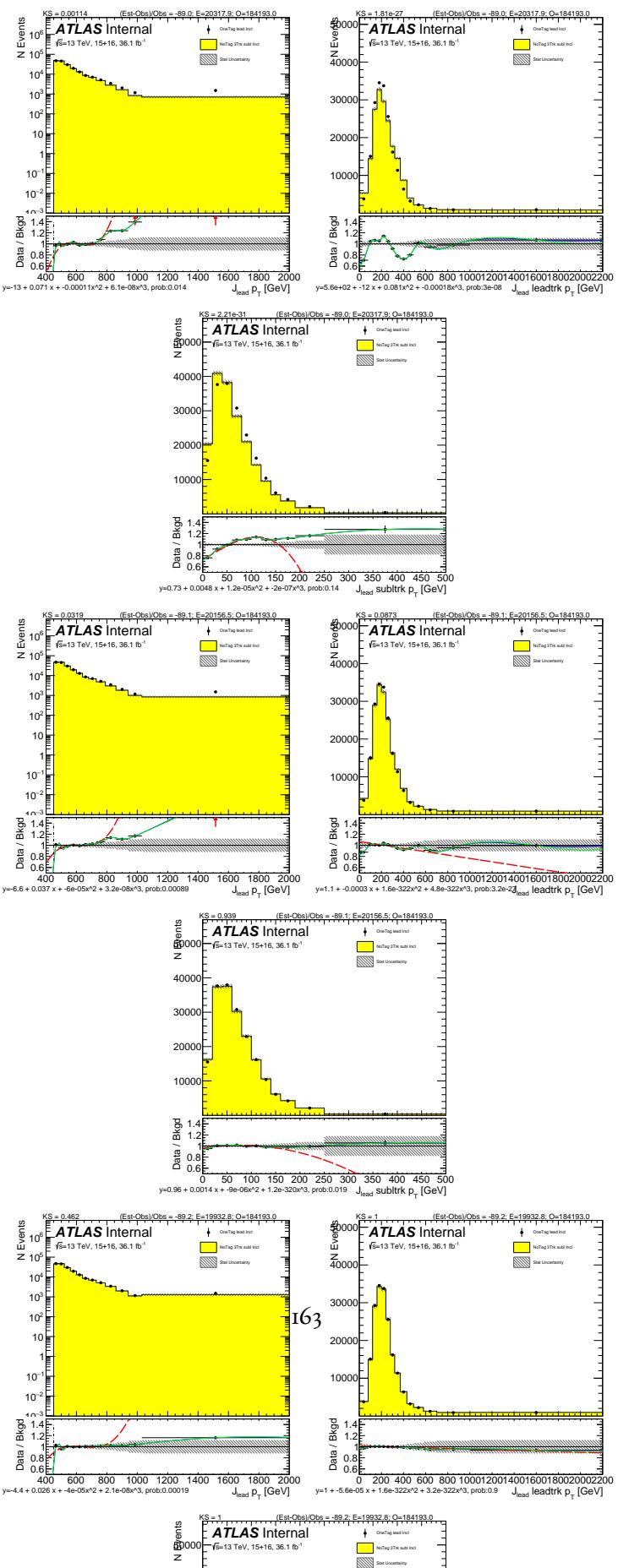
It should be noted that in some plots, like Figure 7.9 and 7.10, the last ratio bin sometimes still doesn't converge to unity. This is a feature from the limited statistics from the last bin, especially in the  $4b$  case, where only 20% number of events in  $2b$  is used for background prediction and therefore reweighted. One could choose a different binning and use more iterations to help this converge to one, yet the last bin's few event will also likely to end up with a large unphysical weight and therefore harm the background prediction later.

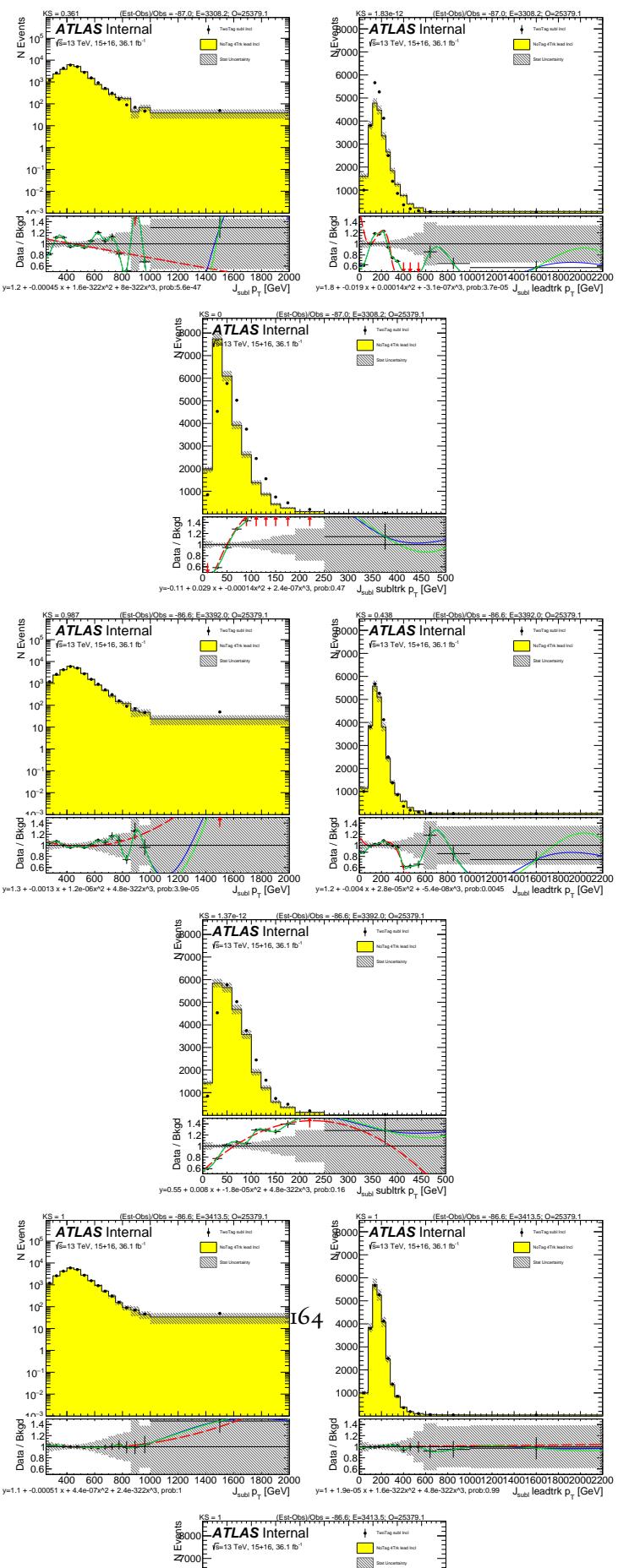
For the distribution of weights and the weight as a function of different kinematic ranges, see Appendix ??.

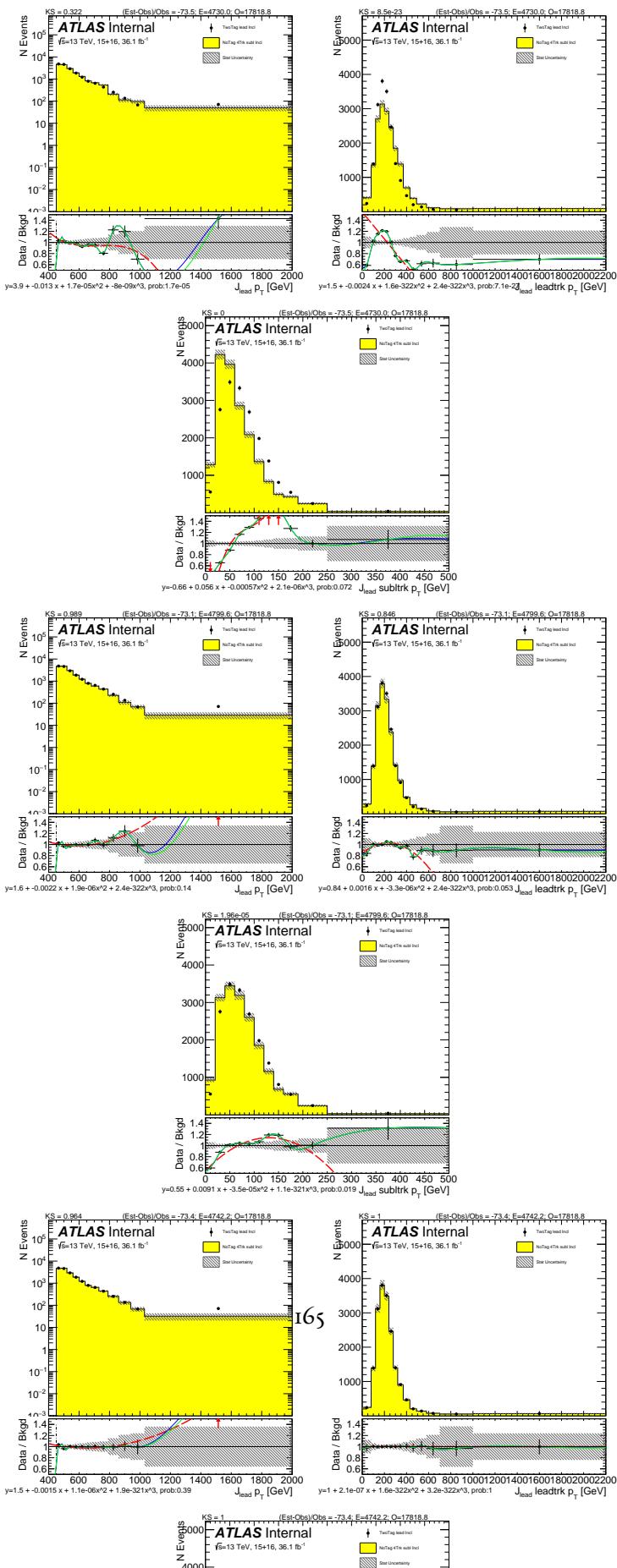






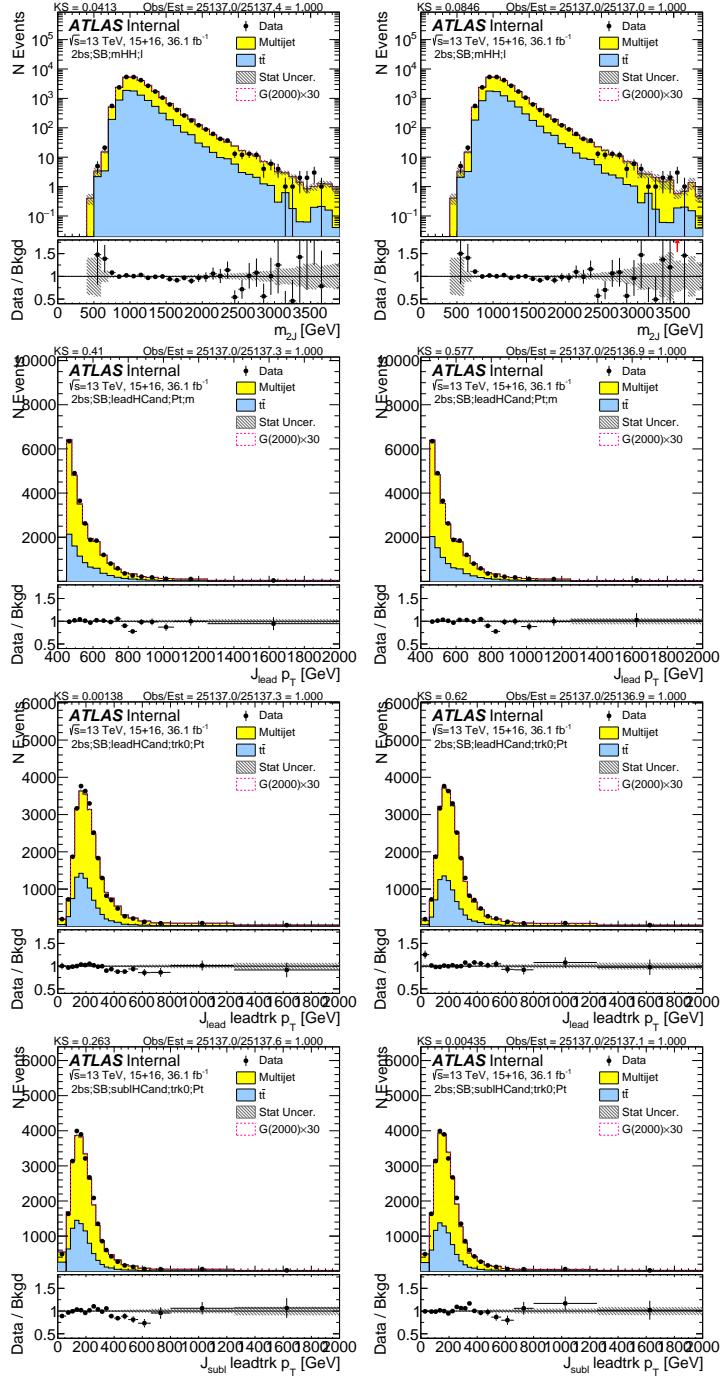




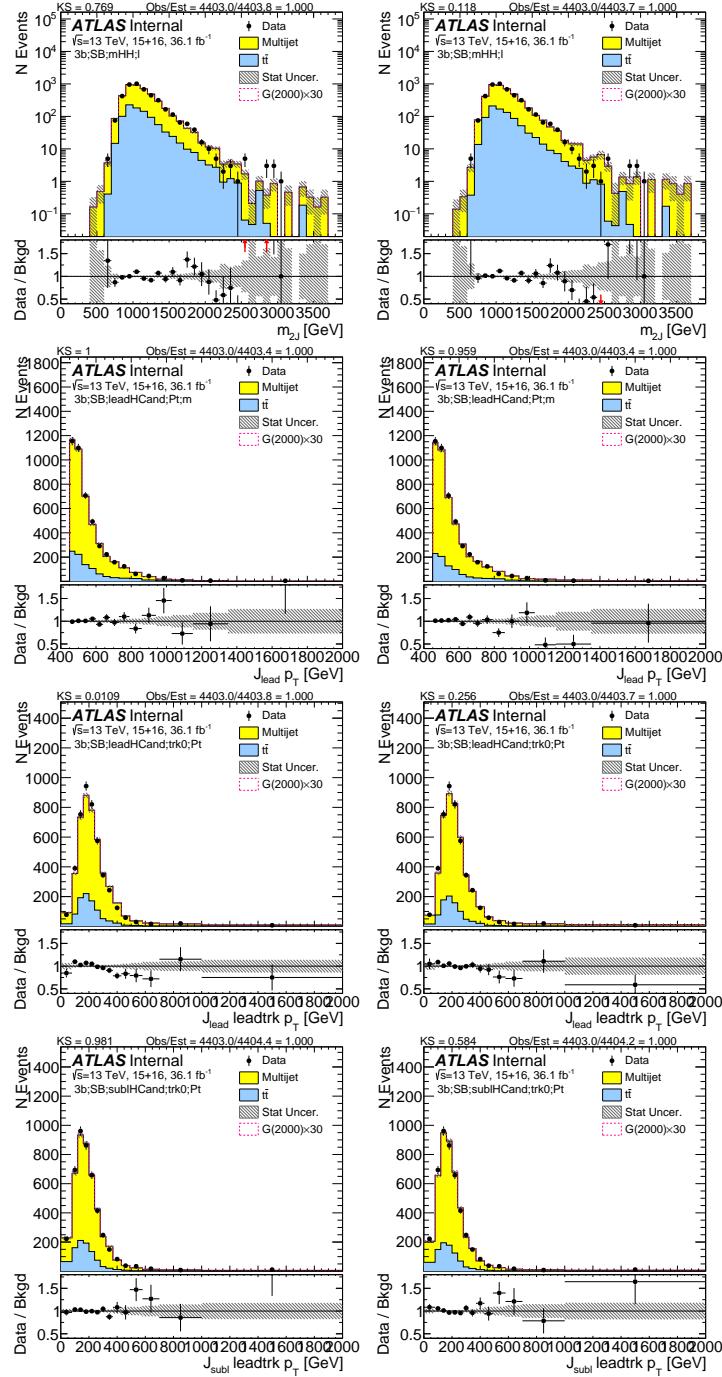


## REWEIGHTING RESULTS COMPARISON IN SIDEBAND AND CONTROL REGION

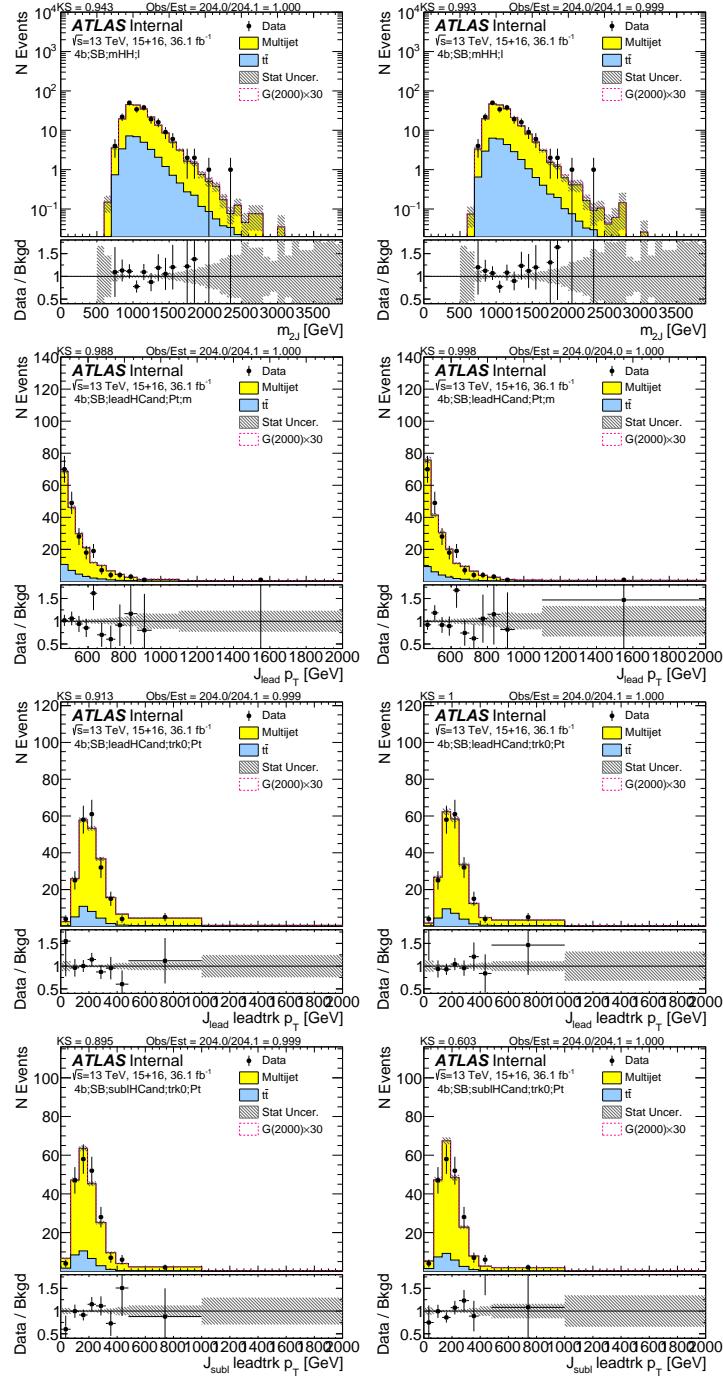
A comparison of the Sideband shapes before and after reweighting for  $2bs$ ,  $3b$  and  $4b$  can be seen in Figures 7.11, 7.12, and 7.13. Also, a comparison of the Control Region shapes before and after reweighting for  $2bs$ ,  $3b$  and  $4b$  can be seen in Figures 7.14, 7.15, and 7.16. In almost all cases, both the reweighted/non-reweighted prediction agrees fairly well with the data, and the reweighted plots' KS score improved from non-reweighted distributions.



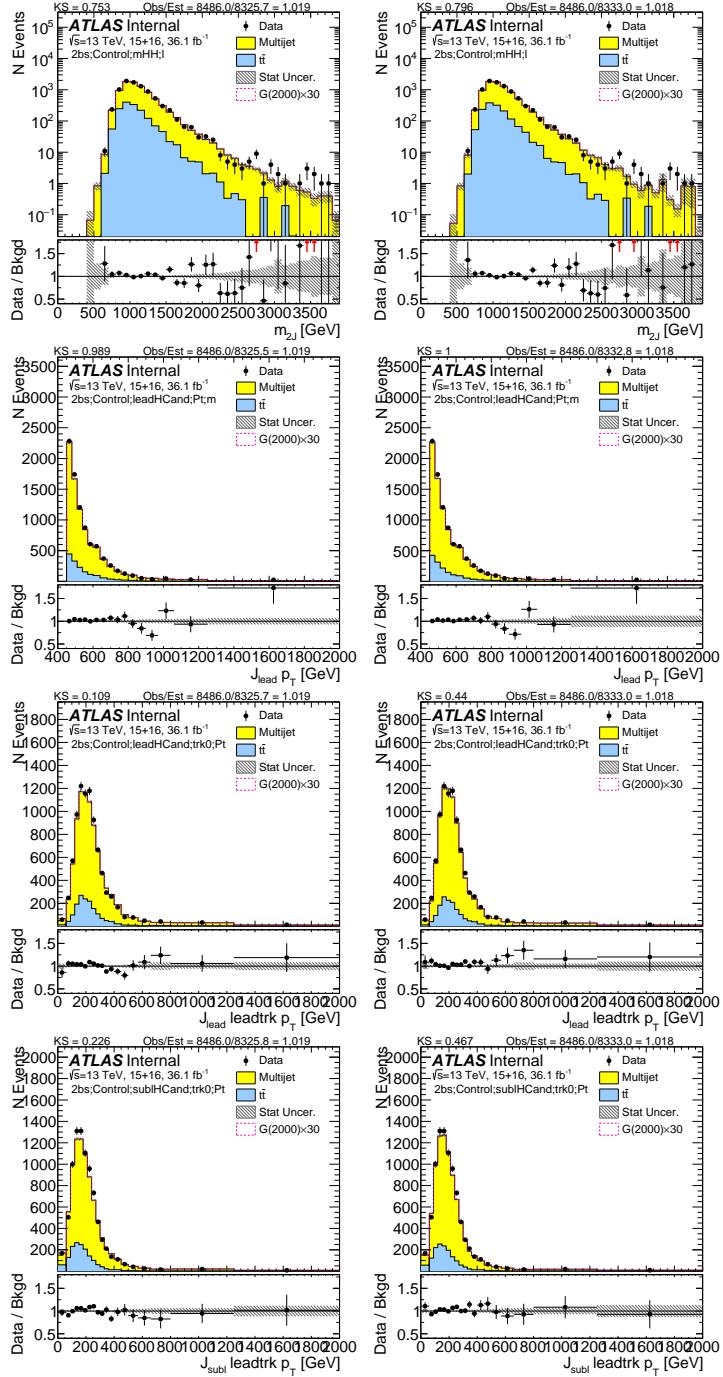
**Figure 7.11:** Reweighted 2bs Sideband region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



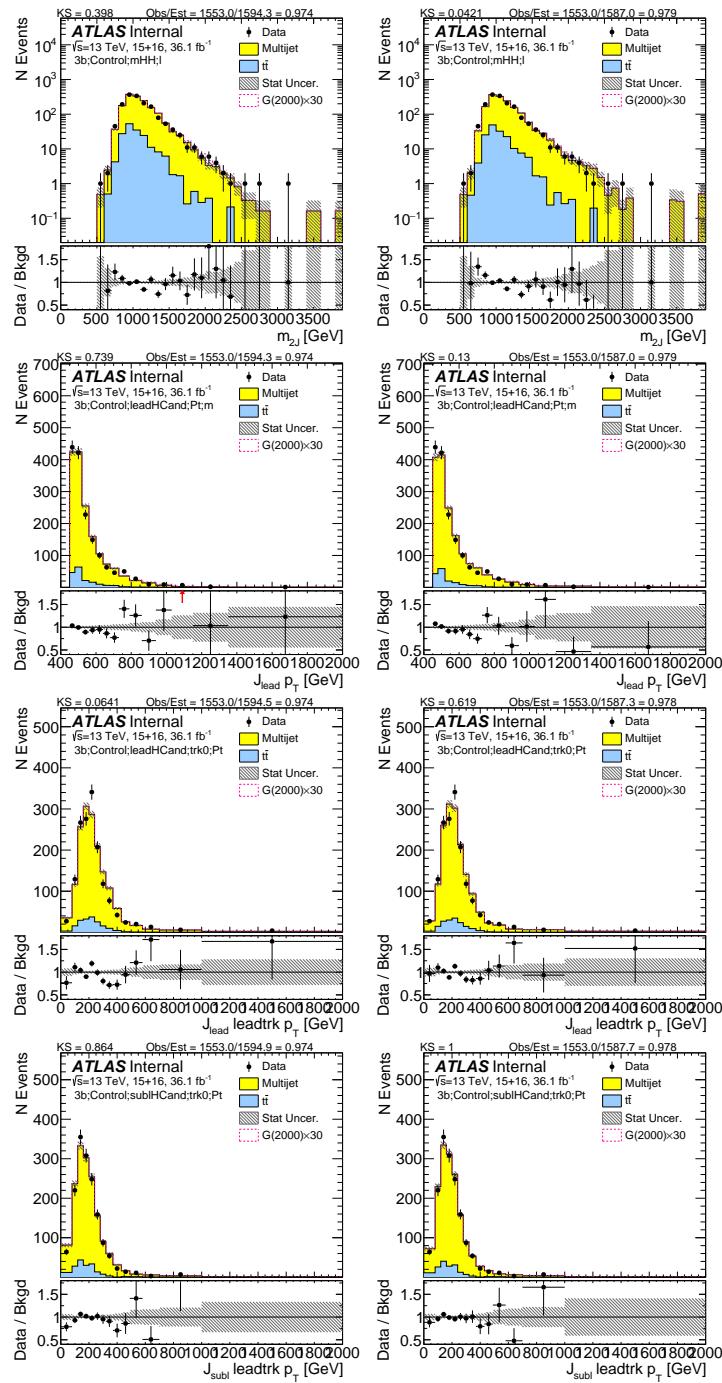
**Figure 7.12:** Reweighted 3b Sideband region predictions comparison. Top row is the dijet Mass, second row is lead-ing large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



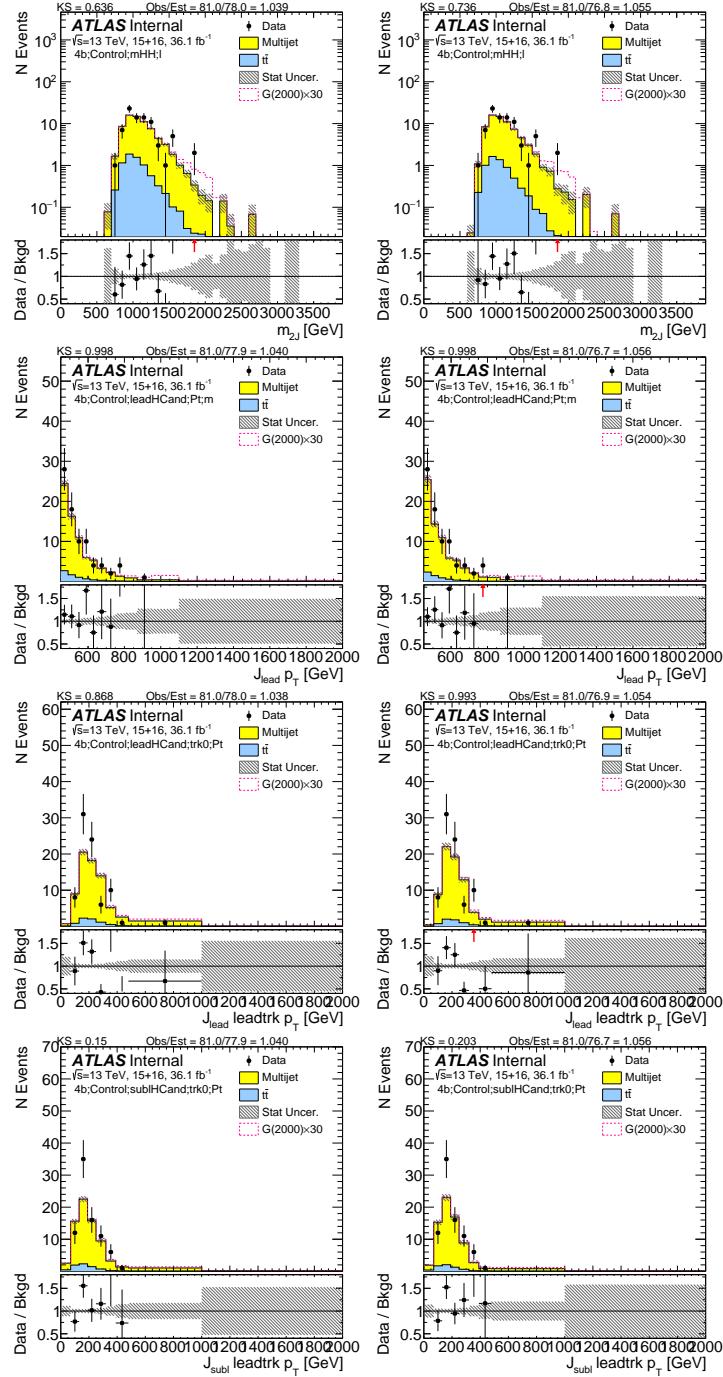
**Figure 7.13:** Reweighted  $4b$  Sideband region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



**Figure 7.14:** Reweighted 2bs Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



**Figure 7.15:** Reweighted  $3b$  Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.



**Figure 7.16:** Reweighted 4*b* Control region predictions comparison. Top row is the dijet Mass, second row is leading large- $R$  jet  $p_T$ , third row is the leading large- $R$  jet's leading trackjet  $p_T$  and the last row subleading large- $R$  jet's leading trackjet  $p_T$ . On the left are the distributions before reweighting, and on the right are the distributions after reweighting.

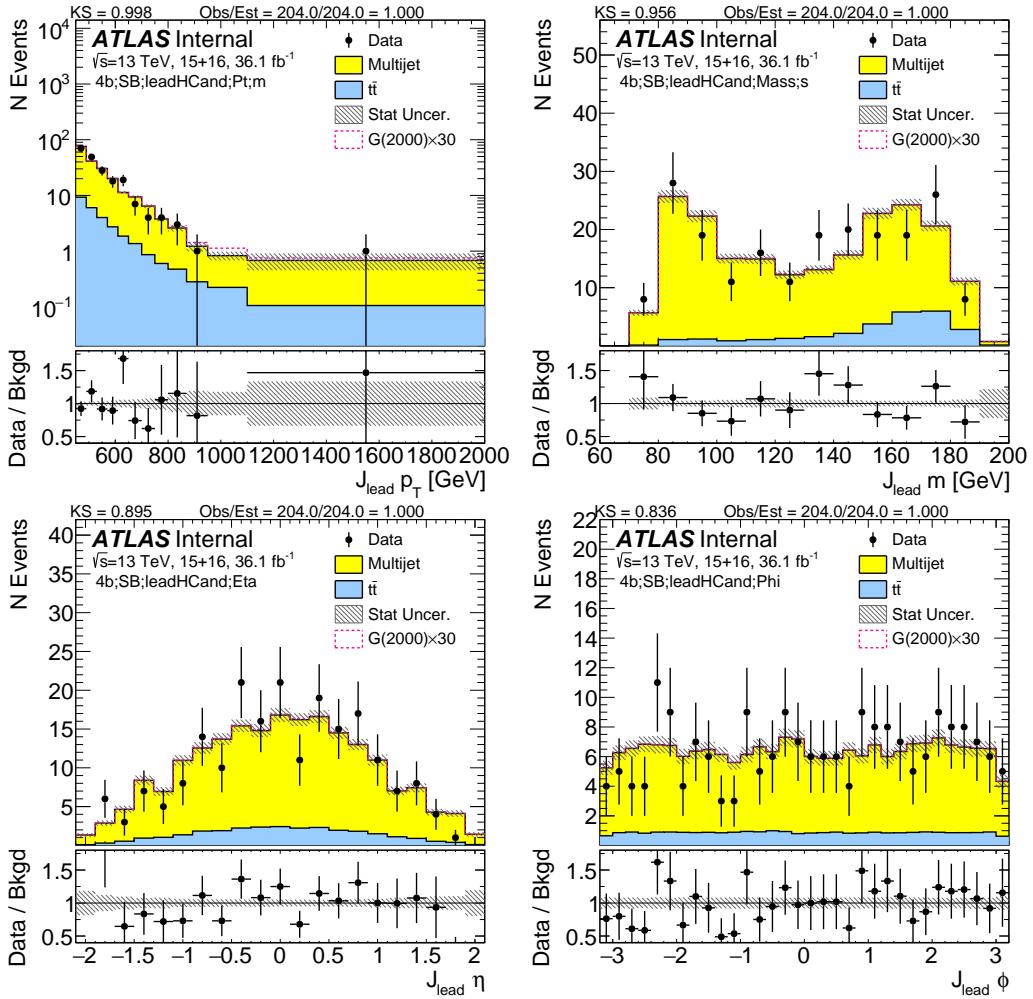
### 7.1.7 PREDICTIONS IN THE SIDEBAND REGION (SB)

This section shows comparisons of data with the prediction of QCD multi-jets and  $t\bar{t}$  in the sideband region (SB), which is identical to the signal region (SR) except the large- $R$  jets are required to have masses far from the Higgs mass. The definition of the sideband and control regions is discussed in Section 7.1.2. The predicted and observed event yields are summarized in Tables ?? and ??.

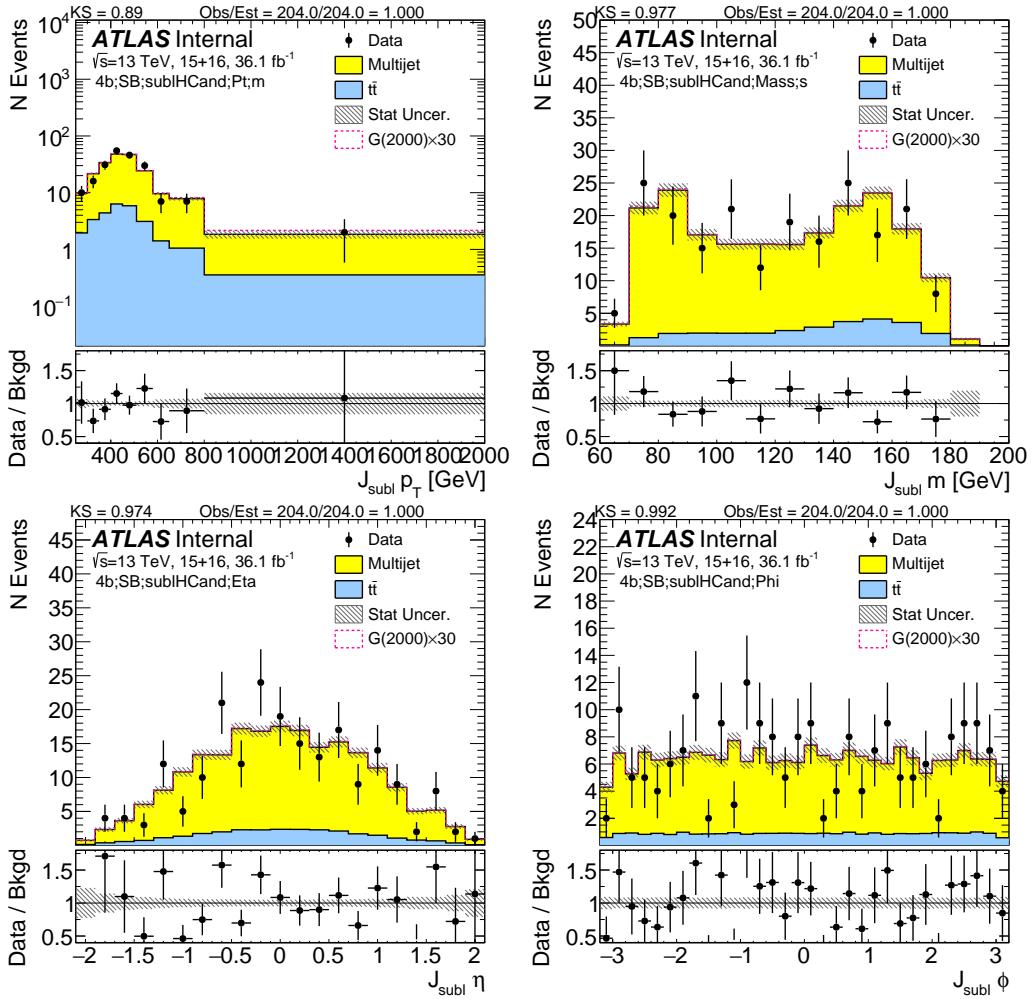
Figures 7.17, 7.18, 7.19, and 7.20 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $4b$  selection. The predicted normalization agrees perfectly by construction, but the shapes are a feature of the prediction. The quality of the prediction is generally good, and no clear systematic biases are observed.

Figures 7.21, 7.22, 7.23, and 7.24 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $3b$  selection. The predicted normalization agrees perfectly by construction, but the shapes are a feature of the prediction. The quality of the prediction is generally good, and no clear systematic biases are observed.

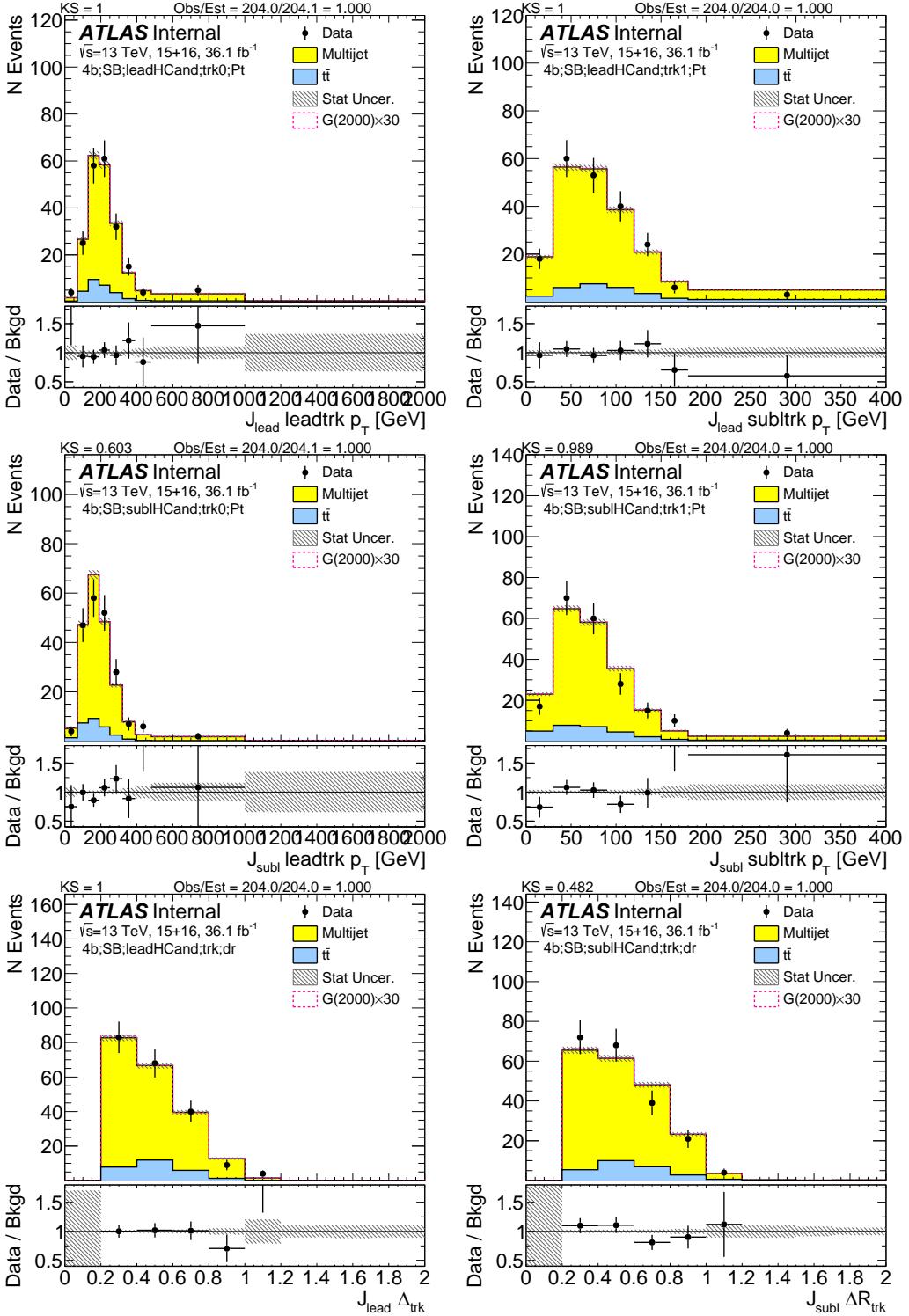
Figures 7.25, 7.26, 7.27, and 7.28 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $2bs$  selection. The predicted normalization agrees perfectly by construction, but the shapes are a feature of the prediction. The quality of the prediction is generally good, and no clear systematic biases are observed.



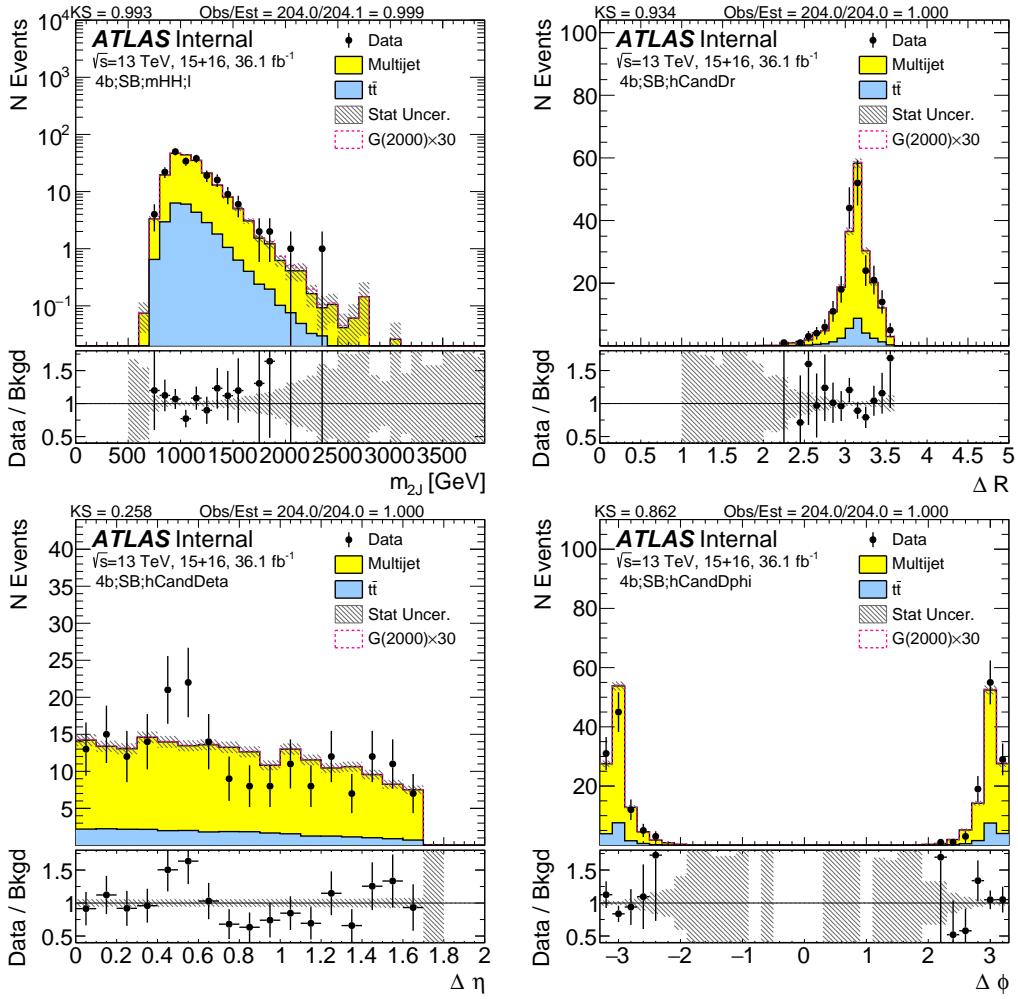
**Figure 7.17:** Kinematics of the lead large- $R$  jet in data and prediction in the sideband region after requiring 4  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



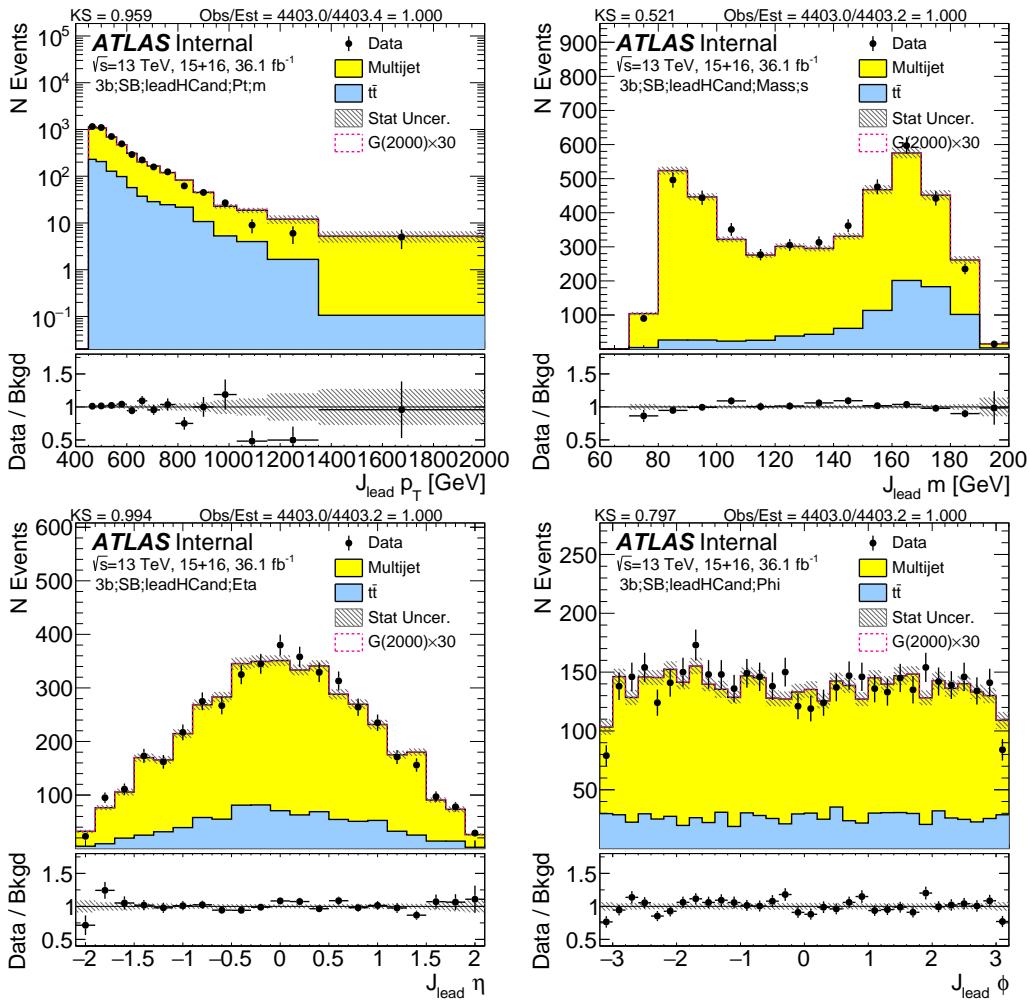
**Figure 7.18:** Kinematics of the sub-lead large- $R$  jet in data and prediction in the sideband region after requiring 4  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



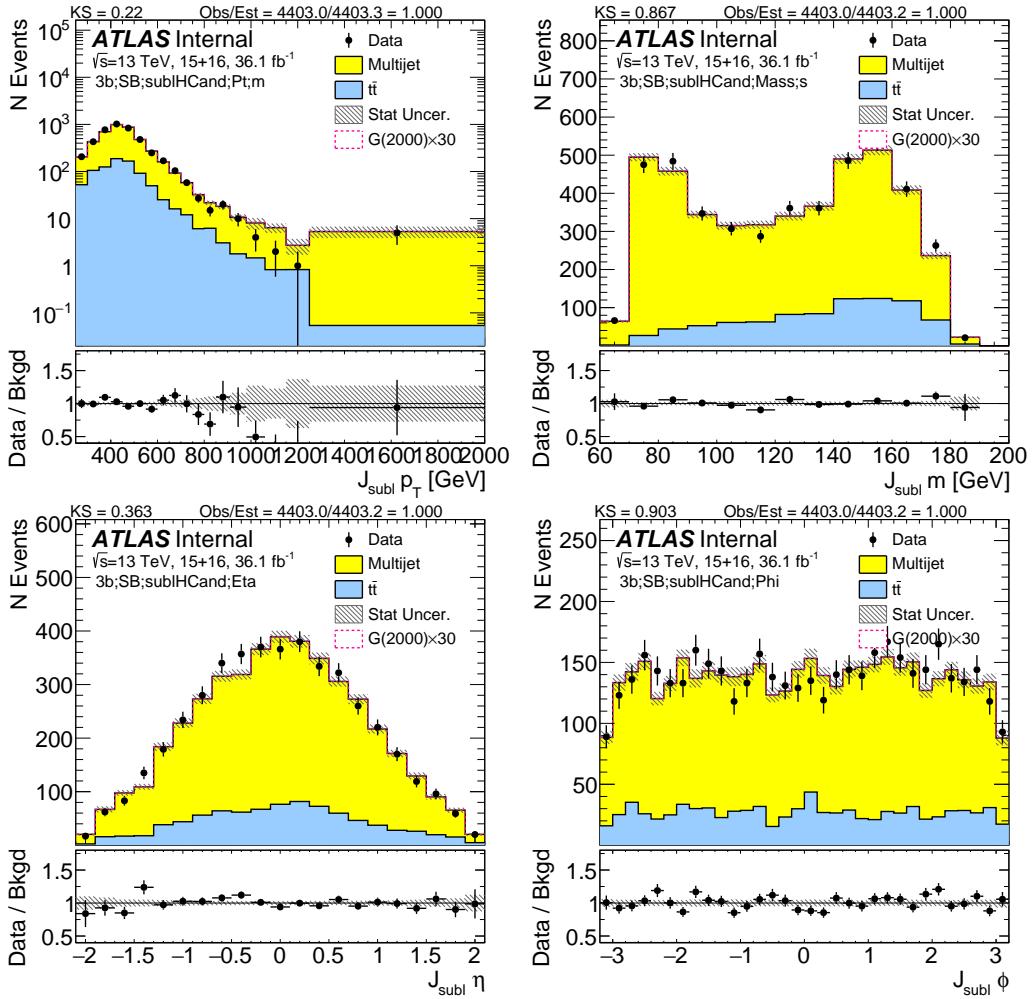
**Figure 7.19:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the sideband region after requiring 4  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. The normalization agrees by construction, and the shapes are a feature of the prediction.



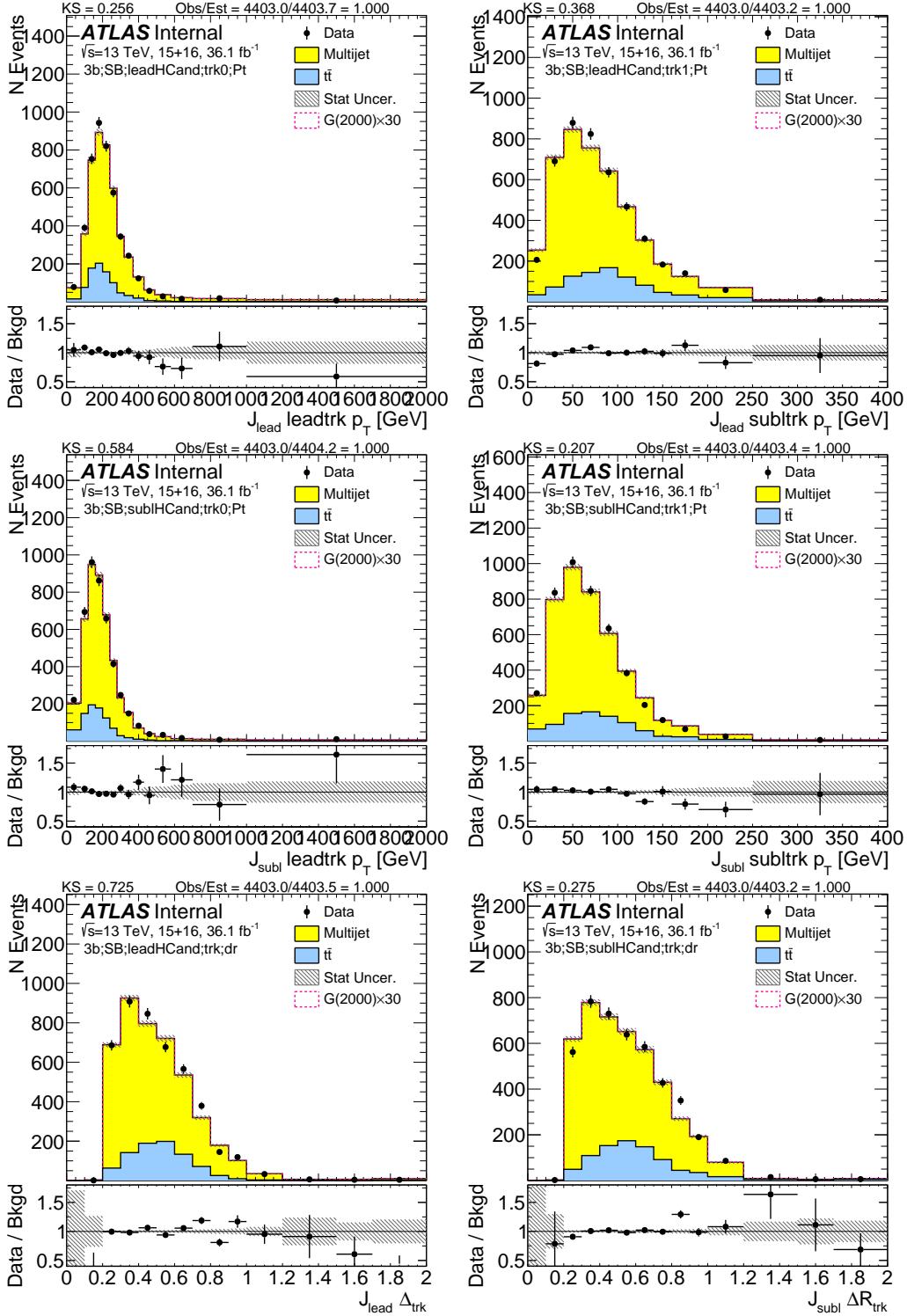
**Figure 7.20:** Kinematics of the large- $R$  jet system in data and prediction in the sideband region after requiring 4  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



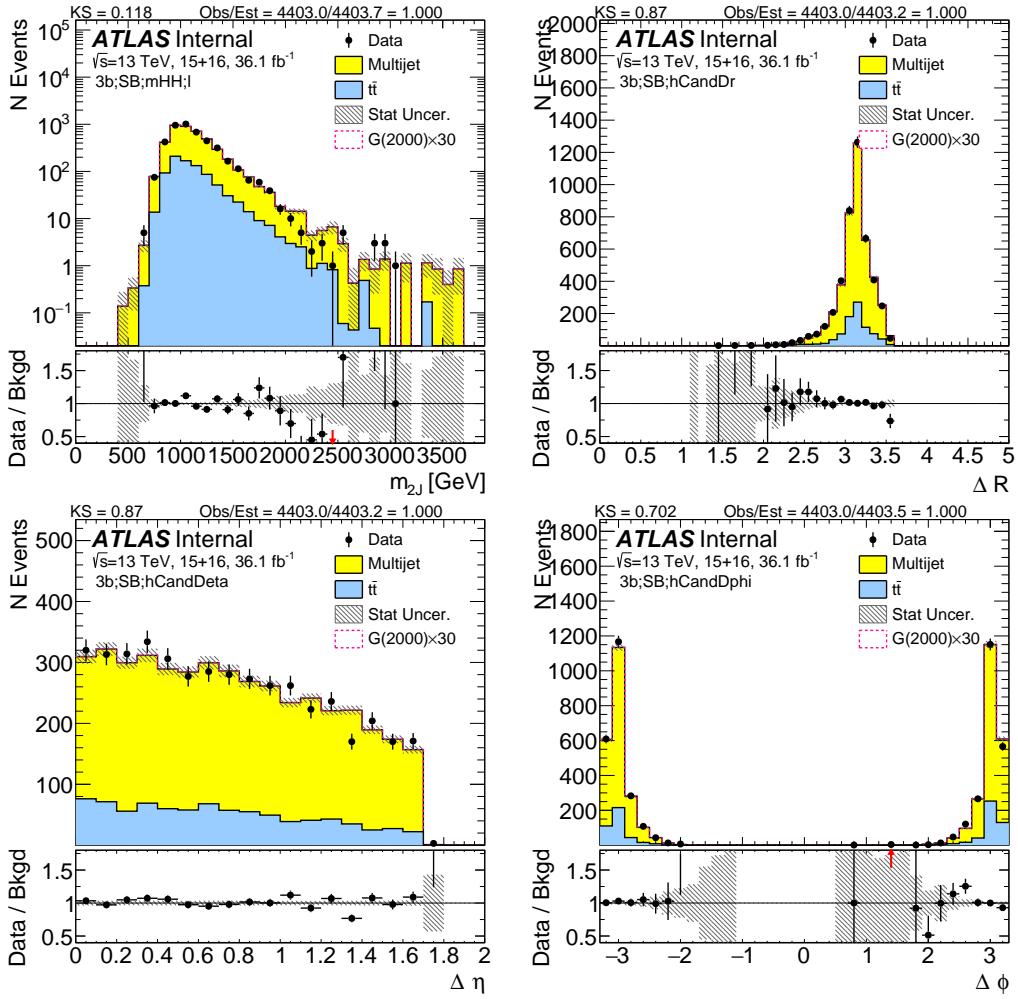
**Figure 7.21:** Kinematics of the lead large- $R$  jet in data and prediction in the sideband region after requiring 3  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



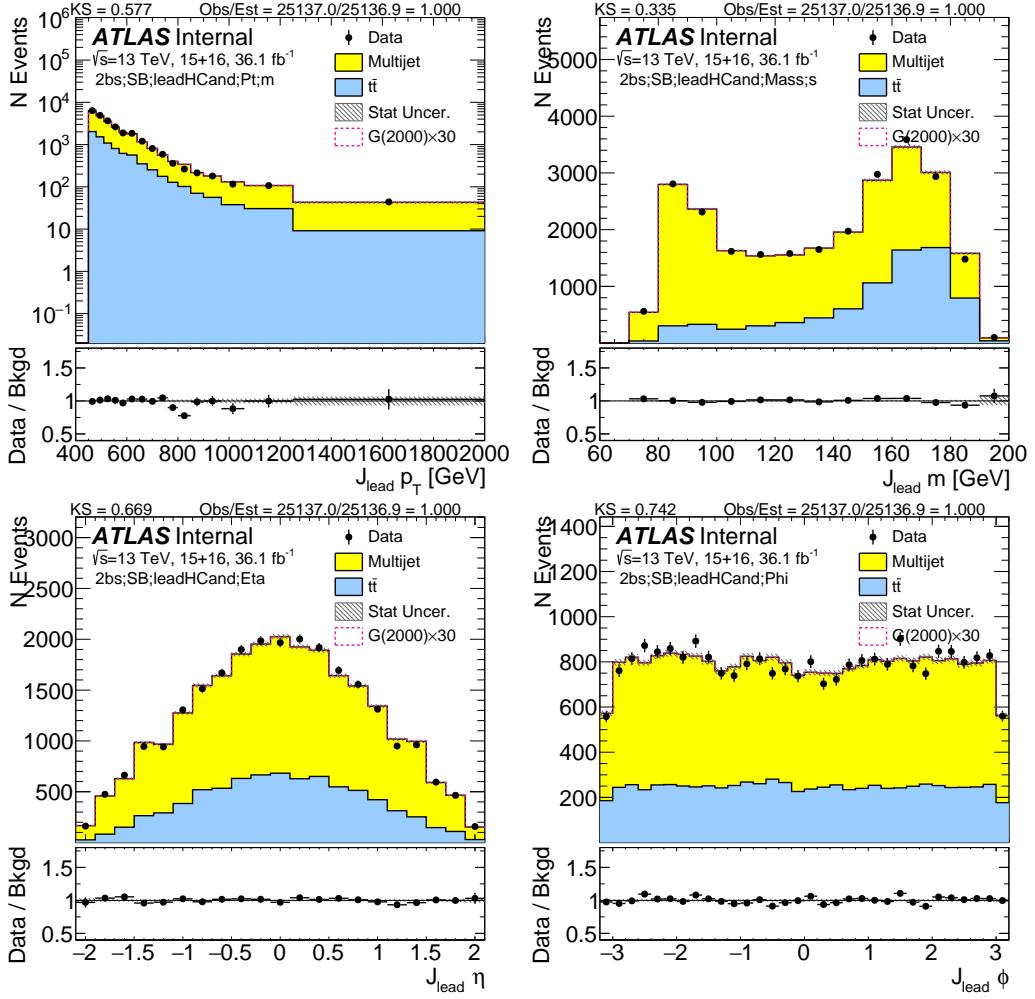
**Figure 7.22:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the sideband region after requiring 3  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



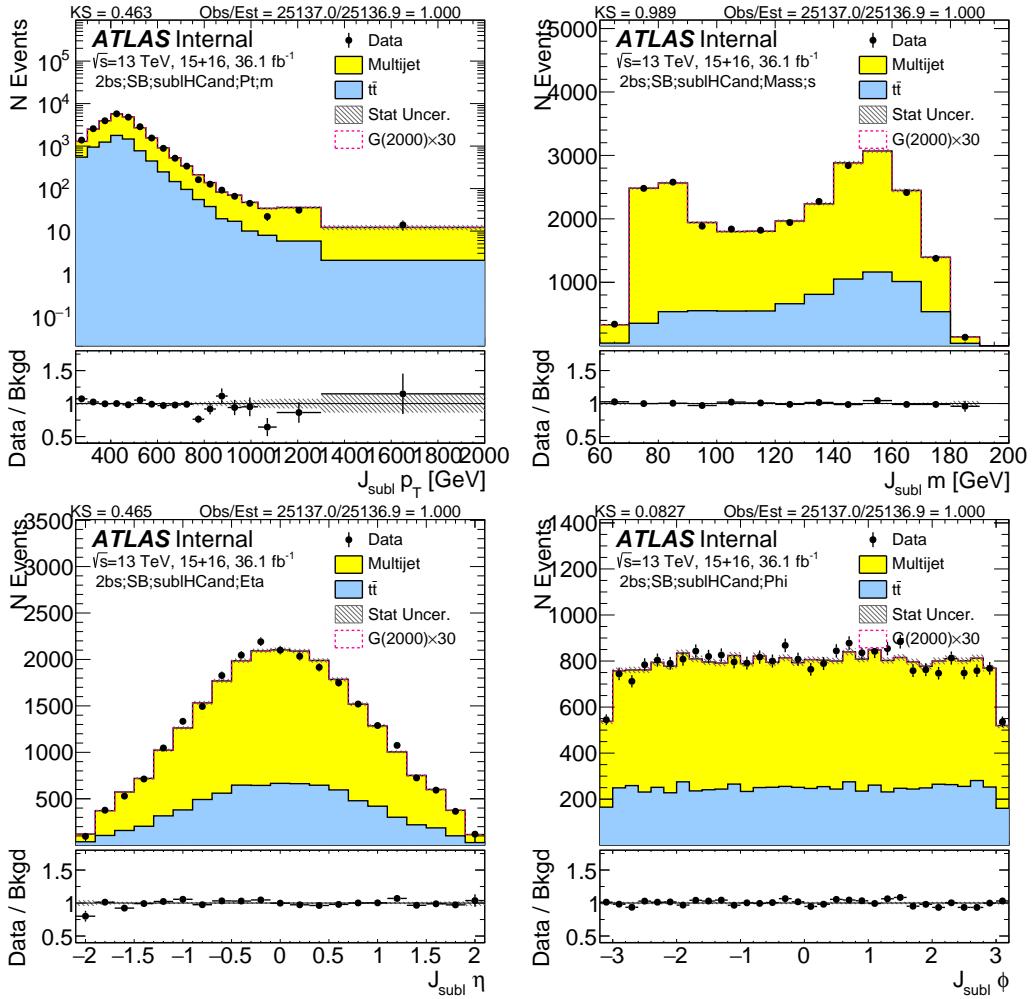
**Figure 7.23:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the sideband region after requiring 3  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. The normalization agrees by construction, and the shapes are a feature of the prediction.



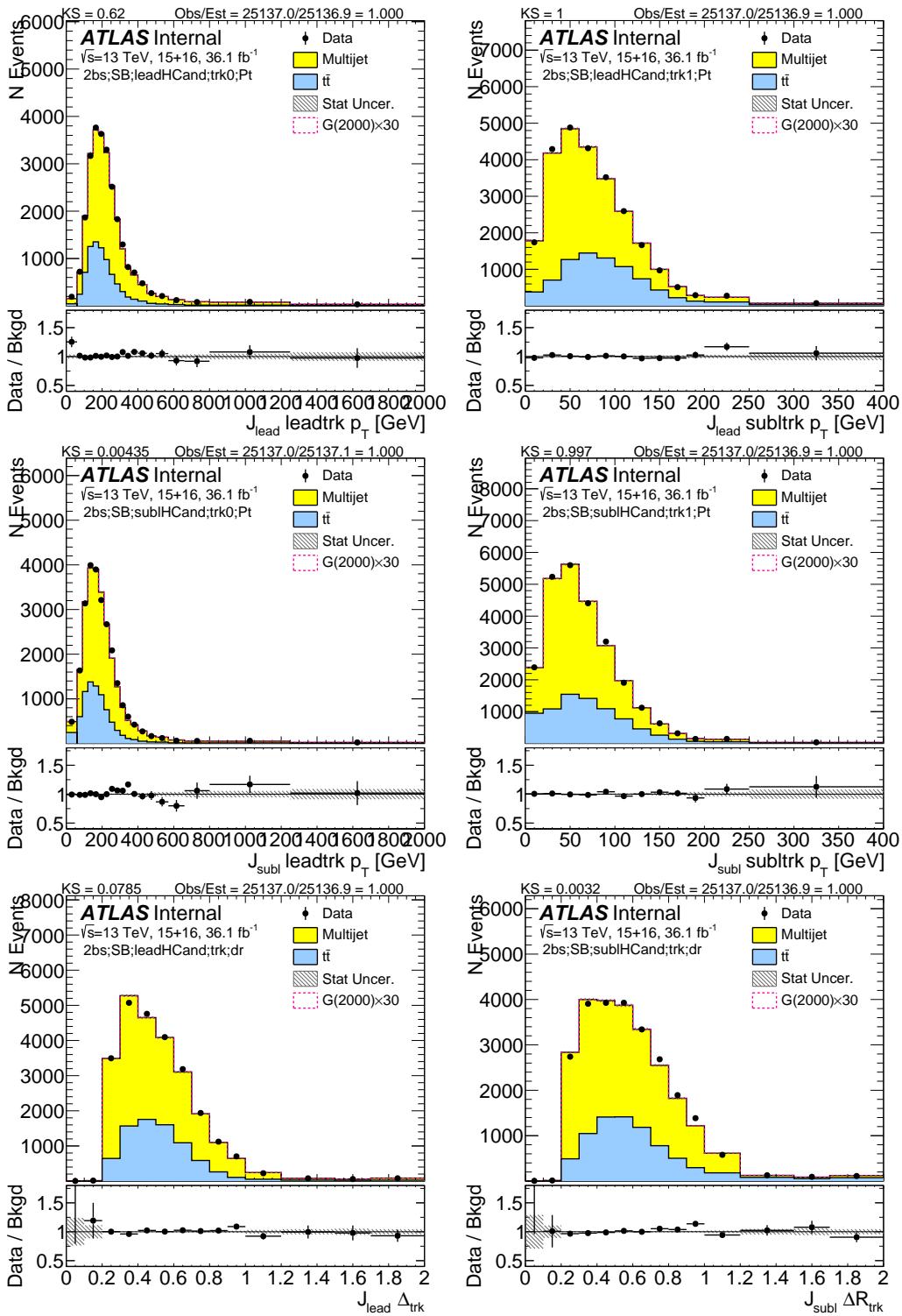
**Figure 7.24:** Kinematics of the large- $R$  jet system in data and prediction in the sideband region after requiring 3  $b$ -tags. The normalization agrees by construction, and the shapes are a feature of the prediction.



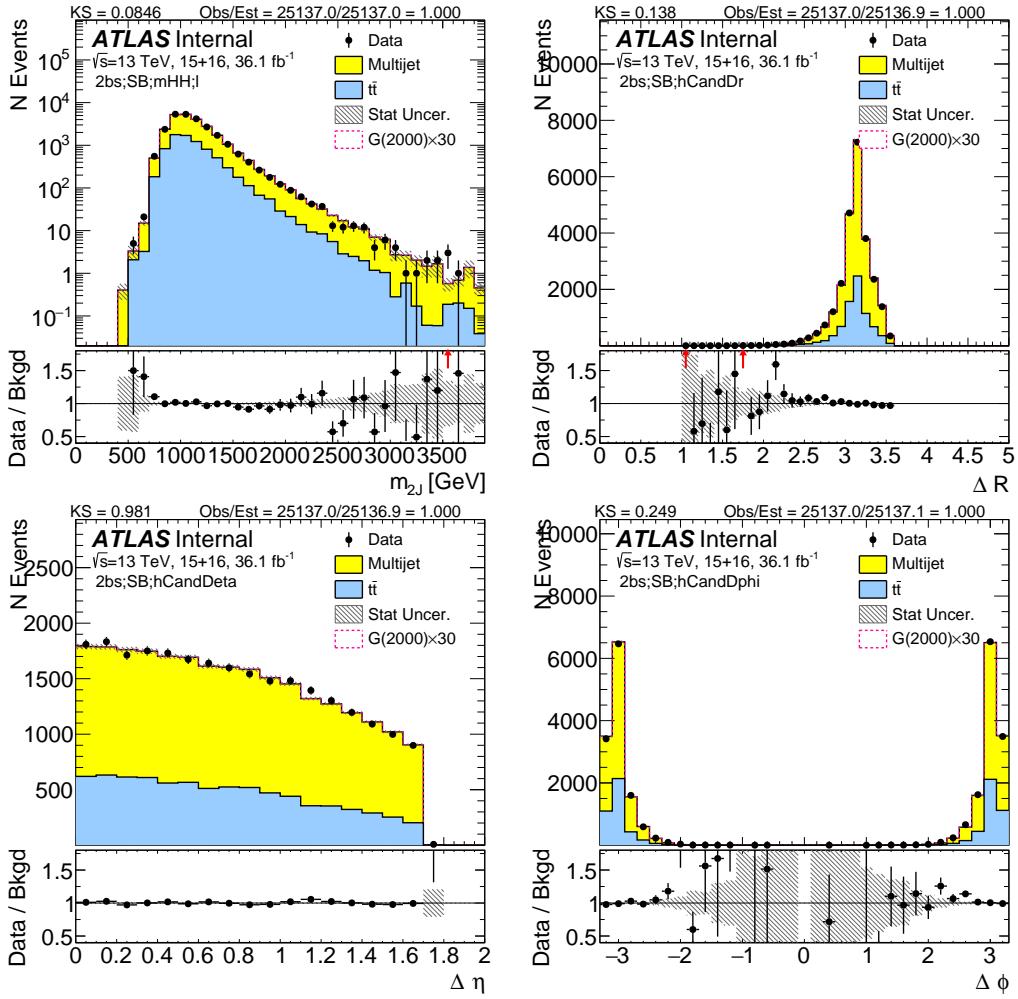
**Figure 7.25:** Kinematics of the lead large- $R$  jet in data and prediction in the sideband region after requiring 2  $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction.



**Figure 7.26:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the sideband region after requiring 2  $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction.



**Figure 7.27:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the sideband region after requiring 2  $b$ -tags split. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. The normalization agrees by construction, and the shapes are a feature of the prediction.



**Figure 7.28:** Kinematics of the large- $R$  jet system in data and prediction in the sideband region after requiring 2  $b$ -tags split. The normalization agrees by construction, and the shapes are a feature of the prediction.

### 7.1.8 PREDICTIONS IN THE CONTROL REGION (CR)

This section shows comparisons of data with the prediction of QCD multi-jets and  $t\bar{t}$  in the control region (CR), which is identical to the signal region (SR) except the large- $R$  jets are required to have masses close but not too close to the Higgs mass. The definition can be seen in Section 7.1.2. The predicted and observed event yields are summarized in Tables ?? and ??.

Figures 7.29, 7.30, 7.31, and 7.32 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $4b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB. The quality of the prediction is generally good, and no clear systematic biases are observed.

Figures 7.33, 7.34, 7.35, and 7.36 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $3b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB. The quality of the prediction is generally good, and no clear systematic biases are observed.

Figures 7.37, 7.38, 7.39, and 7.40 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $2bs$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB. The quality of the prediction is generally good, and no clear systematic biases are observed.

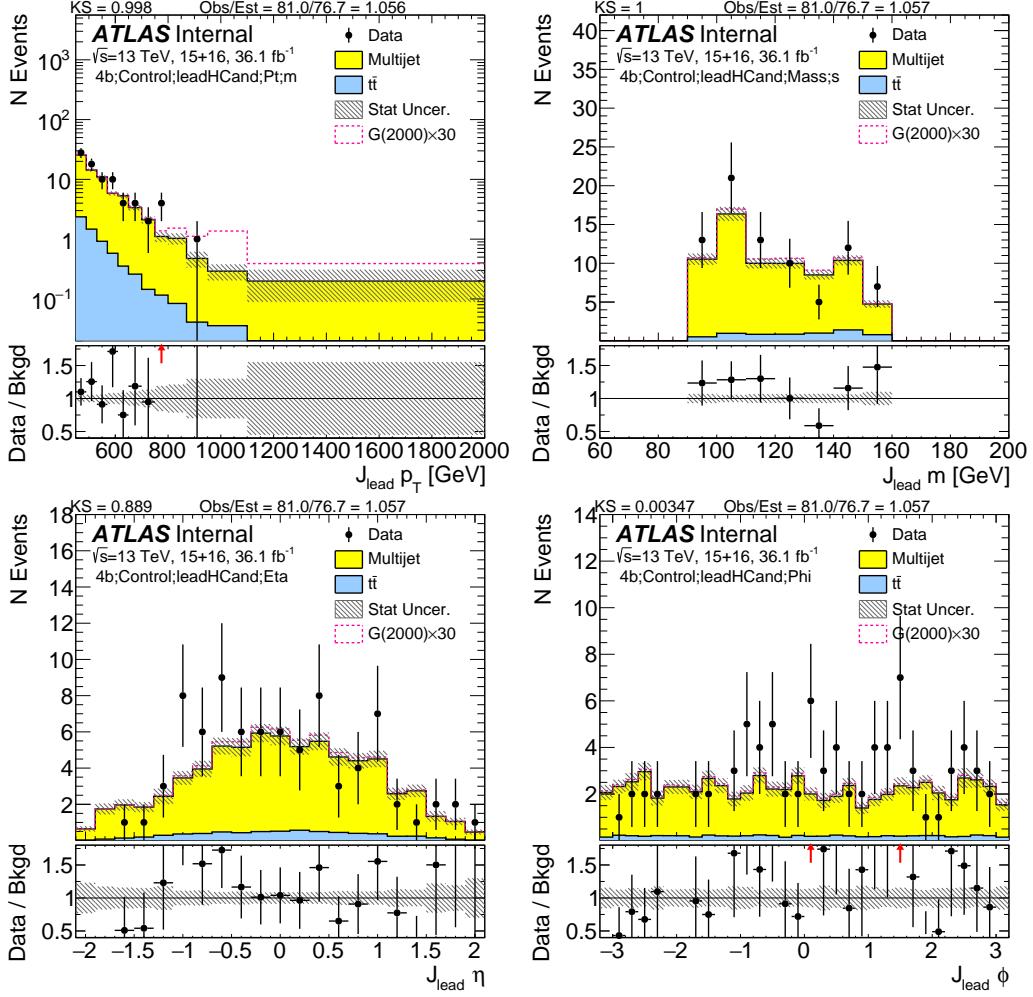


Figure 7.29: Kinematics of the lead large- $R$  jet in data and prediction in the control region after requiring 4  $b$ -tags.

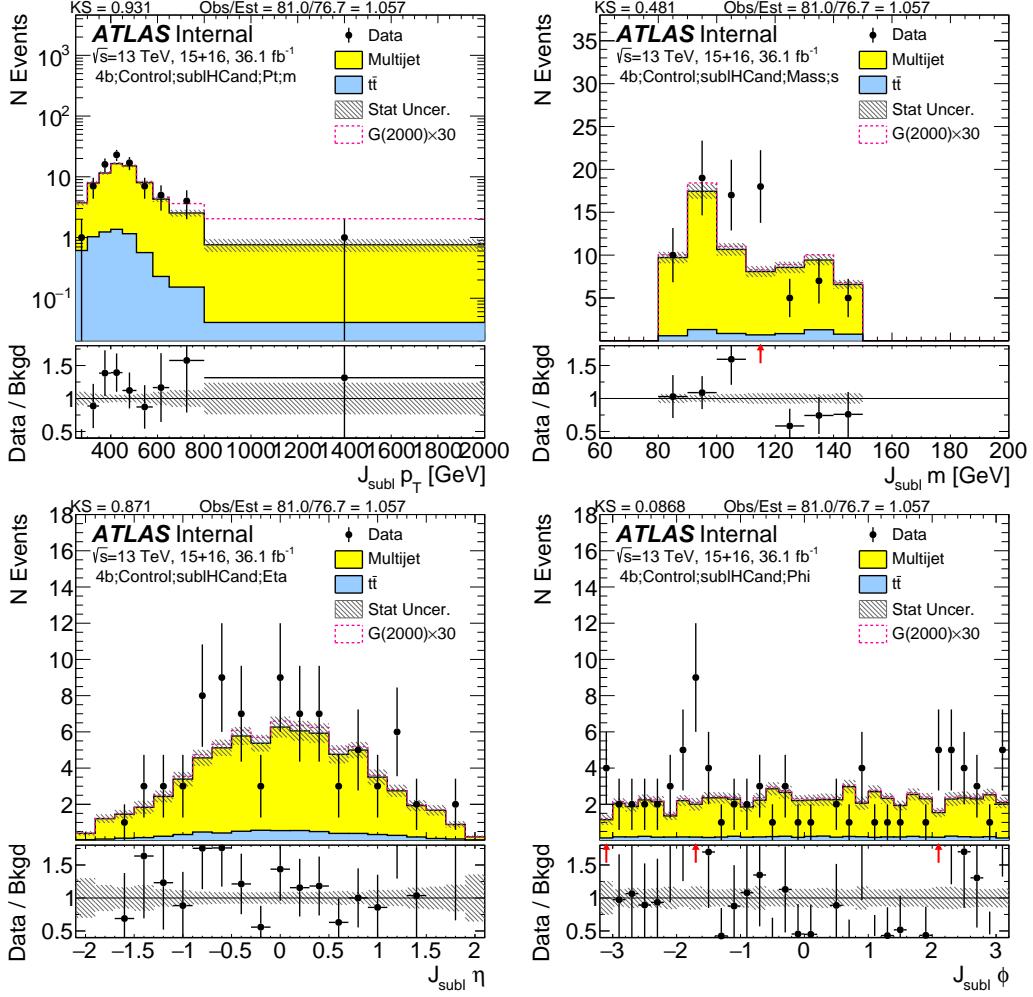


Figure 7.30: Kinematics of the sub-lead large- $R$  jet in data and prediction in the control region after requiring 4  $b$ -tags.

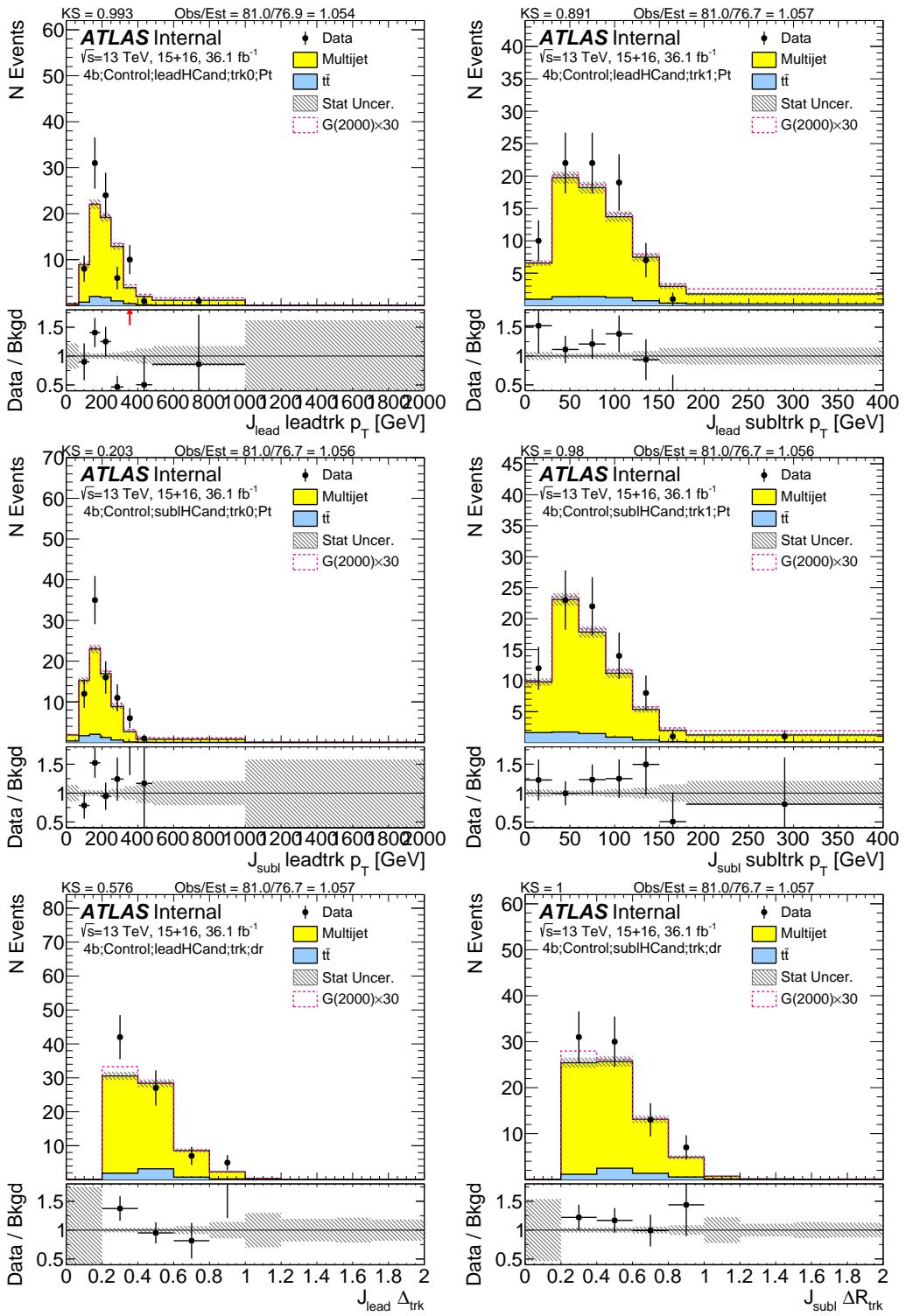


Figure 7.31: First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the control region after requiring 4  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.

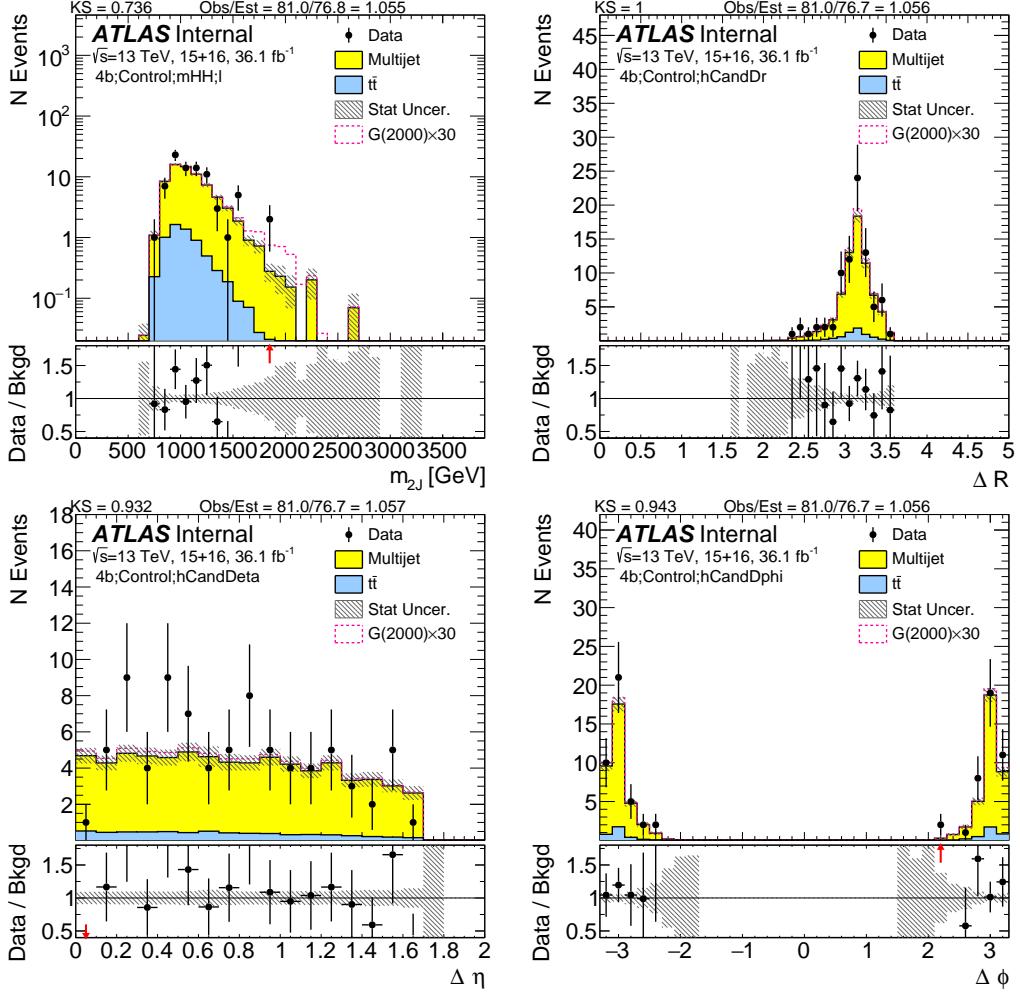


Figure 7.32: Kinematics of the large- $R$  jet system in data and prediction in the control region after requiring 4  $b$ -tags.

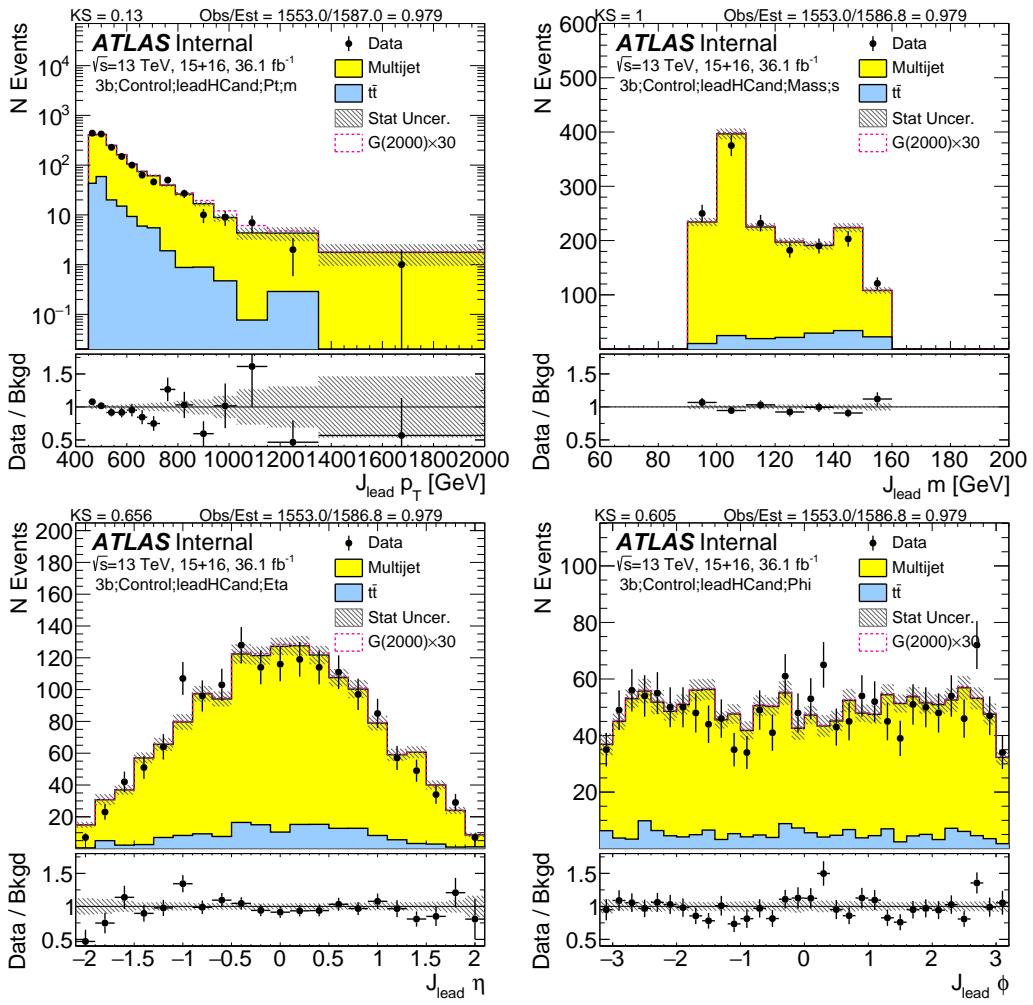


Figure 7.33: Kinematics of the lead large- $R$  jet in data and prediction in the control region after requiring 3  $b$ -tags.

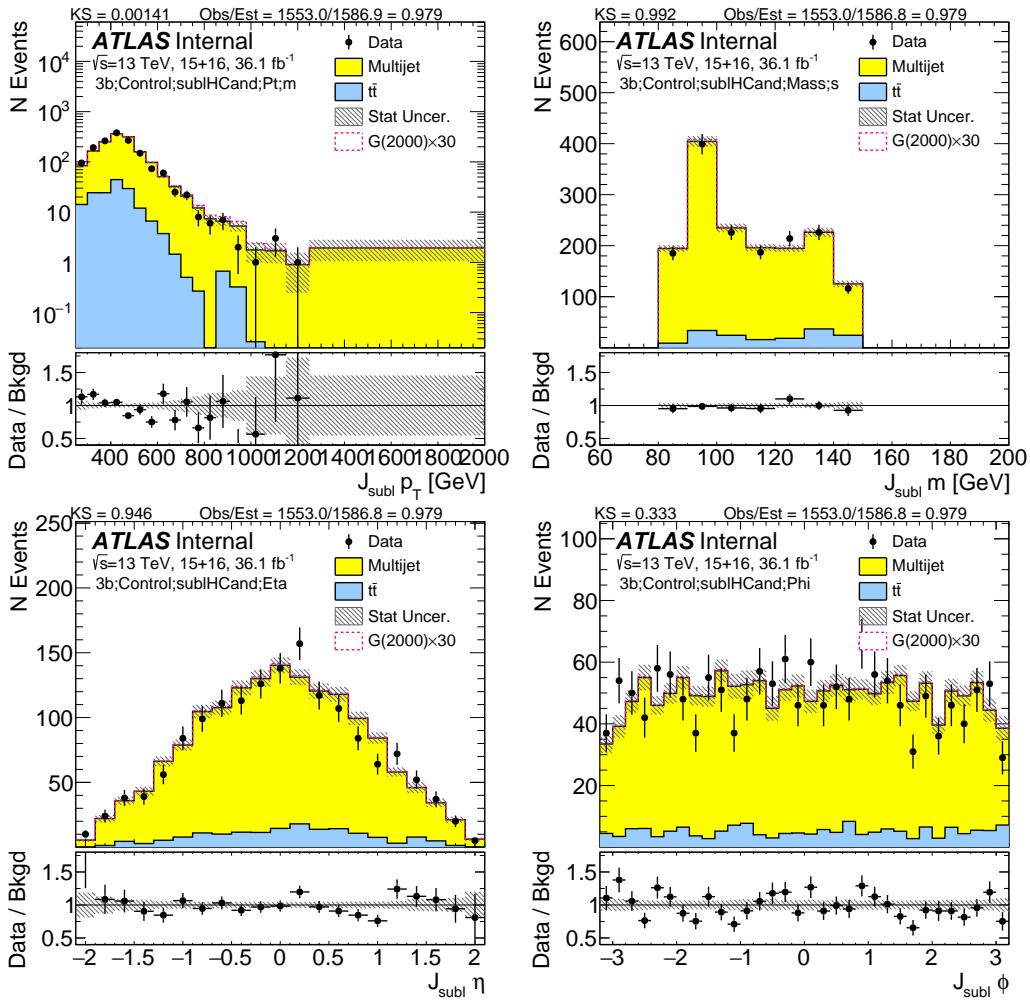
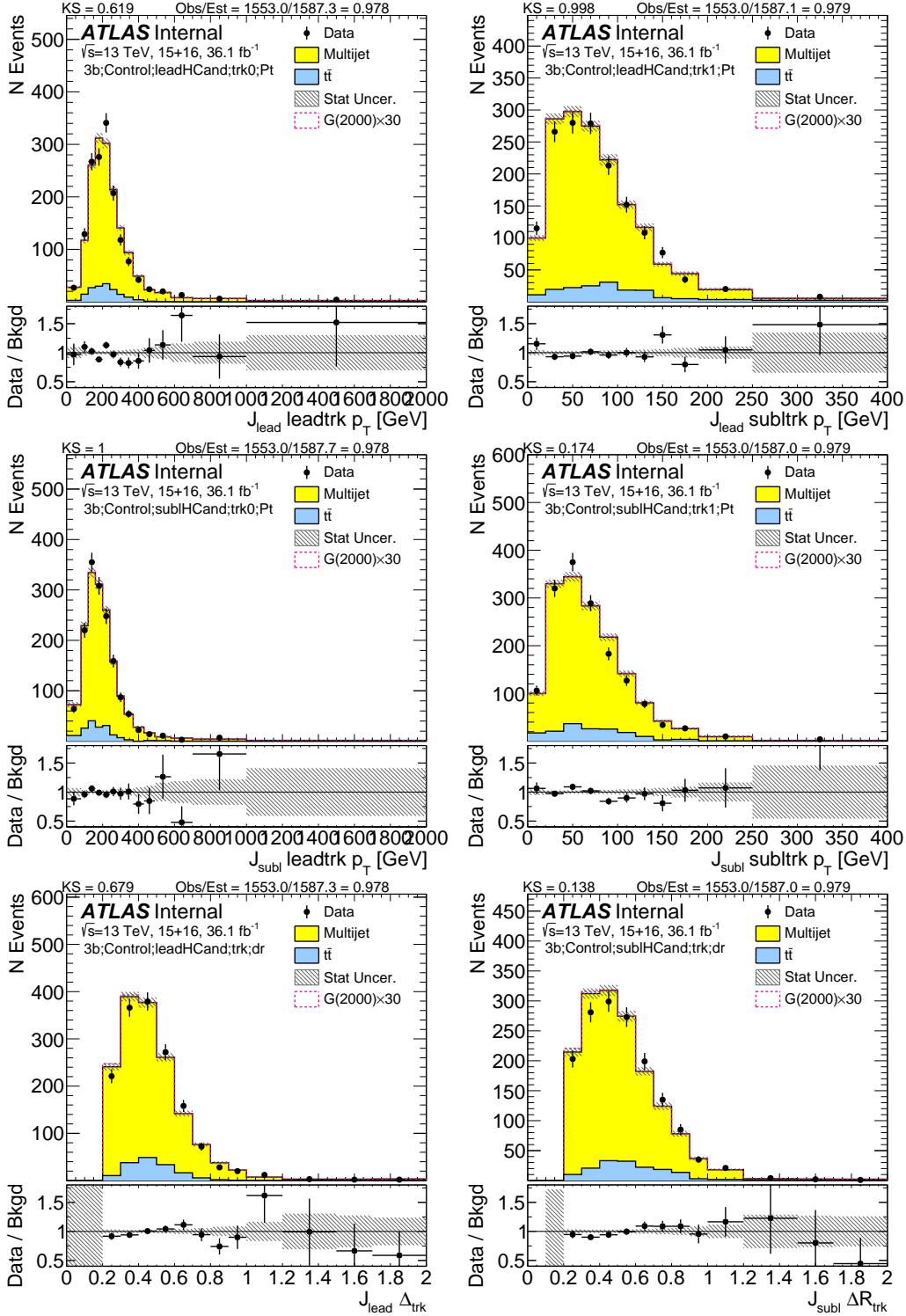


Figure 7.34: Kinematics of the sub-lead large- $R$  jet in data and prediction in the control region after requiring 3  $b$ -tags.



**Figure 7.35:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the control region after requiring 3  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.

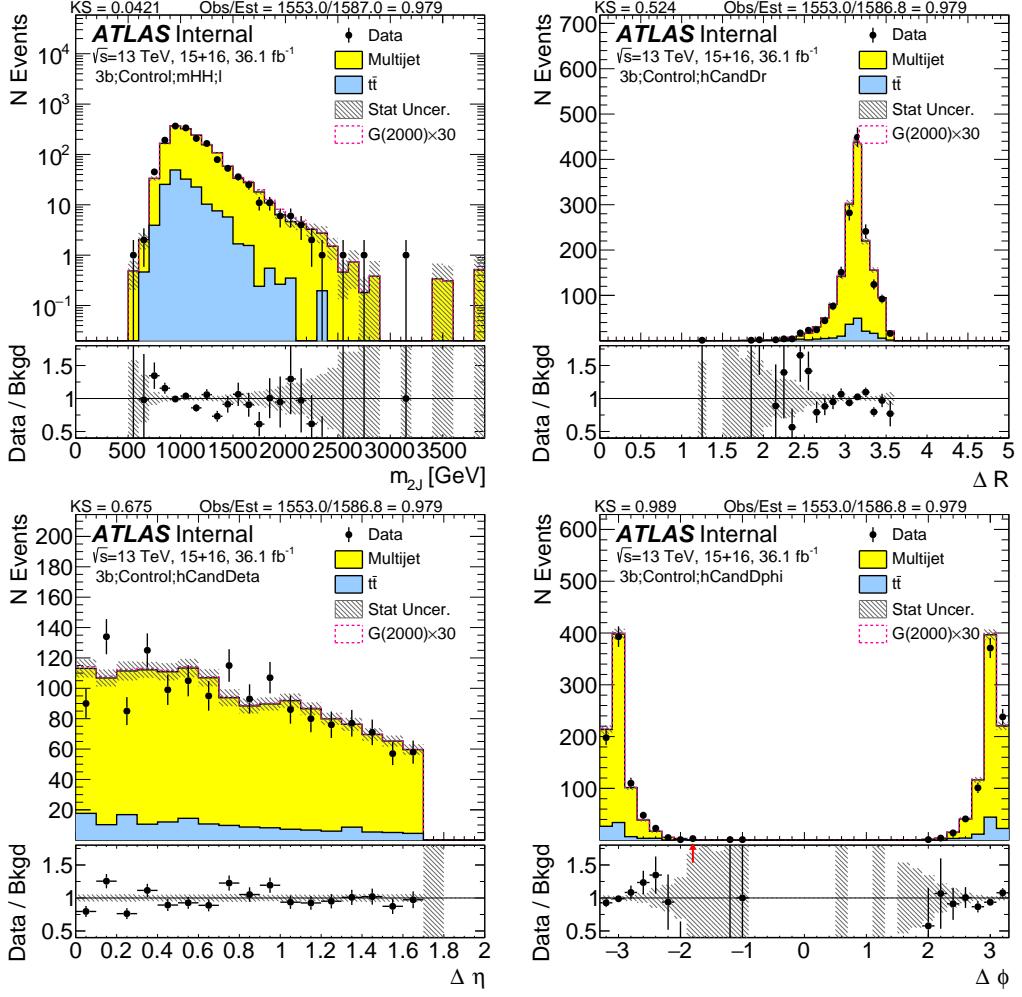
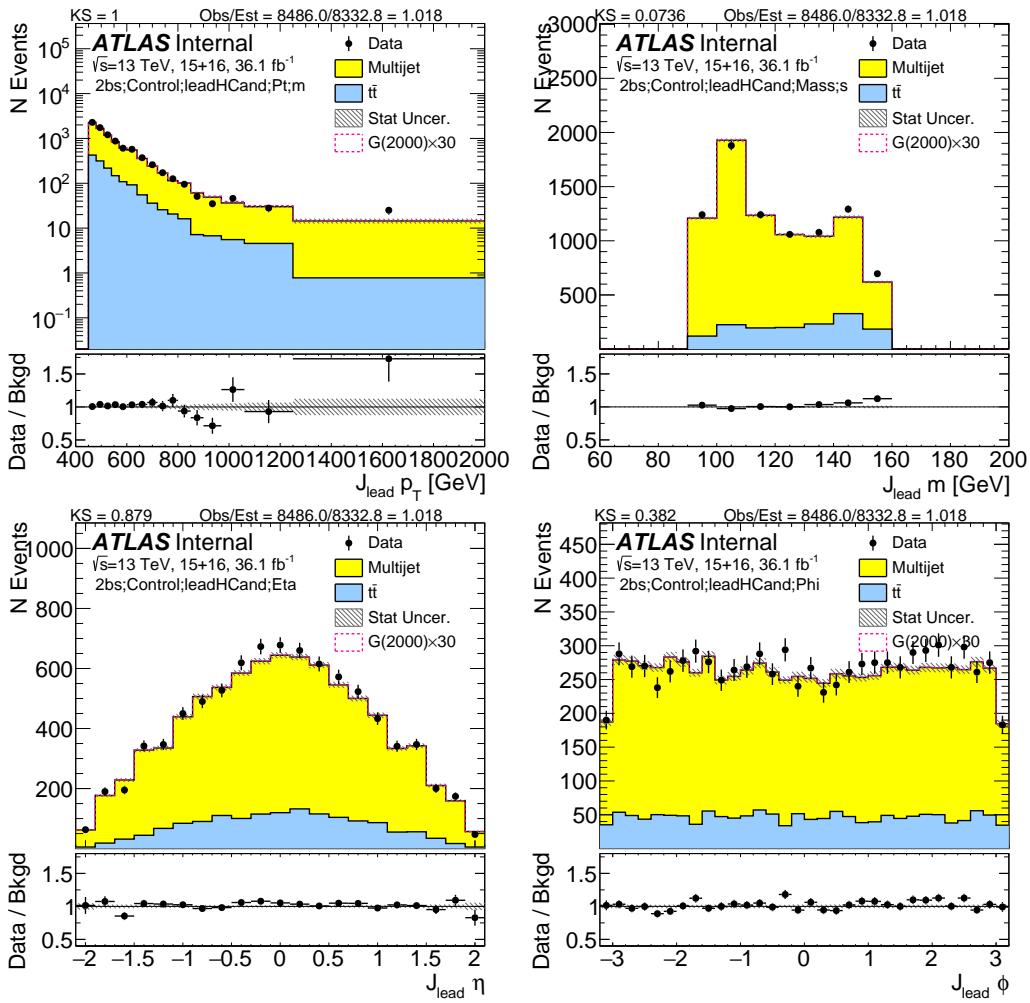
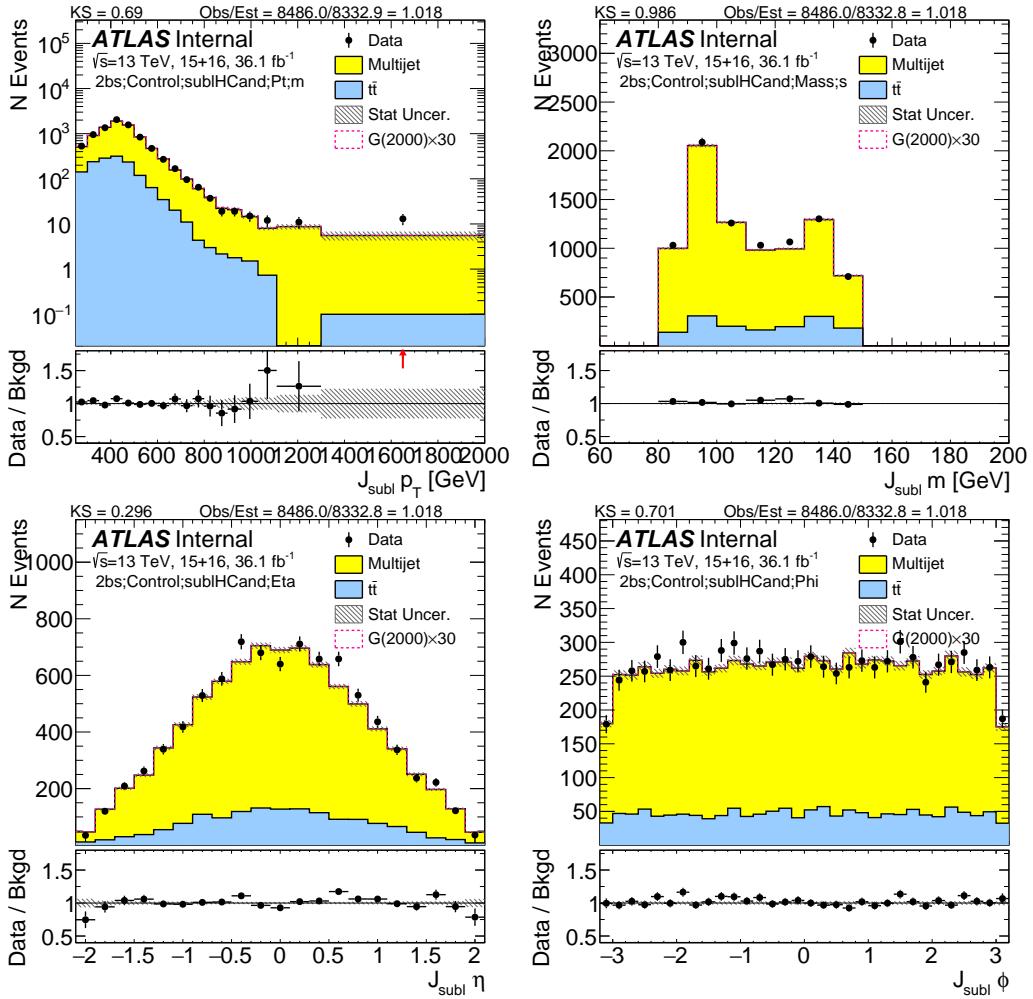


Figure 7.36: Kinematics of the large- $R$  jet system in data and prediction in the control region after requiring 3  $b$ -tags.



**Figure 7.37:** Kinematics of the lead large- $R$  jet in data and prediction in the control region after requiring 2  $b$ -tags split.



**Figure 7.38:** Kinematics of the sub-lead large- $R$  jet in data and prediction in the control region after requiring 2  $b$ -tags split.

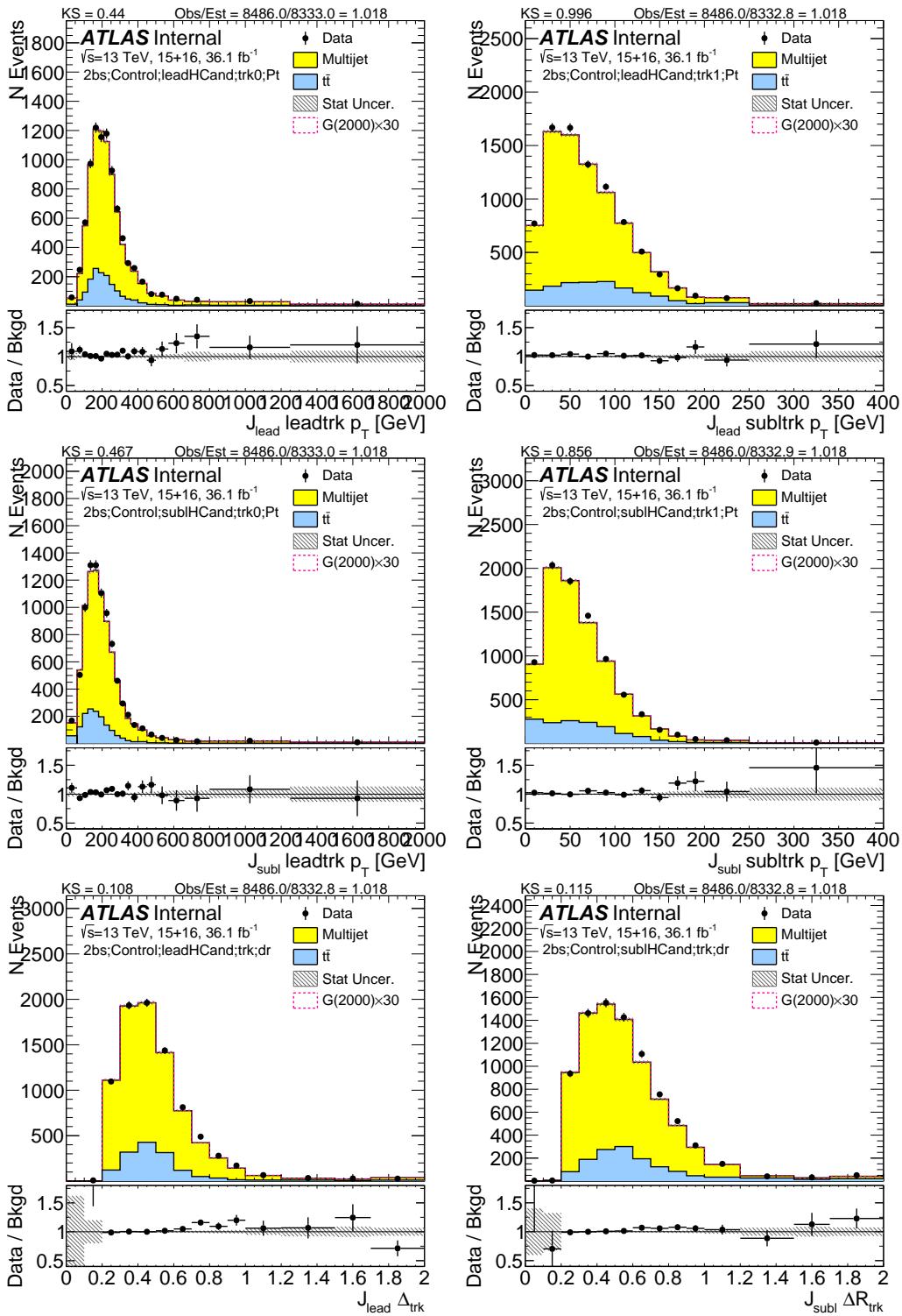
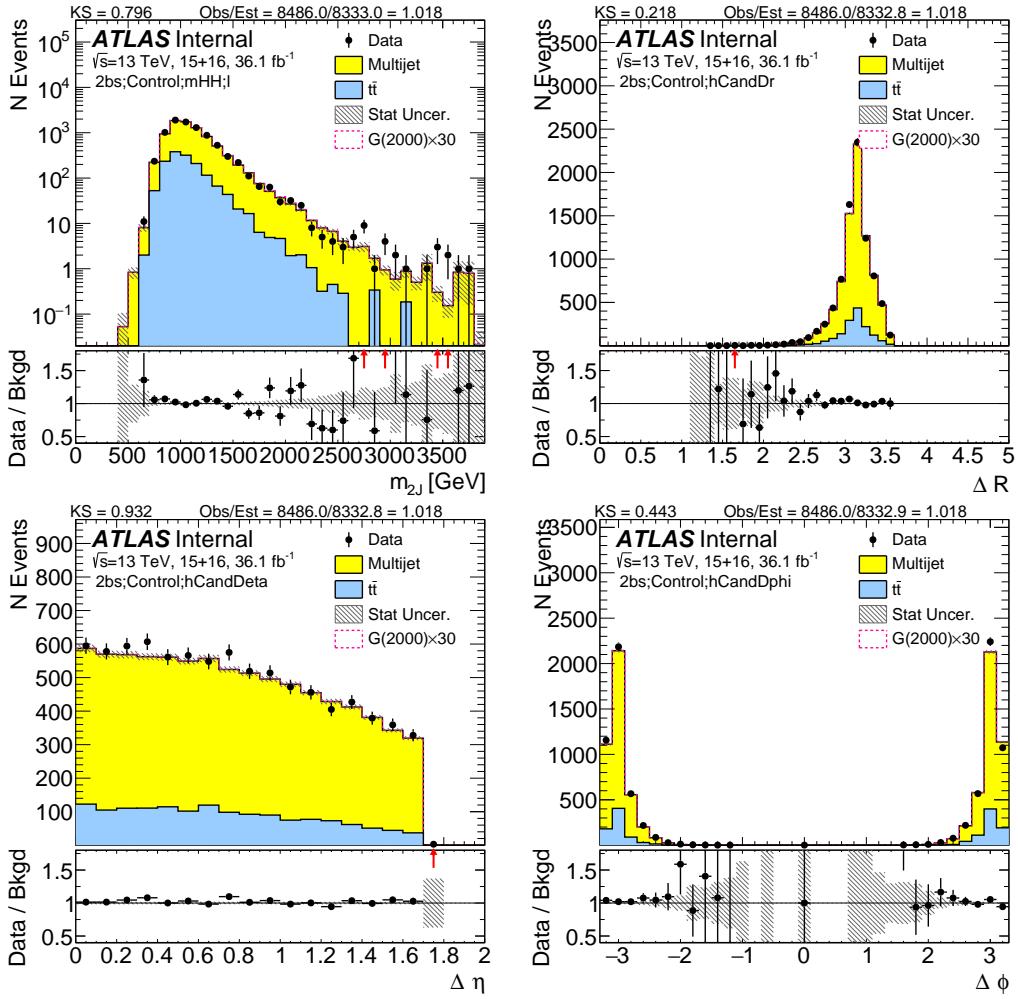


Figure 7.39: First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the control region after requiring 2  $b$ -tags split. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.



**Figure 7.40:** Kinematics of the large- $R$  jet system in data and prediction in the control region after requiring 2  $b$ -tags split.

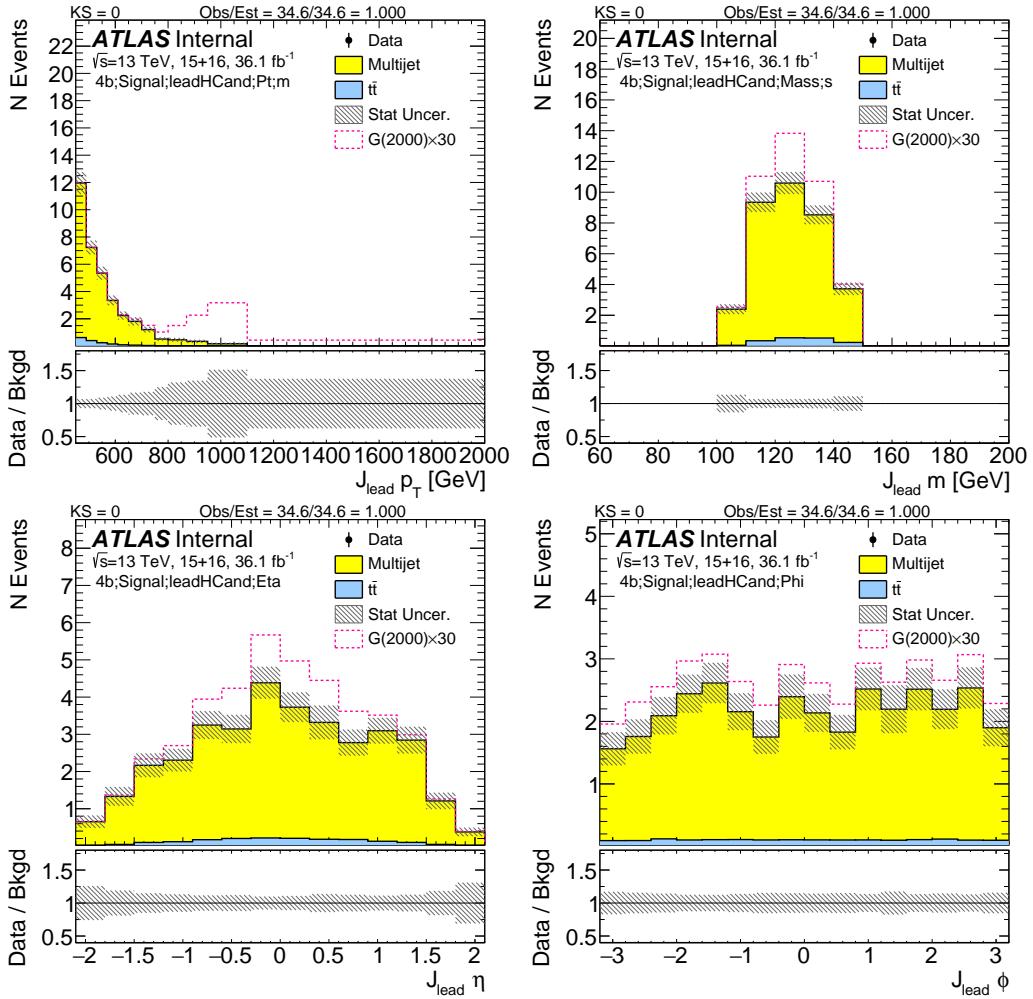
### 7.1.9 SIGNAL REGION PREDICTIONS

This section shows comparisons of data with the prediction of QCD multi-jets and  $t\bar{t}$  in the signal region (SR). Plots shown are blinded. The unblinded data is shown in the Result Section, section ??.

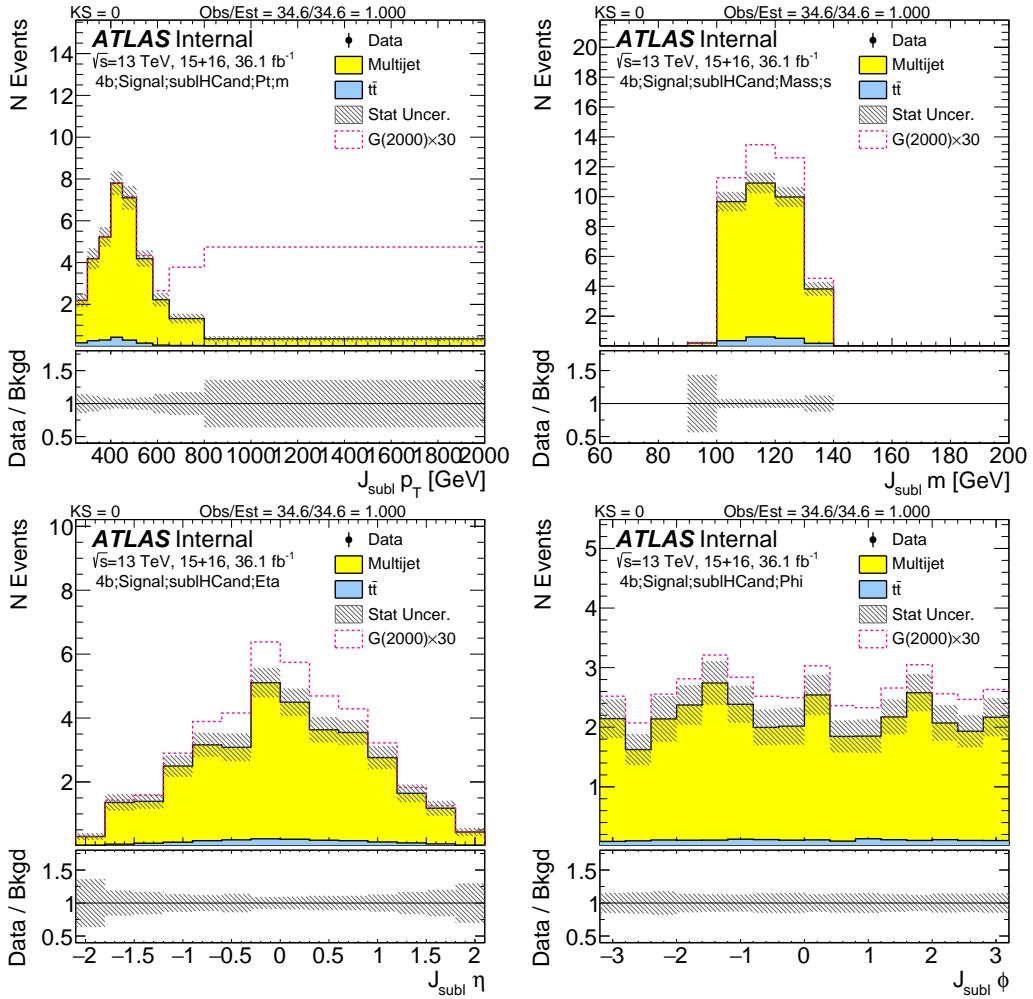
Figures 7.41, 7.42, 7.43, and 7.44 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $4b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB.

Figures 7.45, 7.46, 7.47, and 7.48 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $3b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB.

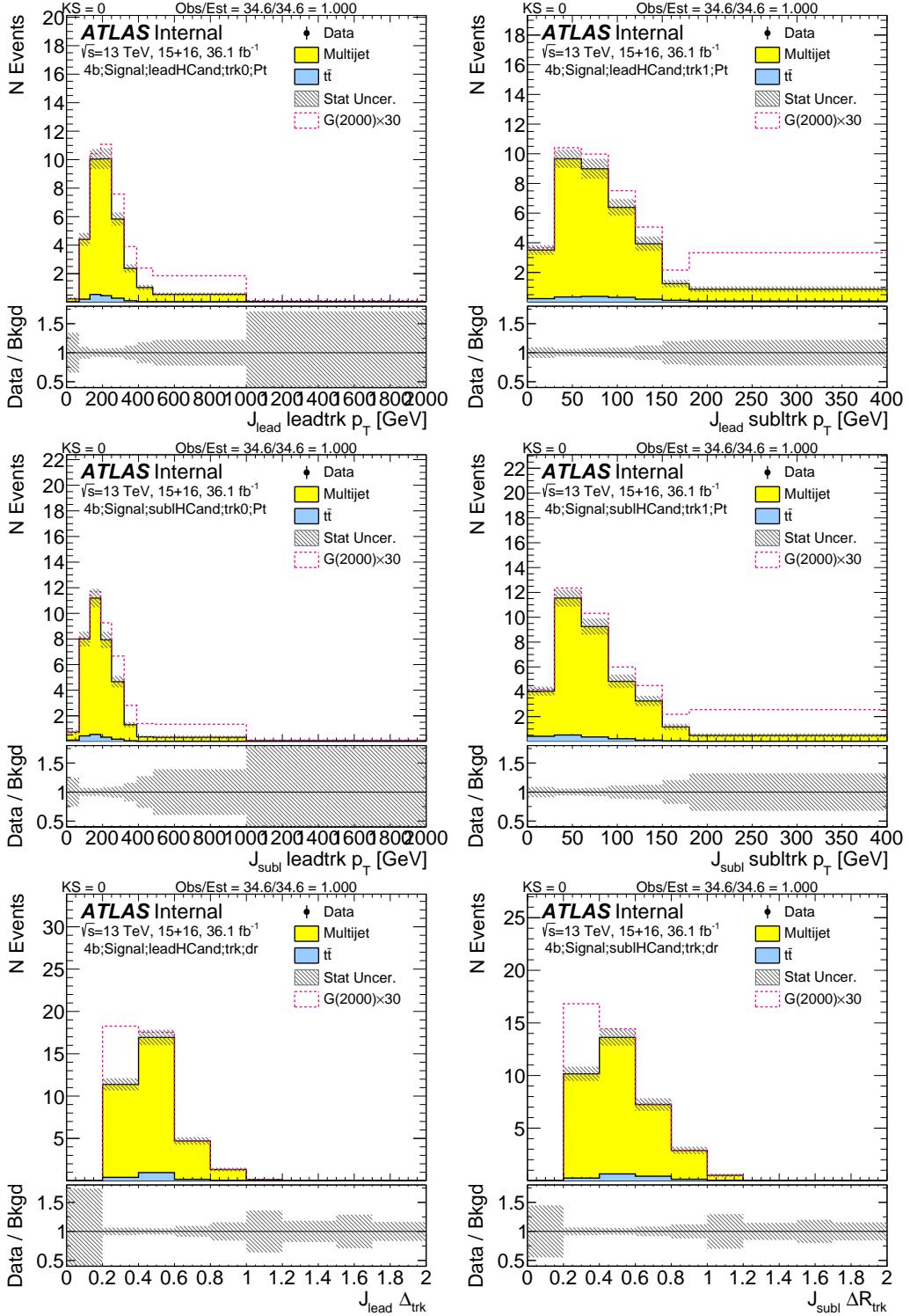
Figures 7.49, 7.50, 7.51, and 7.52 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $2b$  selection. The shapes and normalization are a feature of the prediction, where the normalization is derived in the SB.



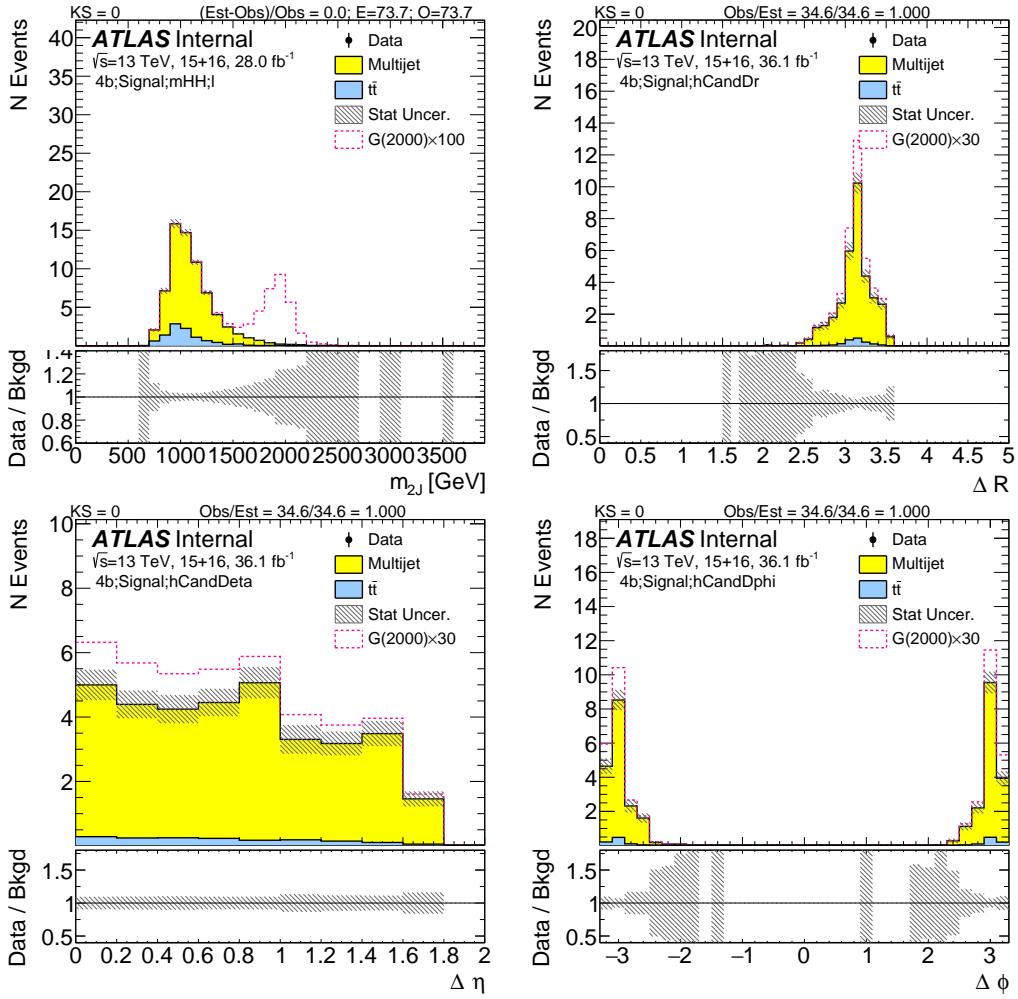
**Figure 7.41:** Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring 4  $b$ -tags. Data is blinded, and will be added after unblinding.



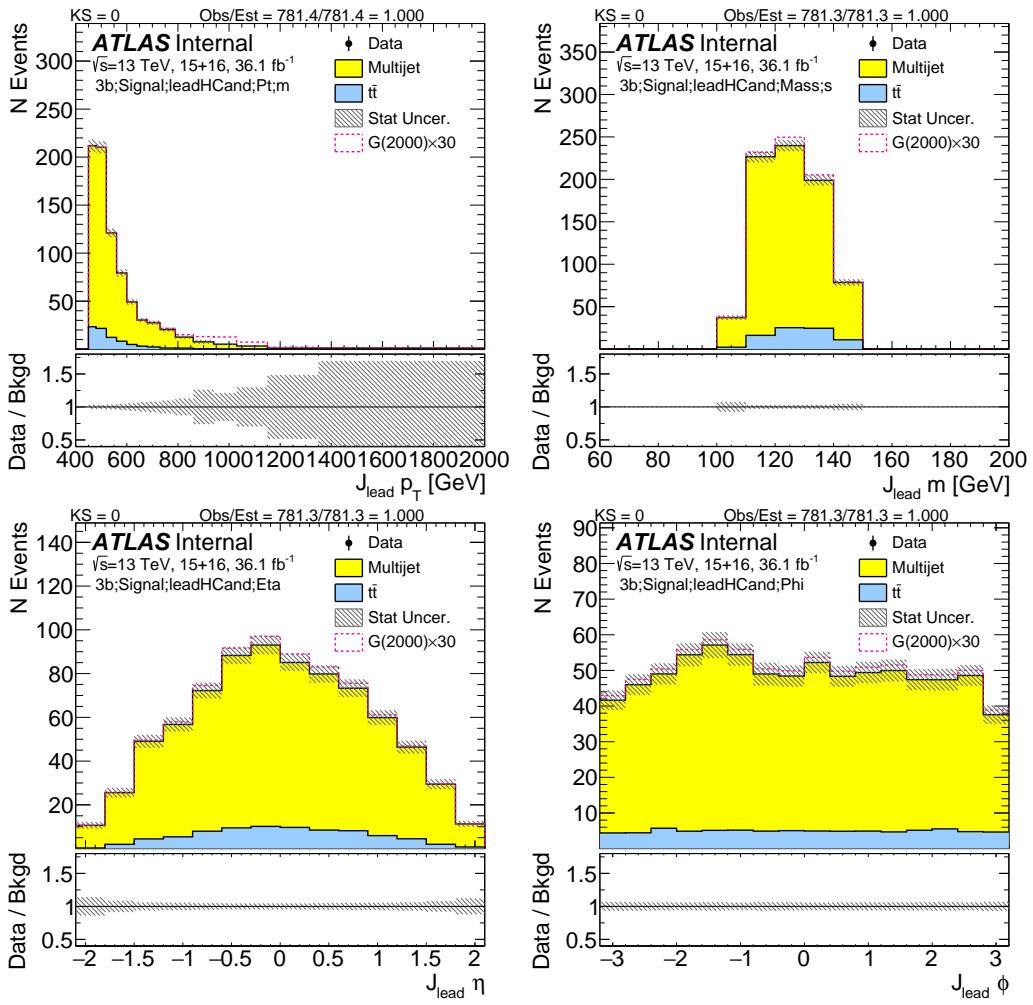
**Figure 7.42:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the signal region after requiring 4  $b$ -tags. Data is blinded, and will be added after unblinding.



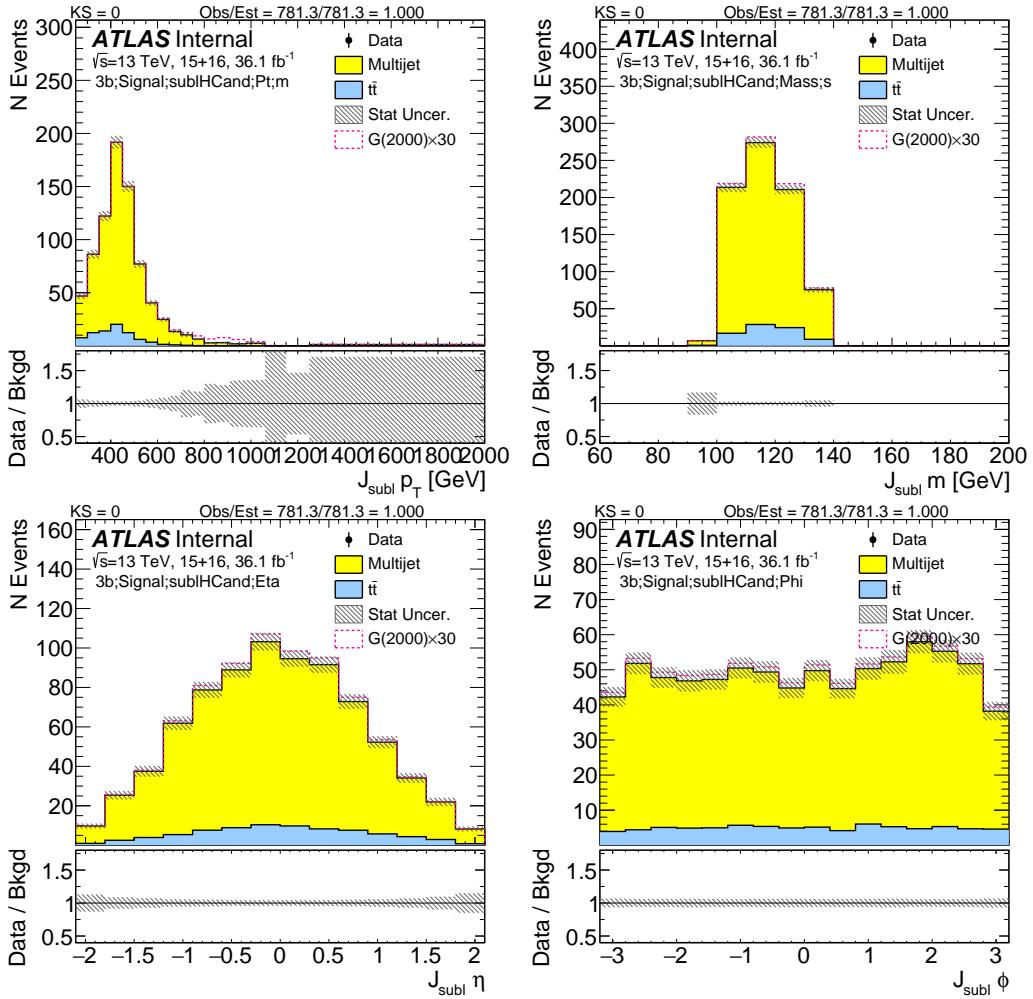
**Figure 7.43:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 4  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. Data is blinded, and will be added after unblinding.



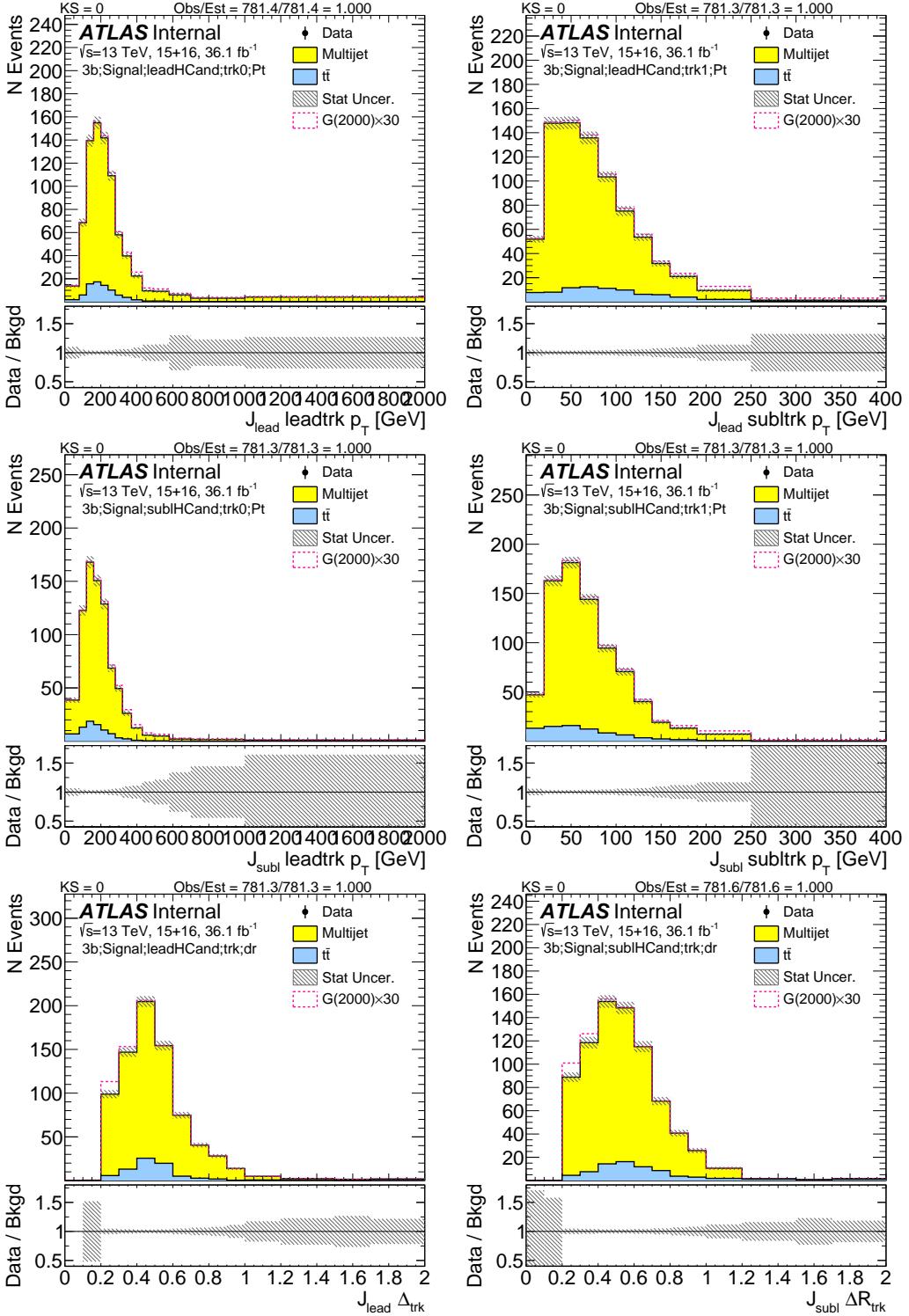
**Figure 7.44:** Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 4  $b$ -tags. Data is blinded, and will be added after unblinding.



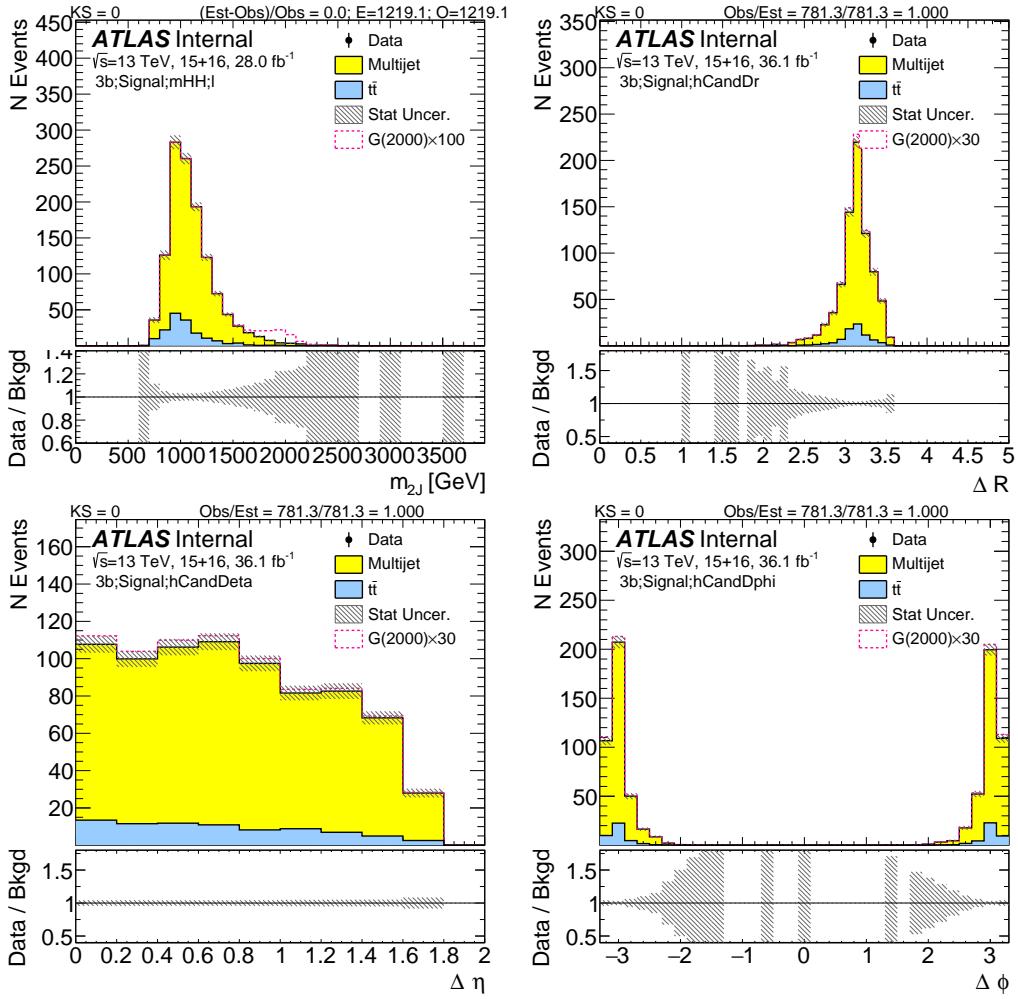
**Figure 7.45:** Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags. Data is blinded, and will be added after unblinding.



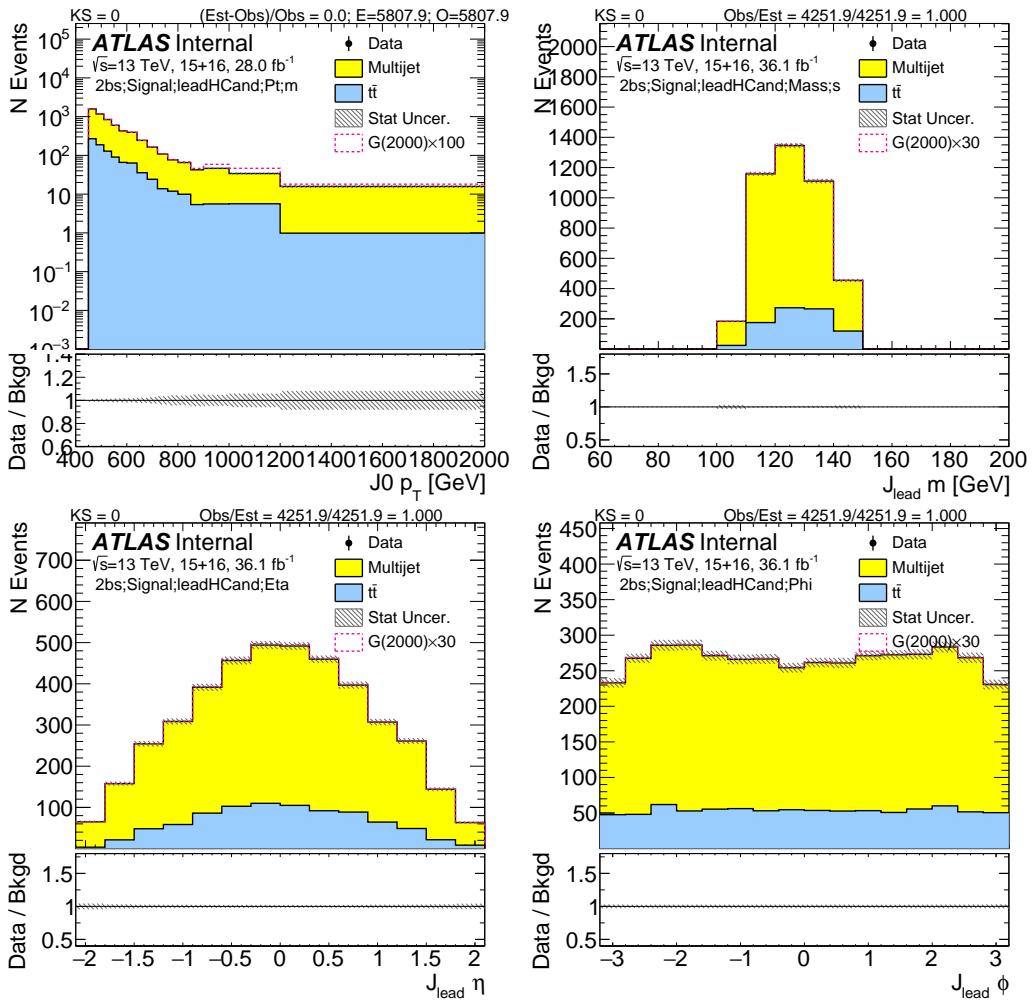
**Figure 7.46:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags. Data is blinded, and will be added after unblinding.



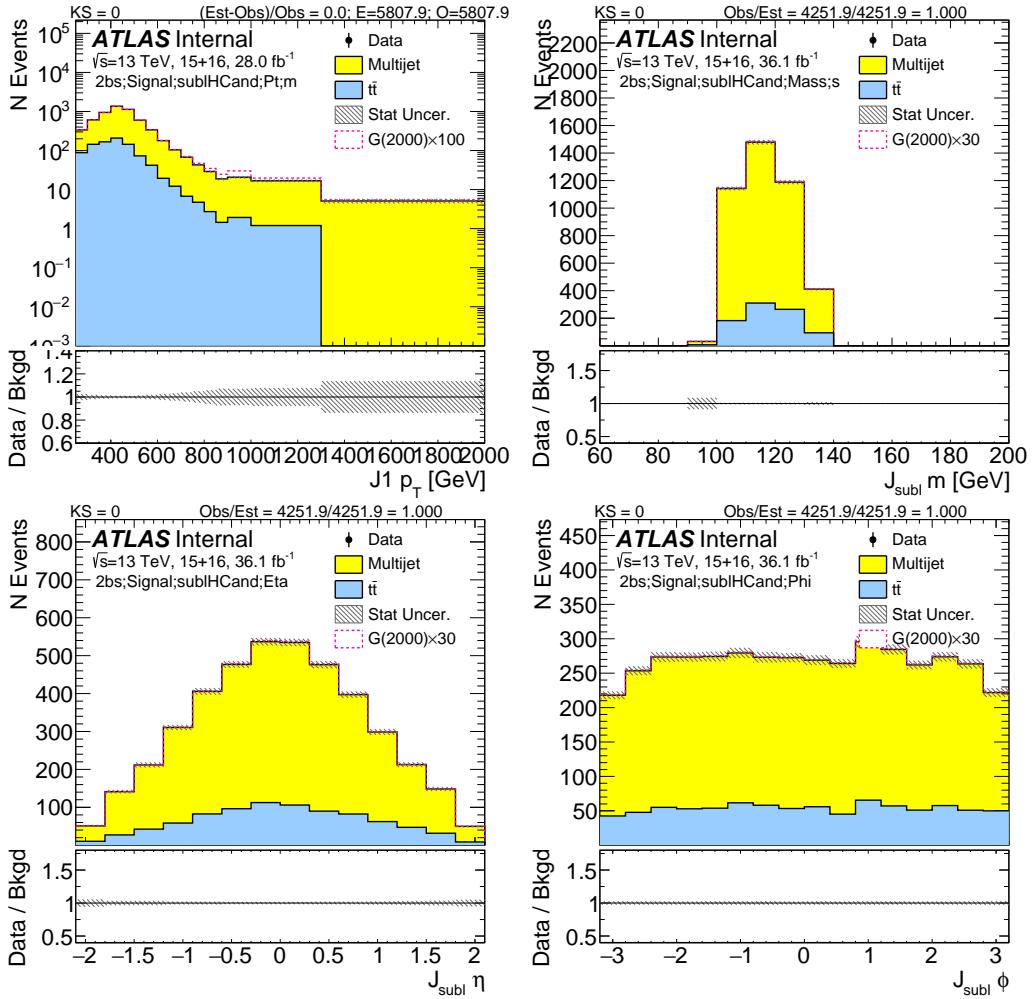
**Figure 7.47:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. Data is blinded, and will be added after unblinding.



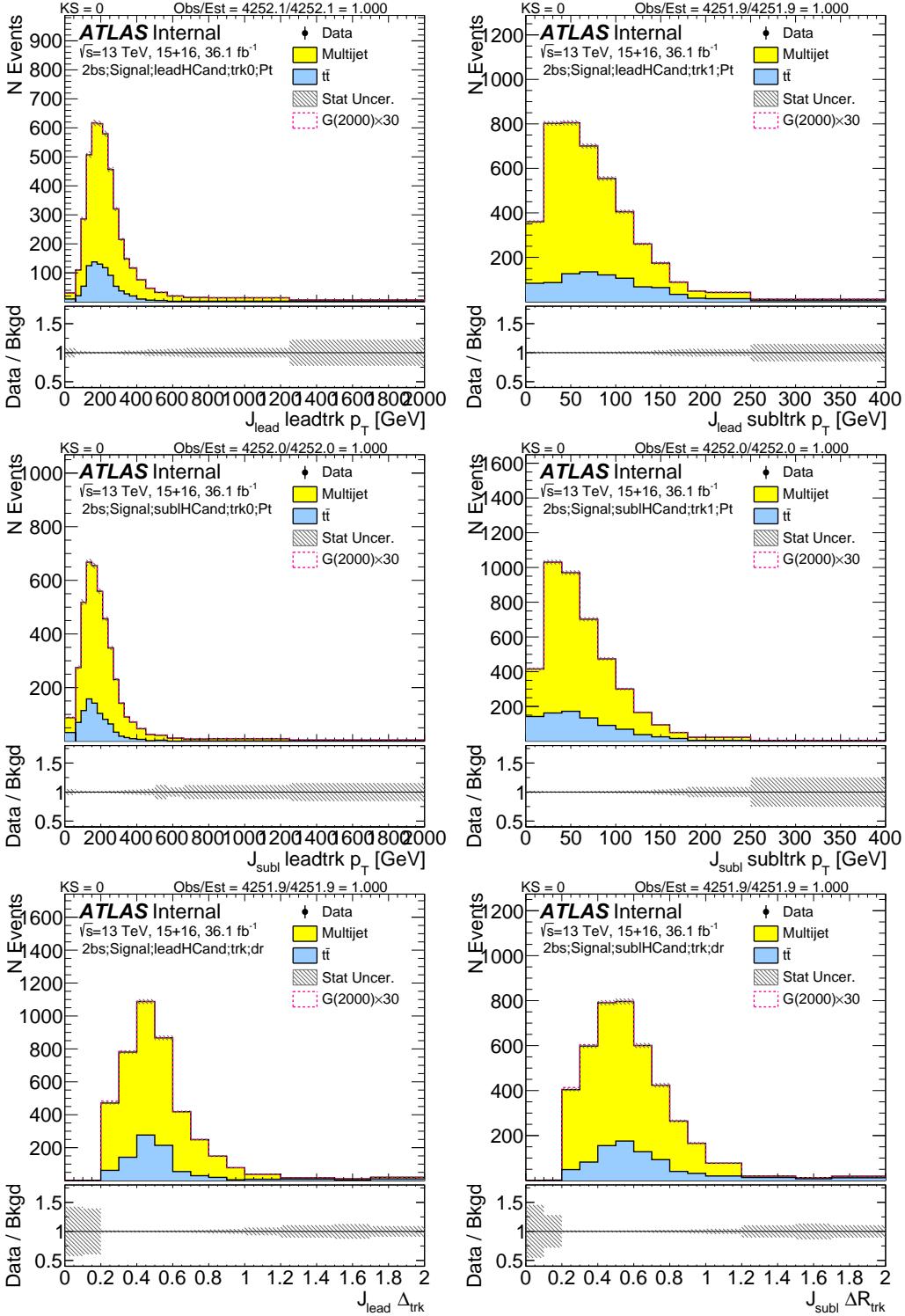
**Figure 7.48:** Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 3  $b$ -tags. Data is blinded, and will be added after unblinding.



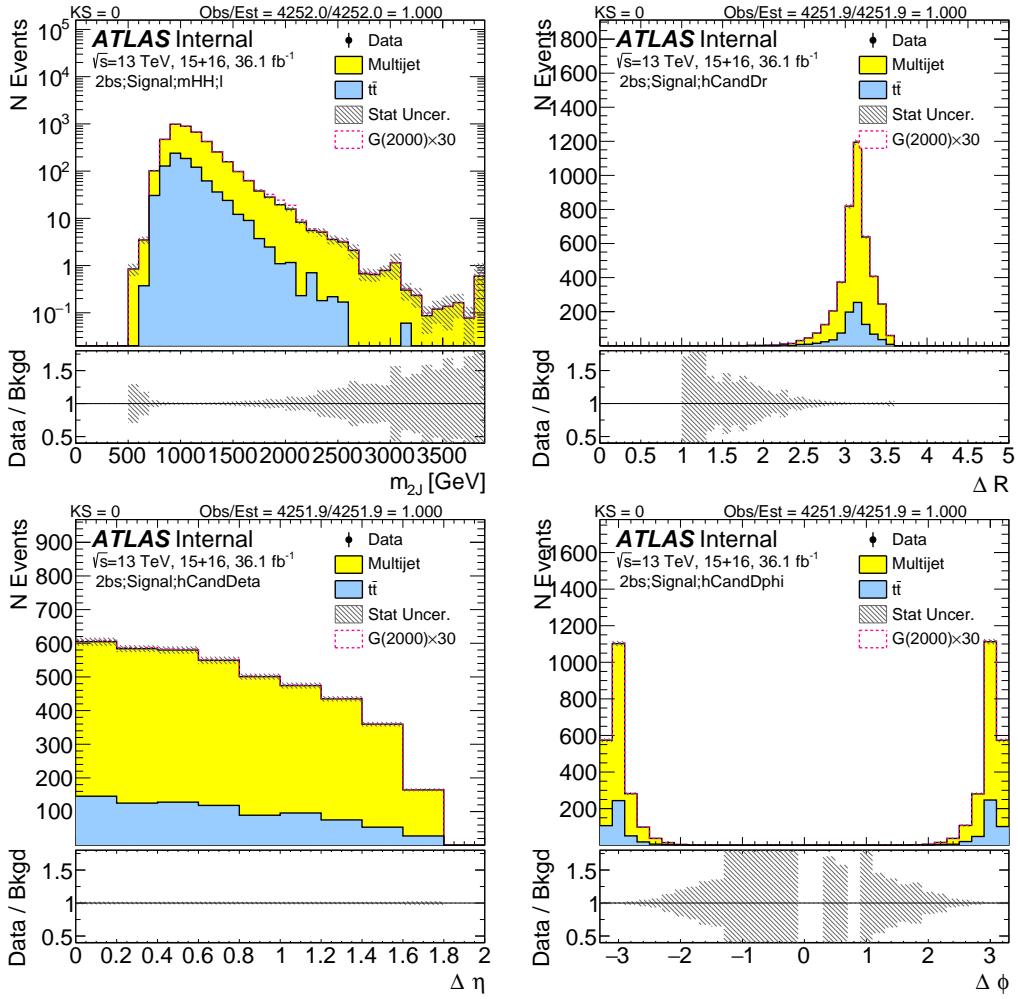
**Figure 7.49:** Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split. Data is blinded, and will be added after unblinding.



**Figure 7.50:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split. Data is blinded, and will be added after unblinding.



**Figure 7.51:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet. Data is blinded, and will be added after unblinding.



**Figure 7.52:** Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 2  $b$ -tags split. Data is blinded, and will be added after unblinding.

## SIGNAL REGION SMOOTHING

Due to the improved  $1/2b$  statistics at high di-large  $-R$  jet invariant mass above 1500 GeV and the limited  $t\bar{t}$  statistics above 1100 GeV, different fits are performed to smooth the di-large  $-R$  jet mass distribution in the signal region. The  $1/2b$  QCD background is fit with the MJ8 functional form:

$$y = \frac{a}{\frac{x^2}{\sqrt{s}}} \left(1 - \frac{x}{\sqrt{s}}\right)^{b-c} \log\left(\frac{x}{\sqrt{s}}\right) \quad (7.3)$$

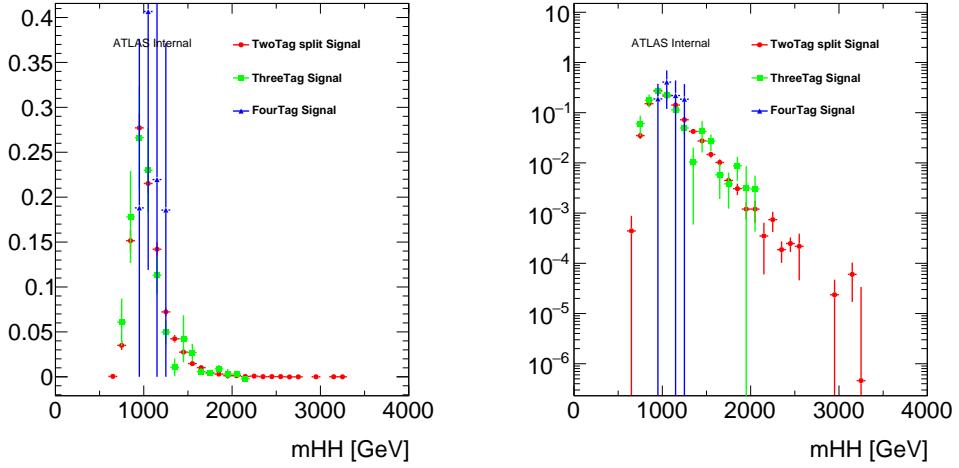
where  $\sqrt{s} = 13000$  GeV, in the range  $1200 < M_{JJ} < 3000$  GeV, and the three free parameters are  $a$ ,  $b$  and  $c$ . This form is used in fitting, as it was seen to be easier for the fits to converge. The signal region  $t\bar{t}$  distribution is fitted also with the dijet functional form, also in the range  $1200 < M_{JJ} < 3000$  GeV, without parameter constraints. The values of the estimated fit parameters in the  $4b$  and  $3b$  and  $2bs$  signal regions can be found in Table 7.3.

Given that the similar  $1/2b$  sample is used for deriving the QCD shape for the  $4/3/2bs$  signal regions, it is not surprising that the slope parameter ( $a$ ) is similar in the  $4/3/2bs$  signal regions for each the QCD backgrounds.

Due to the very limited statistics of the  $4b$   $t\bar{t}$  sample, the  $4b$   $t\bar{t}$  dijet mass shape is used from the  $3b$   $t\bar{t}$  dijet mass shape normalized to the number of  $4b$   $t\bar{t}$  events. A comparison of the shape is shown in Figure 7.53. Good agreement between the  $4b$  and  $3b$  signal region plot is shown.

Figure 7.54 shows the smoothing fits for the QCD background and the  $t\bar{t}$  background in the  $4b$  signal region. Figure 7.55 shows the same for the  $3b$  signal region. Figure 7.56 shows the same for the  $2bs$  signal region. The smoothing statistical uncertainties are also shown on these two plots. More additional uncertainties, such as uncertainty from choice of smoothing function, will be discussed in the Section 8.0.3.

The final smoothed background predictions for the  $4b$  and  $3b$  and  $2bs$  signal regions can be found in Figure 7.57. This includes smoothing statistical uncertainties only. More details on other system-



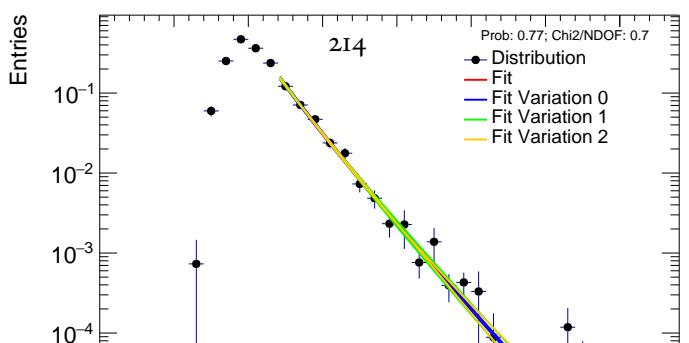
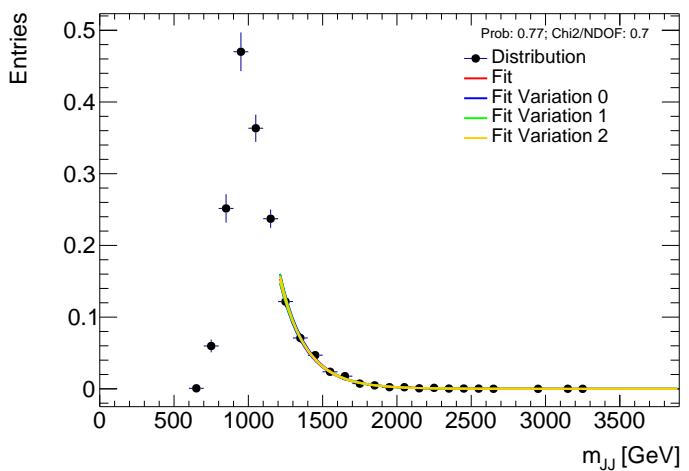
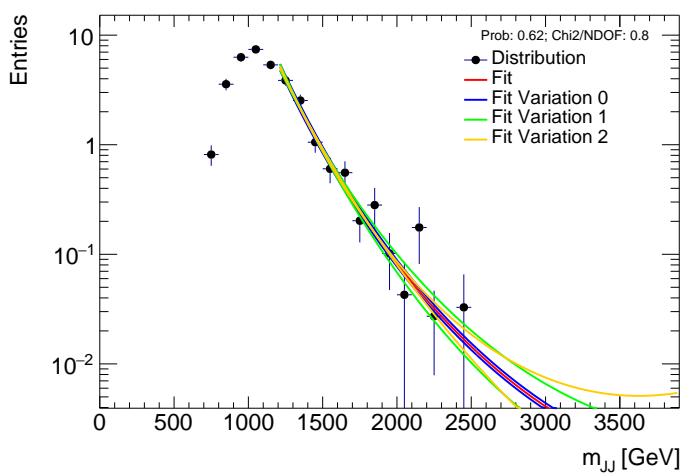
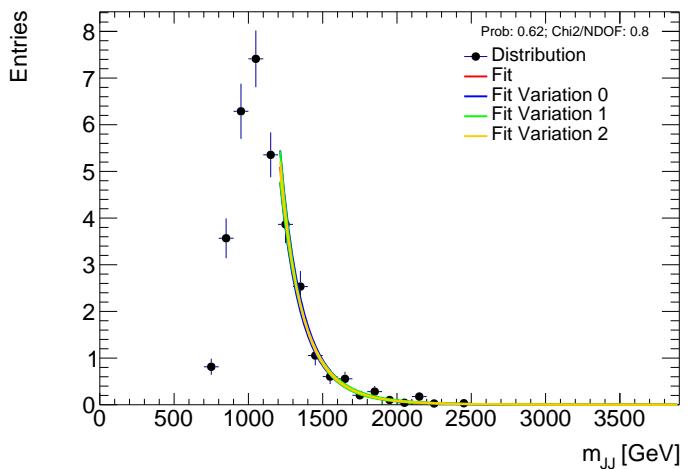
**Figure 7.53:** Comparison of the  $4b$ ,  $3b$  and  $2bs$  signal region  $t\bar{t}$  dijet mass shape. On the left is the linear scale, and on the right is the log scale. Both distributions are normalized to 1 for comparison.

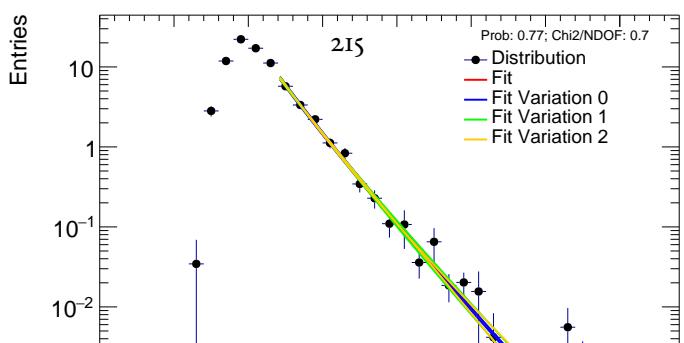
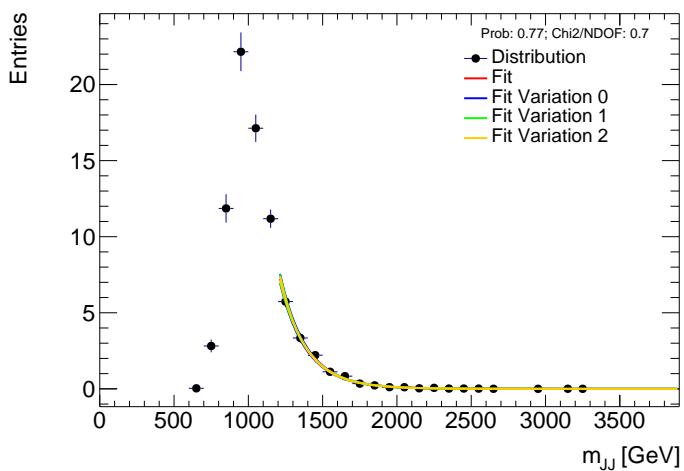
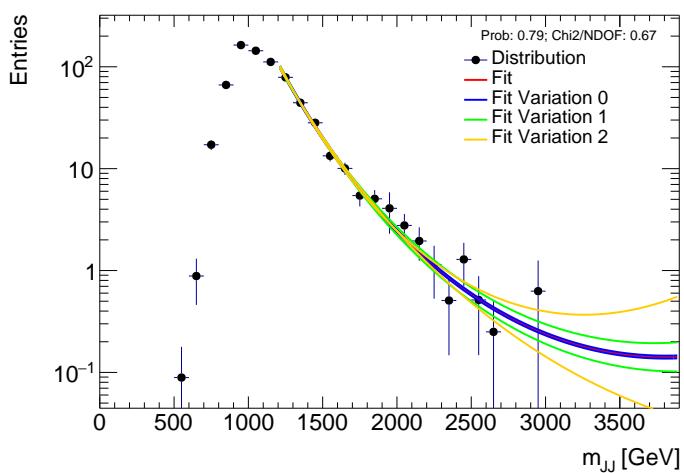
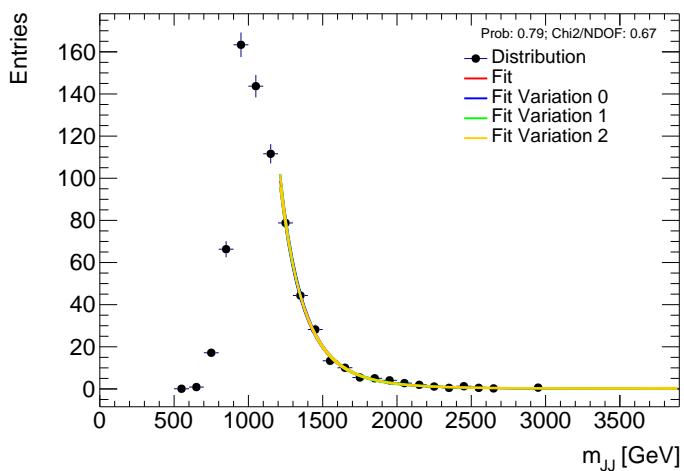
atics, including smoothing systematics, shape uncertainties and other sources of uncertainties would be discussed in Section 8.0.3.

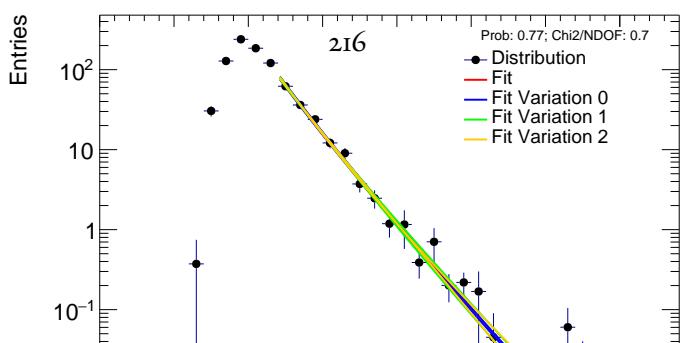
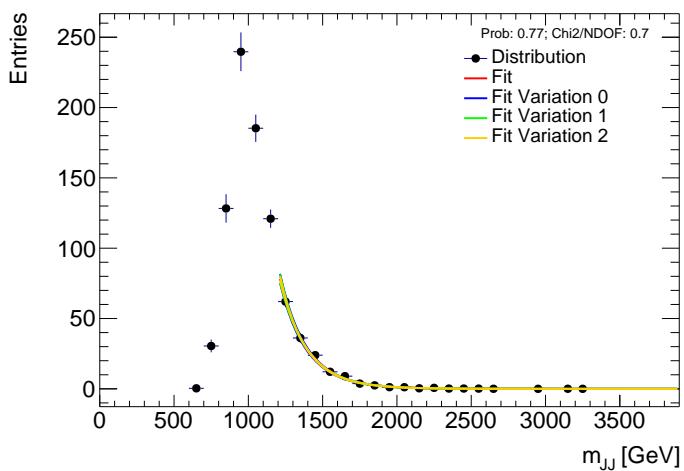
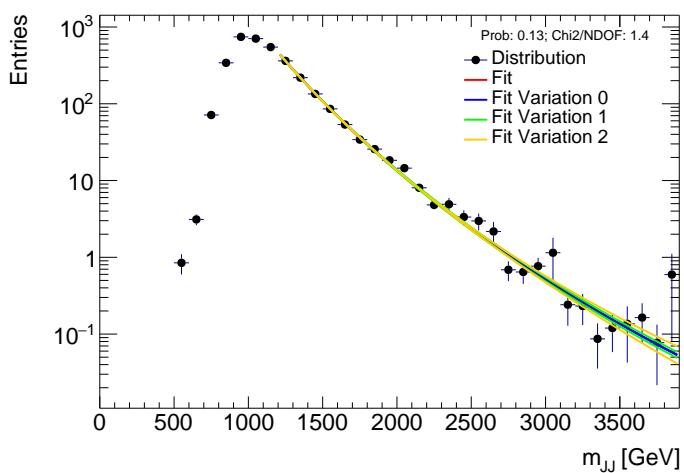
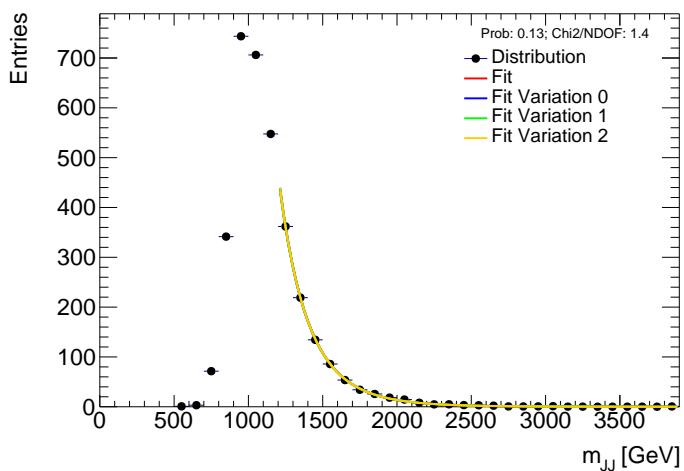
Uncertainties on the fit parameters are propagated as systematic uncertainties, though they are essentially replacing the bin-by-bin statistical uncertainties of the background estimates (which are not used once smoothing is applied). Correlations in the fit parameters of the backgrounds are taken into account when propagating the uncertainties, as described in Appendix ??.

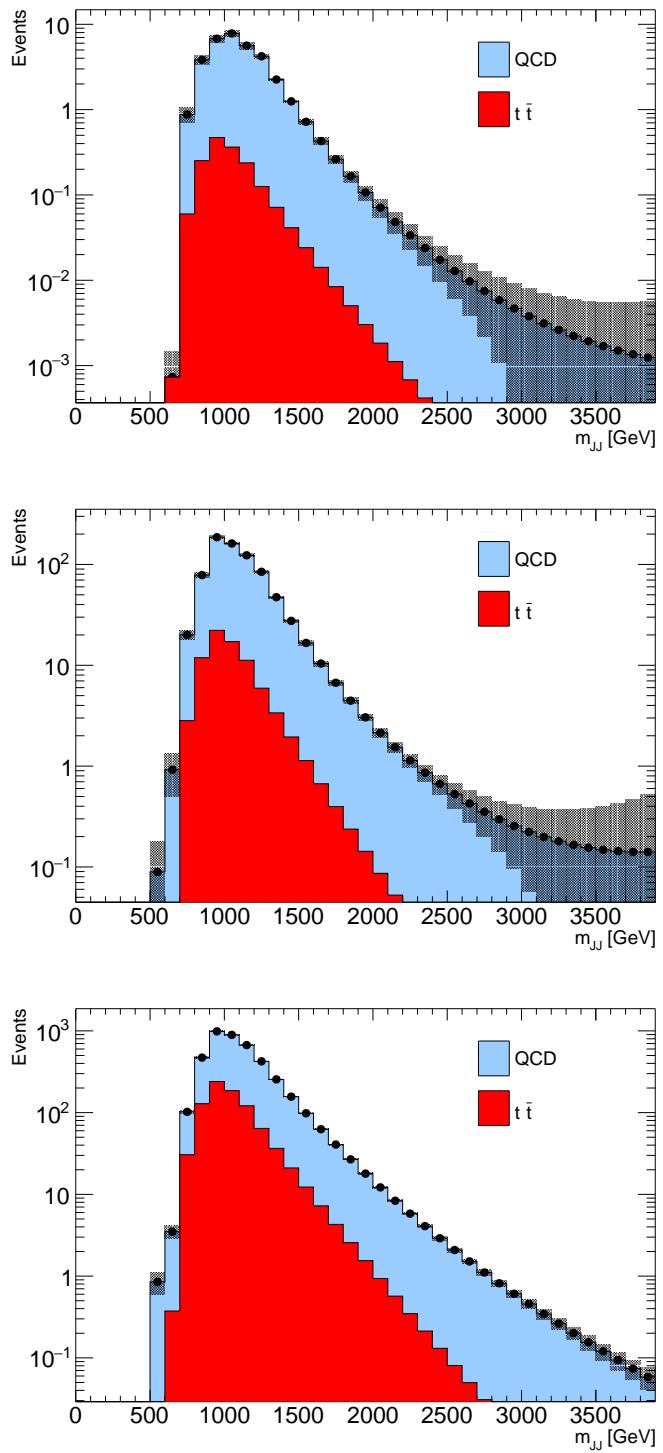
Region	$a_{t\bar{t}}$	$b_{t\bar{t}}$	$c_{t\bar{t}}$	$a_{qcd}$	$b_{qcd}$	$c_{qcd}$
FourTag	-1.02 $\pm$ 1.22	35.62 $\pm$ 10.83	9.05 $\pm$ 9.59	7.75 $\pm$ 1.8	-18.0 $\pm$ 16.75	54.36 $\pm$ 14.28
ThreeTag	2.83 $\pm$ 1.22	35.62 $\pm$ 10.83	9.05 $\pm$ 9.59	10.42 $\pm$ 1.73	-27.1 $\pm$ 15.78	56.91 $\pm$ 13.89
TwoTag split	5.21 $\pm$ 1.22	35.62 $\pm$ 10.83	9.05 $\pm$ 9.59	7.74 $\pm$ 0.36	7.22 $\pm$ 3.11	24.54 $\pm$ 2.78

**Table 7.3:** Smoothing parameters in  $4b$  and  $3b$  and  $2bs$  signal regions, the correlation between parameters is almost always 0.99.









**Figure 7.57:** Smoothed background estimations the  $4b$  (top),  $3b$  (middle), and  $2bs$  (bottom) signal regions. Only smoothing statistical uncertainties are shown here.

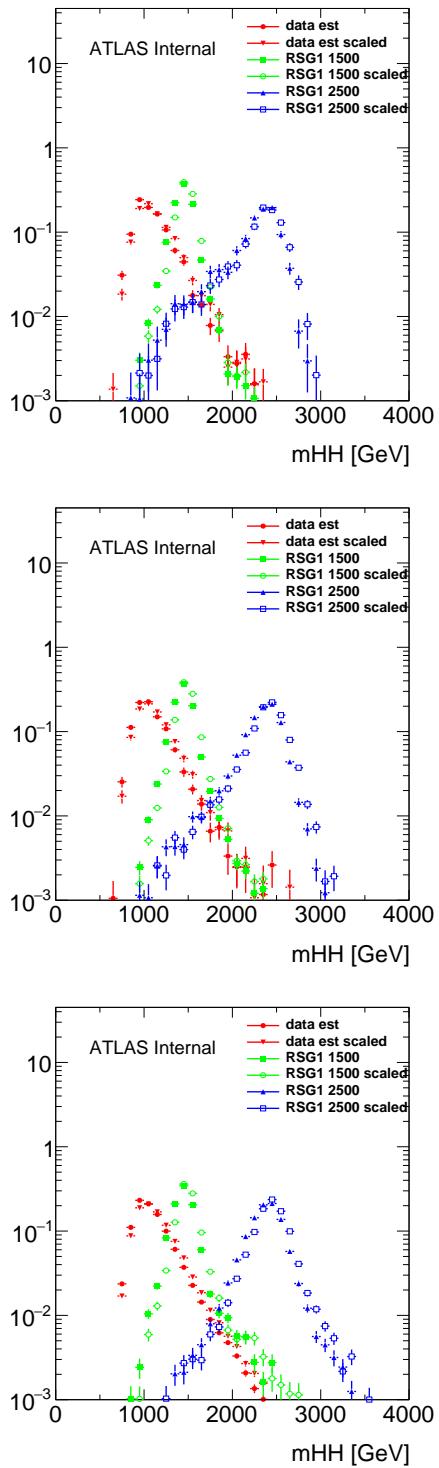
## SCALED DIJET MASS DISTRIBUTION IN SIGNAL REGION

As is done in the resolved analysis, we also consider the scaled  $M_{jj}$  distribution. In this case, the two higgs candidate 4-vectors are scaled by  $m_h/m_j$ , where  $m_h = 125$  GeV, and  $m_j$  is the *large – Rjet* mass of the Higgs candidate. While this distribution is expected to have less impact on the boosted analysis, because the mass correction is small relative to the signal masses being considered, we investigate this variable for possible improvements and for consistency with the resolved analysis. The scaled dijet mass distribution can be found in Figure 7.58. Its impact of the boosted analysis limit can be found in Appendix ??.

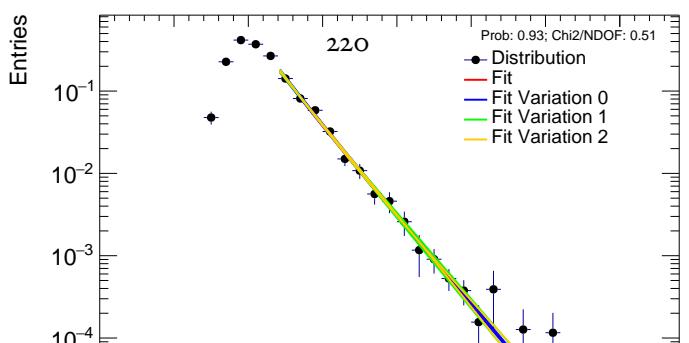
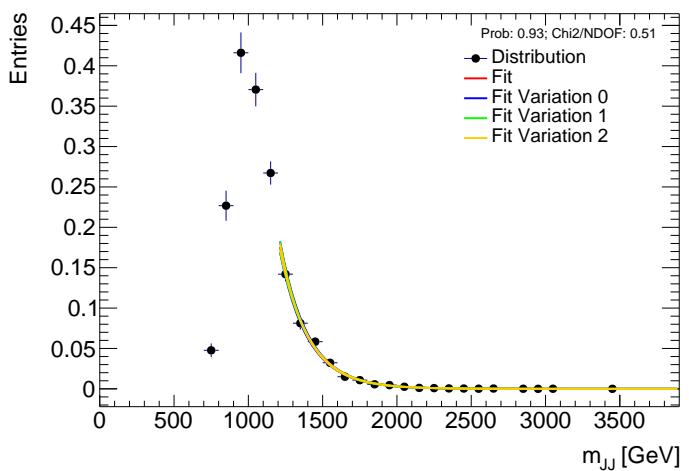
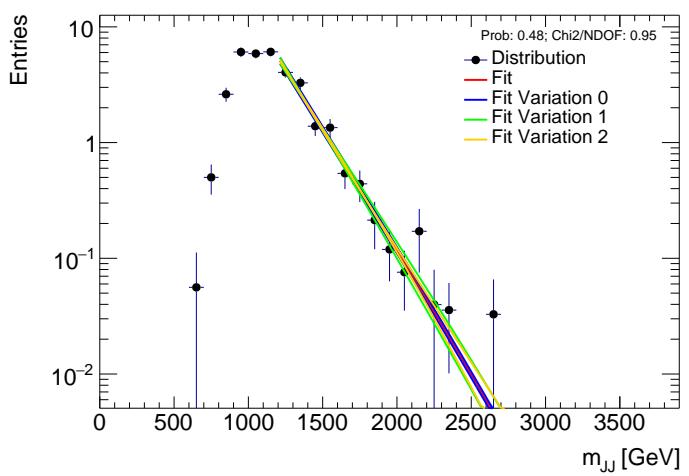
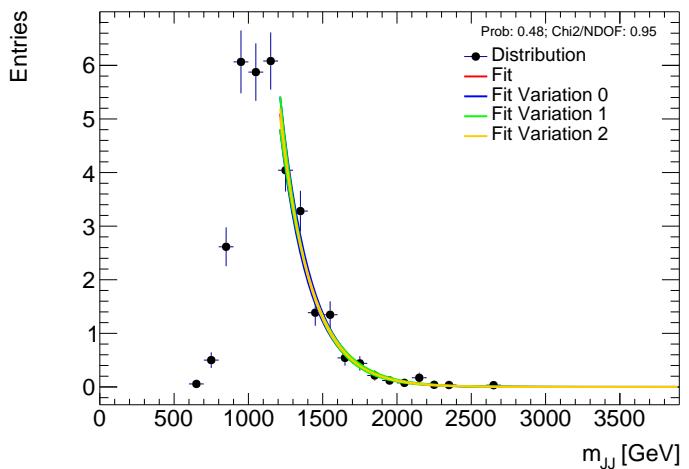
For determining the choice of the final discriminant, both the expected limits on the nominal and scaled dijet mass distribution have been computed. Since the scaled dijet mass distribution based limits are consistent (slightly better at low mass and slightly worse at high mass, with differences of the order of 10%) than the nominal dijet mass limits, we proceed to use the scaled dijet mass distribution for consistency with the resolved analysis.

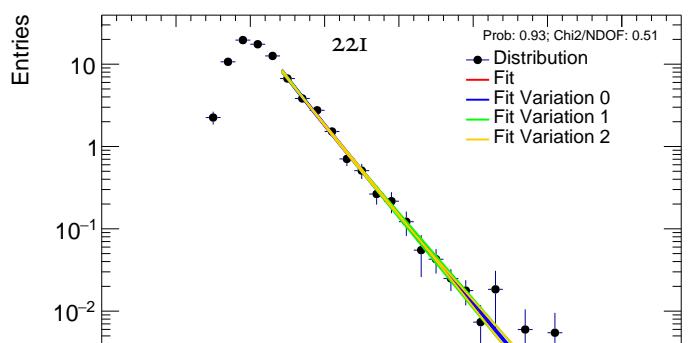
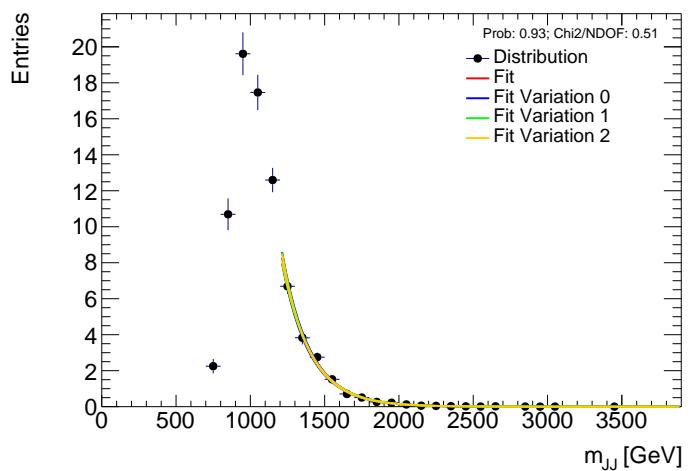
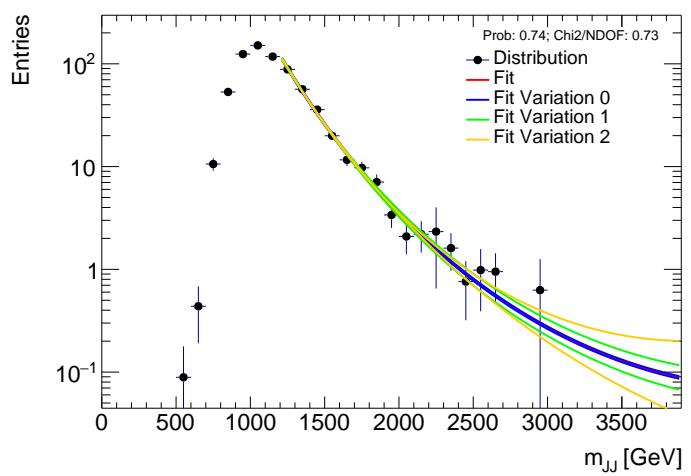
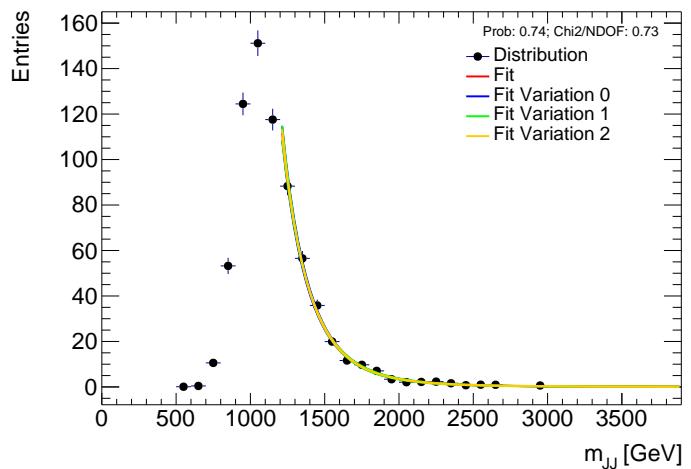
As is done for the dijet mass distribution, the scaled dijet mass distribution is smoothed. The smoothing is performed between 1200 GeV and 3000 GeV for the QCD and  $t\bar{t}$ . The smoothed distributions can be seen in Figures 7.59, 7.60, and 7.61. The values of the estimated fit parameters in the  $4b$  and  $3b$  and  $2bs$  signal regions can be found in Table 7.4.

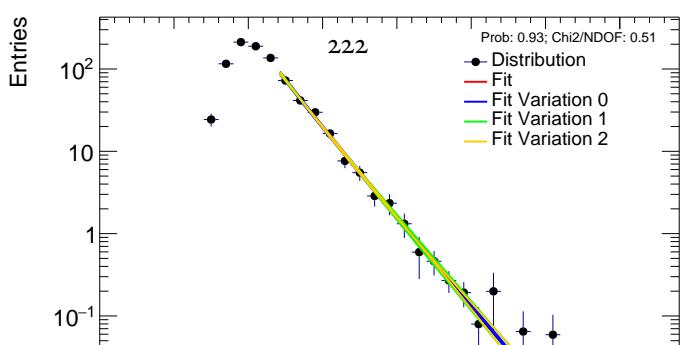
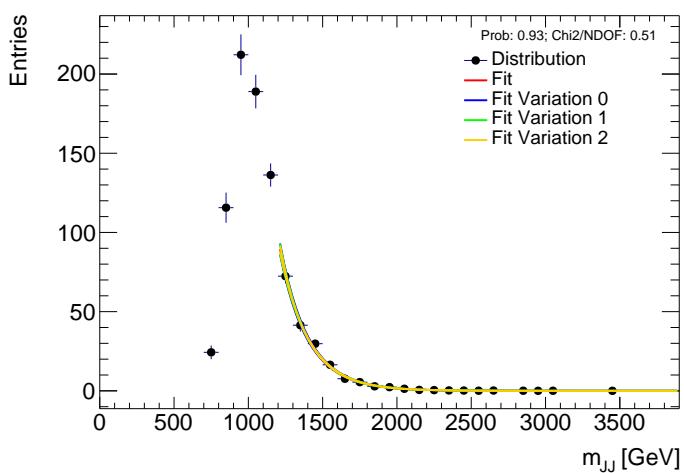
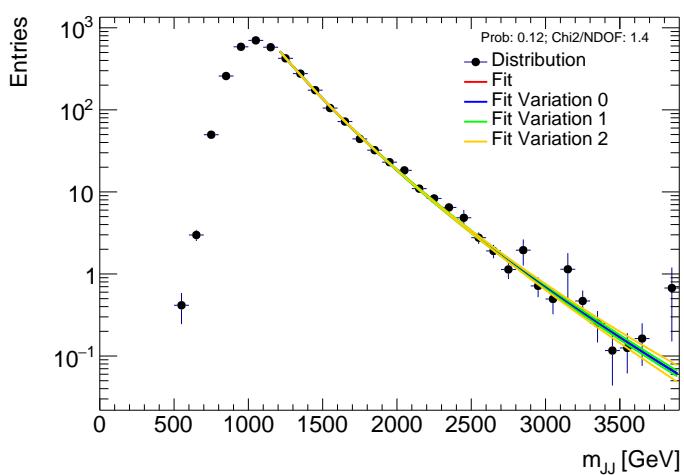
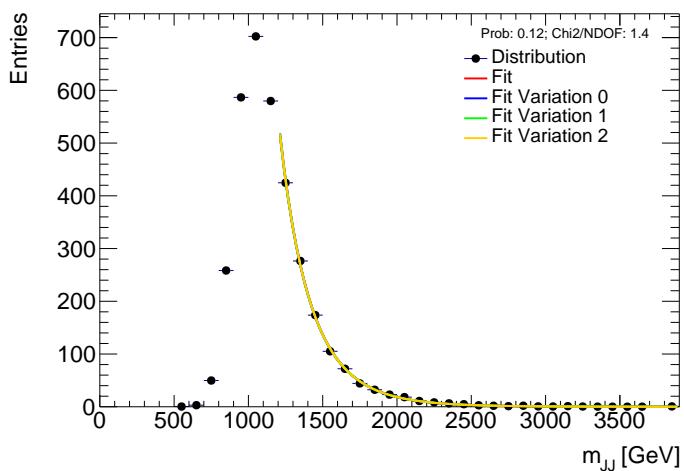
The final signal region prediction, using scaled di-jet mass distribution, with only statistical uncertainties, are shown in Figure 7.63(Figure 7.64) as before(after) smoothing.

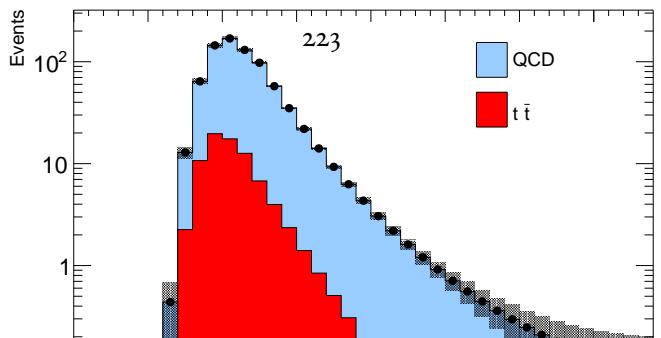
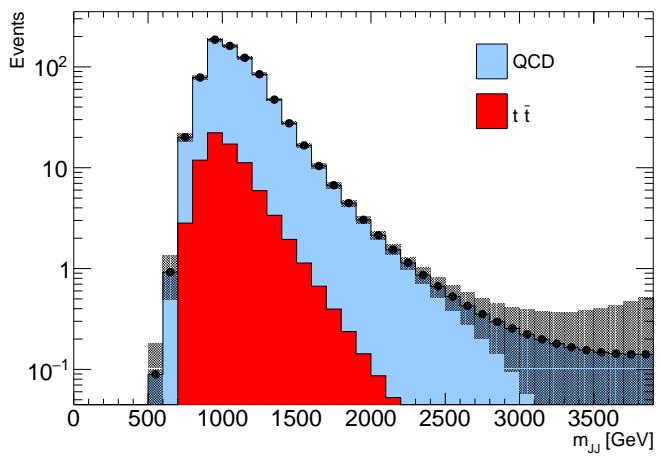
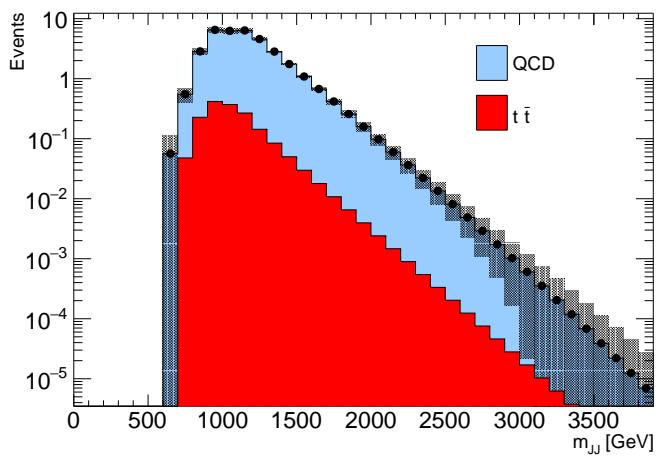
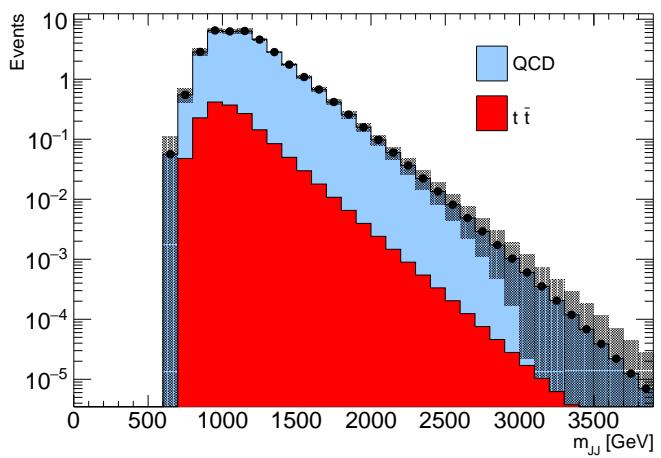


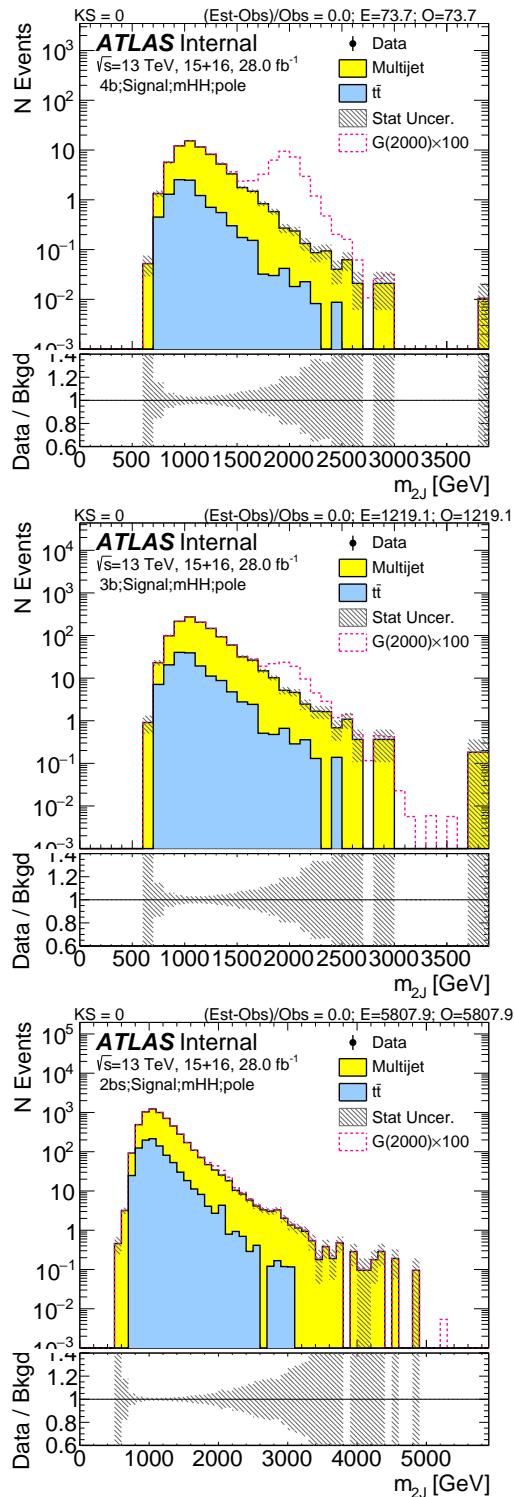
**Figure 7.58:** Normalized Scaled dijet mass distributions for the  $4b$  (top),  $3b$  (middle), and  $2bs$  (bottom) signal regions. For comparison, the unscaled distributions are shown on the same plot.



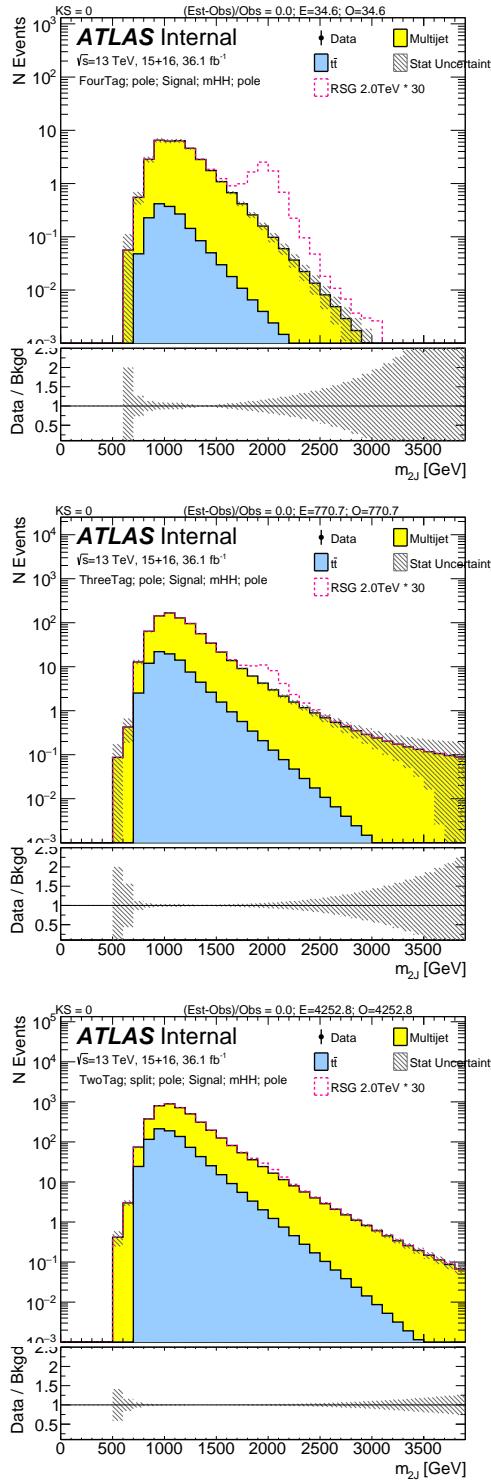








224  
**Figure 7.63:** Background prediction for  $4b$  (top),  $3b$  (middle), and  $2bs$  (bottom) signal region using scaled di-jet mass before smoothing. The uncertainty band includes only statistical uncertainties.



**225**  
**Figure 7.64:** Background prediction for  $4b$  (top),  $3b$  (middle), and  $2bs$  (bottom) signal region using scaled di-jet mass after smoothing. The uncertainty band includes only statistical uncertainties.

Region	$a_{t\bar{t}}$	$b_{t\bar{t}}$	$c_{t\bar{t}}$	$a_{qcd}$	$b_{qcd}$	$c_{qcd}$
FourTag	-2.02 ± 1.17	42.46 ± 9.87	1.31 ± 8.98	-0.49 ± 1.59	53.06 ± 15.2	-11.1 ± 12.8
ThreeTag	1.84 ± 1.17	42.45 ± 9.88	1.32 ± 8.98	8.51 ± 0.98	-13.8 ± 9.14	42.58 ± 7.87
TwoTag split	4.22 ± 1.17	42.45 ± 9.88	1.32 ± 8.98	7.06 ± 0.32	11.54 ± 2.77	19.05 ± 2.48

**Table 7.4:** Smoothing parameters in  $4b$  and  $3b$  and  $2bs$  signal regions for scaled mass distributions, the correlation between parameters is almost always 0.99.

*Madness and genius are separated only by degrees of  
success.*

Tony

# 8

## Systematics

All backgrounds are data-driven, with the exception of  $Z+jets$ , which is very small. The  $t\bar{t}$  simulation is only used to determine the shape of expected events, where the  $t\bar{t}$  normalization is derived in data. MC uncertainties are applied to all simulated samples and propagated anywhere a MC based shape is used.

### 8.0.1 MC UNCERTAINTIES

MC based uncertainties are propagated in the analysis using standard CP group recommendations. These uncertainties can change both the shape and normalization of the signal and of the MC-based background prediction ( $Z+jets$ ). The multijet and  $t\bar{t}$  backgrounds normalisations are estimated with a data driven method (the likelihood fit). Since the shape for the  $t\bar{t}$  component is taken from MC,

the fit is redone for each MC variations.

**LUMINOSITY UNCERTAINTY** : The uncertainty on the combined 2015+2016 integrated luminosity is 2.1%, assuming uncorrelated uncertainties between years. It is derived, following a methodology similar to that detailed in Refs.<sup>?</sup> and<sup>?</sup>, from a preliminary calibration of the luminosity scale using x-y beam-separation scans performed in August 2015 and May 2016. This uncertainty is applicable to the backgrounds with normalizations determined from simulation, and further propagated to the multijet prediction through data-driven background estimation procedure. It is expected to have a small impact on this analysis. This uncertainty is also applied to the signal normalization prediction

**LARGE-R JET RESOLUTION AND SCALE UNCERTAINTIES** : The uncertainties on the jet energy and mass (JES, JMS scale) are evaluated by the combined performance groups using track-to-calorimeter double ratios between data and MC, measured in dijet data<sup>??</sup>. Discrepancies observed between data and MC are assigned as uncertainties on the energy/mass scales of the jet. Different correlation scenarios are supported by the Jet Substructure group, where all uncertainties can be decorrelated, fully correlated, or only correlated between energy and mass. Currently the latter is being used, known as the “medium” configuration. The uncertainties on the jet energy, mass resolutions are estimated by applying a Gaussian smearing which degrades the nominal resolution by an absolute 2% for  $p_T$ , relative 20% for mass.

The uncertainties in our kinematic regime for signal yield predictions are below 7% for JES/JMS. For  $t\bar{t}$  yield predictions the uncertainties are  $\sim$ 7-24% for JES/JMS. Details are listed in section 8.0.4. The uncertainties in our kinematic regime for signal yield predictions are below 7% / 15% for JER/JMR. For  $t\bar{t}$  yield predictions the uncertainties are  $\sim$ 4-27% for JER/JMR. Details are listed in section 8.0.4.

**B-TAGGED TRACK JET SCALE FACTOR UNCERTAINTIES** The uncertainties related to the  $b$ -tagging efficiency calibrations as measured in  $t\bar{t}$  events for track-jets are considered, using the official pre-

scriptions. The procedure to define these calibrations is similar to that described in reference<sup>48</sup>.

The effect of the different experimental uncertainties on the signal yield is shown in section 8.0.4.

The signal yield uncertainty due to  $b$ -tagging is less than 30% for the signal, and less than 12% for the  $t\bar{t}$  background yield. The main difference with respect to the previous result is on the  $b$ -tagging uncertainties, which have been reduced by approximately 50%. The total effect of the  $b$ -tagging uncertainty on the expected limits is shown in Sec. ??, along with other uncertainties.

## $t\bar{t}$ MC UNCERTAINTY

In addition to the  $t\bar{t}$ fit uncertainties, following the recommendations, extra  $t\bar{t}$ MC samples are used with different variations: Hadronization, Fragmentation, Matrix Element and Additional Radiation. The top quark mass variations are also considered. The MC samples used are:

```

mc15_13TeV.410001.PowhegPythiaEvtGen_P2012radHi_ttbar_hdamp345_down_nonallhad.merge.DAOD_EXOT8.e3783_s2608_r7725_r7676_p2949
mc15_13TeV.410002.PowhegPythiaEvtGen_P2012radLo_ttbar_hdamp172_up_nonallhad.merge.DAOD_EXOT8.e3783_s2608_r7725_r7676_p2949
mc15_13TeV.410003.aMcAtNloHerwigppEvtGen_ttbar_nonallhad.merge.DAOD_EXOT8.e4441_s2726_r7772_r7676_p2949
mc15_13TeV.410004.PowhegHerwigppEvtGen_UEEE5_ttbar_hdamp172p5_nonallhad.merge.DAOD_EXOT8.e3836_a766_a821_r7676_p2949
mc15_13TeV.410008.aMcAtNloHerwigppEvtGen_ttbar_allhad.merge.DAOD_EXOT8.e3964_s2726_r7772_r7676_p2949
mc15_13TeV.410022.Sherpa_CT10_ttbar_SingleLeptonP_MEPS_NLO.merge.DAOD_EXOT8.e3957_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410022.Sherpa_CT10_ttbar_SingleLeptonP_MEPS_NLO.merge.DAOD_EXOT8.e3959_a766_a818_r7676_p2949
mc15_13TeV.410023.Sherpa_CT10_ttbar_SingleLeptonM_MEPS_NLO.merge.DAOD_EXOT8.e3957_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410023.Sherpa_CT10_ttbar_SingleLeptonM_MEPS_NLO.merge.DAOD_EXOT8.e3959_a766_a818_r7676_p2949
mc15_13TeV.410024.Sherpa_CT10_ttbar_AllHadron_MEPS_NLO.merge.DAOD_EXOT8.e3957_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410024.Sherpa_CT10_ttbar_AllHadron_MEPS_NLO.merge.DAOD_EXOT8.e3959_a766_a818_r7676_p2949
mc15_13TeV.410037.PowhegPythiaEvtGen_P2012_ttbar_hdamp170_nonallhad.merge.DAOD_EXOT8.e4529_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410038.PowhegPythiaEvtGen_P2012_ttbar_hdamp171p5_nonallhad.merge.DAOD_EXOT8.e4529_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410039.PowhegPythiaEvtGen_P2012_ttbar_hdamp173p5_nonallhad.merge.DAOD_EXOT8.e4529_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410040.PowhegPythiaEvtGen_P2012_ttbar_hdamp175_nonallhad.merge.DAOD_EXOT8.e4529_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410041.PowhegPythiaEvtGen_P2012_ttbar_hdamp177p5_nonallhad.merge.DAOD_EXOT8.e4529_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410042.PowhegPythiaEvtGen_P2012_ttbar_hdamp170_allhad.merge.DAOD_EXOT8.e4510_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410043.PowhegPythiaEvtGen_P2012_ttbar_hdamp171p5_allhad.merge.DAOD_EXOT8.e4510_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410044.PowhegPythiaEvtGen_P2012_ttbar_hdamp173p5_allhad.merge.DAOD_EXOT8.e4510_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410045.PowhegPythiaEvtGen_P2012_ttbar_hdamp175_allhad.merge.DAOD_EXOT8.e4510_s2608_s2183_r7725_r7676_p2949

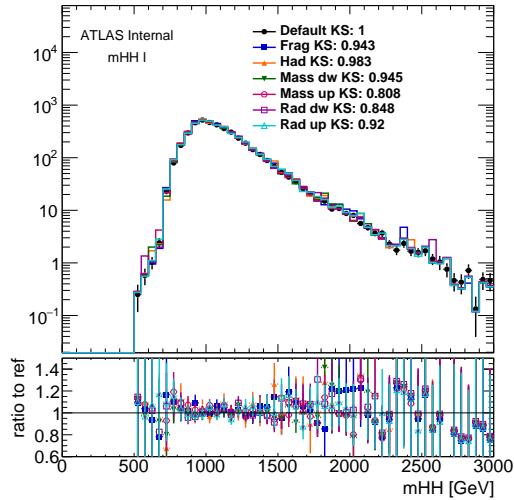
```

```

mc15_13TeV.410046.PowhegPythiaEvtGen_P2012_ttbar_hdamp177p5_allhad.merge.DAOD_EXOT8.e4510_s2608_s2183_r7725_r7676_p2949
mc15_13TeV.410161.PowhegPythiaEvtGen_P2012radHi_ttbar_hdamp345_down_allhad.merge.DAOD_EXOT8.e4837_s2726_r7772_r7676_p2949
mc15_13TeV.410162.PowhegPythiaEvtGen_P2012radLo_ttbar_hdamp172p5_up_allhad.merge.DAOD_EXOT8.e4837_s2726_r7772_r7676_p2949
mc15_13TeV.410163.PowhegHerwigppEvtGen_UEEE5_ttbar_hdamp172p5_allhad.merge.DAOD_EXOT8.e4836_s2726_r7772_r7676_p2949

```

These  $t\bar{t}$ samples are used to replace the normal had and nonhad MCs, stiched with Mtt slices samples, and the variation in the  $t\bar{t}$ yield and background predictions are considered. The variation in total background, with different  $t\bar{t}$ MC sample as input is tested. This is shown in Figure 8.1. Therefore, this uncertainty is considered limited by the MC statistics and dropped from the final list.



**Figure 8.1:** Total background estimation ( $\text{qcd} + t\bar{t}$ ) with different  $t\bar{t}$ MC variations. The different variations agree with the default within the statistical uncertainties.

### 8.0.2 THEORETICAL UNCERTAINTIES

The theoretical uncertainties on the acceptance times efficiency ( $\mathcal{A} \times \varepsilon$ ) are evaluated by analysis of specially-generated, particle-level signal samples. The generation of these samples follows the

configuration of the baseline samples, but with modifications to probe the following theoretical uncertainties: uncertainties in the parton density functions; uncertainties due to missing higher order terms in the matrix elements; and uncertainties in the modelling of the underlying event (including multi-parton interactions), of hadronic showers and of initial and final state radiation. Each of the signals is tested.

The estimation of the theoretical uncertainties is performed using a Rivet-based analysis, which replicates the full analysis selection outlined in Section 5. The most important detector effects – b-tagging efficiency and jet mass resolution – are emulated. Jet mass resolution is emulated by smearing the particle-level  $R = 1.0$  jet masses, using the resolutions estimated in<sup>5</sup>. B-tagging efficiency is treated using a truth-tagging approach, which weights events according to the combinatoric probabilities of which jets are b-tagged, using the measured b-tagging efficiencies from the CDI file.

Reasonable agreement is observed between the acceptance times efficiency of the particle-level analysis and of the full, reconstruction-level analysis when measured on independent samples generated using the same configuration (Figure ??), although there are clearly discrepancies. Perfect agreement is not necessary, since the theoretical uncertainties will be calculated using the relative change in  $\mathcal{A} \times \varepsilon$  between variations of the signal sample, as measured by the Rivet-based analysis.

To evaluate the potential effect of missing higher order terms in the matrix element, the renormalisation and factorisation scales used in the signal generation were varied coherently by factors of  $0.5\times$  and  $2\times$  for the signals. The effect is shown as function of resonance mass in Figure ??.

Uncertainties due to modelling of the parton shower and the underlying event (including multi-parton interactions) are evaluated by switching the MC generator used. For the Bulk RS graviton samples, this means switching from Pythia 8 to Herwig++, while for the scalar and non-resonant it is Herwig++ to Pythia 8. Figure ?? shows the impact of these variations on the signal acceptance.

PDF uncertainties are evaluated using the PDF4LHC15\_nlo\_mc set, which combines CT14, MMHT14 and NNPDF3.0 PDF sets<sup>3</sup>. The uncertainty is evaluated by calculating the acceptance

for each PDF replica. The standard deviation of these acceptance values divided by the baseline acceptance is taken as the PDF uncertainty. For each mass point the distribution of these ratio is compatible with a Gaussian centred on one. The calculated PDF uncertainty is shown in Figure ?? as upward and downward shifts from unity. The uncertainty in acceptance due to PDF uncertainties is less than  $\pm 1\%$  across the full mass range considered for the analysis. For this reason, it is neglected in the statistical analysis described in Section ??.

These uncertainties are implemented in the final statistical analysis as normalisation uncertainties on the signals, with the value taken from the polynomial fit. This smooths out statistical fluctuations and allows interpolation between the generated mass points, if needed.

### 8.0.3 BACKGROUND PREDICTION UNCERTAINTY

A statistical uncertainty on the value of  $\mu_{\text{multijet}}$  for the  $4b$  ( $3b, 2bs$ ) Signal Region was determined from the fitting procedure described in Section 7.1.5.

The statistical uncertainty of the  $t\bar{t}$  normalization is accounted for through the uncertainties on  $\alpha_{t\bar{t}}$  from the fit to data, as described in Section 7.1.5. The statistical uncertainty of 69% on  $\alpha_{t\bar{t}}$  are 79% anti-correlated to the value of  $\mu_{\text{multijet}}$  found in the fitting procedure in the  $4 b$ -tag region, the uncertainties of 9% on  $\alpha_{t\bar{t}}$  are 76% anti-correlated to the value of  $\mu_{\text{multijet}}$  found in the fitting procedure in the  $3 b$ -tag region, and the uncertainties of 2.6% on  $\alpha_{t\bar{t}}$  are 75% anti-correlated to the value of  $\mu_{\text{multijet}}$  found in the fitting procedure in the  $2bs$ -tag region.

The background systematic uncertainties in the signal region are divided into the following components:

- Non-closure uncertainty on  $\mu_{\text{multijet}}$  found by comparing the value derived from the sideband to the control region normalization.
- Effects on the QCD prediction from variations of the SideBand and Control Region Definitions

- The impact of the shape uncertainty of the  $t\bar{t}$  distribution in the  $4b$  signal region.
- The impact of the shape uncertainty of the  $1/2b$  QCD distribution derived in the control region.
- The impact of the smoothing function fit range and function choice on the QCD prediction

The  $t\bar{t}$  normalization uncertainty on  $\alpha_{t\bar{t}}$  is derived in the fit to data as described in Section 7.1.5.

This has a negligible impact on the signal sensitivity, but is still propagated as the uncertainty in the  $t\bar{t}$  normalization in the signal region.

#### NON-CLOSURE UNCERTAINTY ON $\mu_{\text{multijet}}$ DETERMINED IN THE CONTROL REGION

A further uncertainty is derived by comparing the value of  $\mu_{\text{multijet}}$  to the overall difference between predicted to observed events in the control region. While the total predicted background (showing stat error only) of  $4b$ :  $76.7 \pm 5.4$  vs obs  $81.0$ ,  $3b$ :  $1565.6 \pm 18.1$  vs obs  $1553.0$ ,  $2bs$ :  $8332.4 \pm 38.8$  vs obs  $8486.0$ , the number events agrees with the total data in the control region within statistical error, we consider an added systematic on the background prediction normalization, taken as the maximum between either the difference between the central value of the prediction to the observed number of events ( $4.3$  events, or  $5\%$ , for  $4b$ ;  $13$  events, or  $1\%$ , for  $3b$ ;  $154$  events, or  $2\%$ , for  $2bs$ ) or the statistical uncertainty of the observed  $4b$  ( $3b$ ) data in the CR ( $11.1\%$  for  $4b$ ;  $2.5\%$  for  $3b$ ; and  $1.1\%$  for  $2bs$ ). For the detailed numbers, please refer to section ??.

Although we have derived our non-closure uncertainty on  $\mu_{QCD}$  from comparison between data and prediction in the control region, we need to test how this number is sensitive to our choice of control region (CR) and sideband region (SB). In addition, we also want to check how our background prediction in signal region is sensitive to the choice of control region and sideband region. All these tests are done on the control/sideband regions after the full reweighting procedure as described in 7.1.6, while applying the nominal reweighting values.

Besides the nominal control region as described above, we design three additional control regions, as illustrated in Figure ??, ??, ??, ??, ??, ??, ??:

- Low-mass CR: the center position of the circle that defines nominal CR is moved down by 3 GeV, in both leading and sub-leading large-jet mass.
- High-mass CR: the center position of the circle that defines nominal CR is moved up by 3 GeV, in both leading and sub-leading large-R jet mass.
- Signal-depletion (Small) CR: the  $X_{bb}$  cut that defines signal region is increased to 2.0 from 1.6. This variation only affect CR, while SR remains unchanged (i.e. signal region is still defined by  $X_{bb} < 1.6$ , while CR is defined as  $X_{bb} > 2.0$  and  $R_{bb} < 33$ ).
- High-mass SB: The signal region and control region remain unchanged, the center position of the circle that defines nominal SB is moved up by 3 GeV in both leading and sub-leading large-R jet mass.
- Low-mass SB: The signal region and control region remain unchanged, the center position of the circle that defines nominal SB is moved up by 3 GeV in both leading and sub-leading large-R jet mass.
- Large SB: The signal region and control region remain unchanged, while the SB is  $33 < R_{bb}$  and  $R_{bb}^{\text{high}} < 61$ .  $\mu_{QCD}$  will change.
- Small SB: The signal region and control region remain unchanged, while the SB is  $33 < R_{bb}$  and  $R_{bb}^{\text{high}} < 55$ .  $\mu_{QCD}$  will change.

The results are summarized in Table 8.1, 8.2 and 8.3, while the details are presented in the Appendix ???. Based on all the variations, a 2.8% normalization uncertainty is assigned to  $2bs$  region, 4.2% to  $3b$  region (which is the statistical uncertainty), and a 12.2% normalization uncertainty is assigned to  $4b$  region.

CR Variations FourTag	Data	Prediction	(Predict - Data)/Data
Nominal	$81.0 \pm 9.0$	$76.77 \pm 5.43$	$-5.22\% \pm 17.23\%$
CR High	$76.0 \pm 8.72$	$71.12 \pm 5.41$	$-6.43\% \pm 17.85\%$
CR Low	$91.0 \pm 9.54$	$79.87 \pm 5.45$	$-12.2\% \pm 15.19\%$
CR Small	$58.0 \pm 7.62$	$55.96 \pm 5.35$	$-3.52\% \pm 21.89\%$
SB Large	$81.0 \pm 9.0$	$74.71 \pm 5.4$	$-7.76\% \pm 16.91\%$
SB Small	$81.0 \pm 9.0$	$74.15 \pm 5.38$	$-8.45\% \pm 16.81\%$
SB High	$81.0 \pm 9.0$	$78.72 \pm 5.46$	$-2.82\% \pm 17.54\%$
SB Low	$81.0 \pm 9.0$	$76.51 \pm 5.38$	$-5.54\% \pm 17.14\%$

**Table 8.1:** Agreement between data and prediction in 4b tag CR. Showing stat uncertainty only.

CR Variations ThreeTag	Data	Prediction	(Predict - Data)/Data
Nominal	$1553.0 \pm 39.41$	$1587.04 \pm 21.4$	$2.19\% \pm 3.97\%$
CR High	$1461.0 \pm 38.22$	$1473.89 \pm 20.77$	$0.88\% \pm 4.06\%$
CR Low	$1628.0 \pm 40.35$	$1697.38 \pm 21.75$	$4.26\% \pm 3.92\%$
CR Small	$1134.0 \pm 33.67$	$1127.34 \pm 17.66$	$-0.59\% \pm 4.51\%$
SB Large	$1553.0 \pm 39.41$	$1574.23 \pm 21.47$	$1.37\% \pm 3.95\%$
SB Small	$1553.0 \pm 39.41$	$1601.44 \pm 21.64$	$3.12\% \pm 4.01\%$
SB High	$1553.0 \pm 39.41$	$1602.74 \pm 21.48$	$3.2\% \pm 4.0\%$
SB Low	$1553.0 \pm 39.41$	$1576.56 \pm 21.5$	$1.52\% \pm 3.96\%$

**Table 8.2:** Agreement between data and prediction in 3b tag CR. Showing stat uncertainty only.

CR Variations TwoTag split	Data	Prediction	(Predict - Data)/Data
Nominal	$8486.0 \pm 92.12$	$8332.97 \pm 38.84$	$-1.8\% \pm 1.52\%$
CR High	$8174.0 \pm 90.41$	$7937.59 \pm 39.61$	$-2.89\% \pm 1.56\%$
CR Low	$8907.0 \pm 94.38$	$8800.86 \pm 39.51$	$-1.19\% \pm 1.49\%$
CR Small	$5999.0 \pm 77.45$	$5873.52 \pm 32.31$	$-2.09\% \pm 1.8\%$
SB Large	$8486.0 \pm 92.12$	$8341.7 \pm 38.44$	$-1.7\% \pm 1.52\%$
SB Small	$8486.0 \pm 92.12$	$8333.25 \pm 39.12$	$-1.8\% \pm 1.53\%$
SB High	$8486.0 \pm 92.12$	$8378.14 \pm 38.45$	$-1.27\% \pm 1.52\%$
SB Low	$8486.0 \pm 92.12$	$8356.86 \pm 39.06$	$-1.52\% \pm 1.53\%$

**Table 8.3:** Agreement between data and prediction in 2bs tag CR. Showing stat uncertainty only.

## VALIDATION OF BACKGROUND ESTIMATION FROM LOW MASS AND HIGH MASS SIGNAL REGION

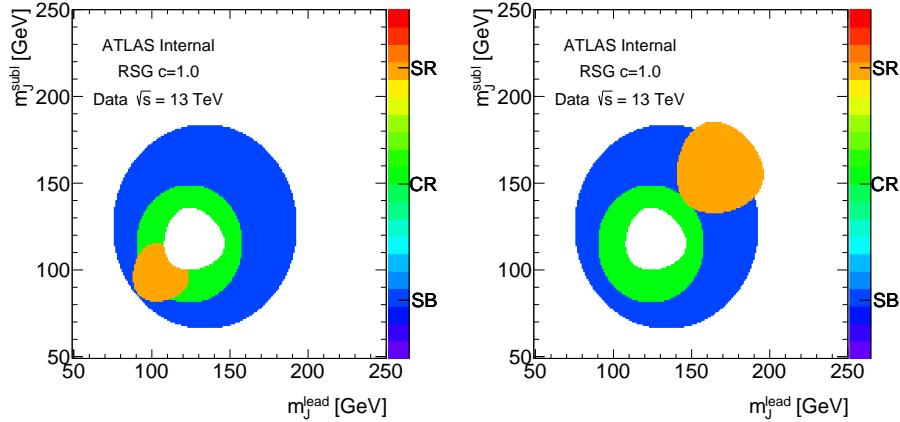
Another check is the so-called "low mass signal region rehearsal" (or ZZ region) and "high mass signal region rehearsal" (or TT region). Instead of a signal region around di-Higgs mass region on leading-subleading large-R jet mass 2D plane, we redefine a separate lower mass (ZZ) and higher mass (TT) signal region:

$$X_{ZZ} = \sqrt{\left(\frac{m(J_1) - 103 \text{ GeV}}{\text{o.}1m(J_1)}\right)^2 + \left(\frac{m(J_2) - 96 \text{ GeV}}{\text{o.}1m(J_2)}\right)^2} < 1.6 \quad (8.1)$$

$$X_{TT} = \sqrt{\left(\frac{m(J_1) - 164 \text{ GeV}}{\text{o.}1m(J_1)}\right)^2 + \left(\frac{m(J_2) - 155 \text{ GeV}}{\text{o.}1m(J_2)}\right)^2} < 1.6 \quad (8.2)$$

which is also illustrated in Figure 8.2. The analysis is repeated, using the same definition of Sideband and Control region as nominal (but with events contained in ZZ signal region excluded) for normalization fit. Then the low mass signal region is unblinded. This helps to validate the background estimation strategy, and the stability for other similar analysis.

The summary of background estimation for ZZ signal region can be found in Table 8.4, 8.5 and 8.6. The difference between data and prediction in ZZ signal region is summarized in Table 8.7 for all the regions. The discrepancy between data and prediction is either covered by statistical uncertainty of data or comparable with data statistical uncertainty in 4b, 3b and 2bs ZZ SR respectively. We further check the kinematic distribution between data and prediction in ZZ SR, as shown in Figure 8.3. The data agrees with prediction well in general, though a few bins might not agree perfectly. The difference from 4b CR region test is 17%, which is smaller than the 4b ZZ region difference. But the statistical uncertainty in ZZ region (yield 37) is much higher compared with our CR regions (min yield 76), hence the CR region with more statistical power is still used for the non-closure un-



**Figure 8.2:** Illustration of ZZ (left) and TT (right) signal region as shown in the orange shaded region. Control region shown in green, and Sideband region in blue. The white circle in the midde is the real Signal region, and it is blinded.

certainty.

The summary of background estimation for TT signal region can be found in Table 8.8, 8.9 and 8.10. The difference between data and prediction in TT signal region is summarized in Table 8.11 for all the regions. The discrepancy between data and prediction is either covered by statistical uncertainty of data or comparable with data statistical uncertainty in  $4b$ ,  $3b$  and  $2bs$  TT SR respectively. We further check the kinematic distribution between data and prediction in TT SR, as shown in Figure 8.4. The data agrees with prediction well in general, though a few bins might not agree perfectly.

Based on all the variation tests done above, we think there is no need to introduce extra uncertainty on non-closure systematics since most of the data/prediction disagreements are well covered by the data statistical uncertainty.

FourTag	Sideband	Control	Signal
QCD Est	$166.65 \pm 2.88$	$45.83 \pm 1.51$	$27.37 \pm 1.16$
$t\bar{t}$ Est.	$27.52 \pm 0.25$	$6.31 \pm 0.14$	$0 \pm 0$
$Z + jets$	$0 \pm 0$	$6.18 \pm 5.12$	$0 \pm 0$
Total Bkg Est	$194.17 \pm 2.89$	$58.32 \pm 5.34$	$27.37 \pm 1.16$
Data	$194.0 \pm 13.93$	$54.0 \pm 7.35$	$37.0 \pm 6.08$
$c = 1.0, m = 1.0 TeV$	$2.45 \pm 0.098$	$4.47 \pm 0.13$	$0.99 \pm 0.063$
$c = 1.0, m = 2.0 TeV$	$0.032 \pm 0.0015$	$0.075 \pm 0.0022$	$0.028 \pm 0.0014$
$c = 1.0, m = 3.0 TeV$	$0.00029 \pm 3.5e-05$	$0.00064 \pm 5e-05$	$0.0002 \pm 2.7e-05$

**Table 8.4:** Background prediction in SR/CR/SB for ZZ SR in  $4b$ -tag region. Uncertainties are stat only.

ThreeTag	Sideband	Control	Signal
QCD Est	$3344.46 \pm 26.85$	$998.41 \pm 14.63$	$637.78 \pm 11.87$
$t\bar{t}$ Est.	$826.66 \pm 25.11$	$136.58 \pm 10.23$	$30.07 \pm 1.24$
$Z + jets$	$32.49 \pm 11.34$	$8.22 \pm 5.29$	$3.3 \pm 2.0$
Total Bkg Est	$4203.61 \pm 38.47$	$1143.2 \pm 18.62$	$671.15 \pm 12.11$
Data	$4203.0 \pm 64.83$	$1108.0 \pm 33.29$	$645.0 \pm 25.4$
$c = 1.0, m = 1.0 TeV$	$7.56 \pm 0.18$	$9.84 \pm 0.2$	$3.05 \pm 0.11$
$c = 1.0, m = 2.0 TeV$	$0.15 \pm 0.0033$	$0.27 \pm 0.0046$	$0.12 \pm 0.003$
$c = 1.0, m = 3.0 TeV$	$0.0034 \pm 0.00012$	$0.0056 \pm 0.00016$	$0.0021 \pm 9.5e-05$

**Table 8.5:** Background prediction in SR/CR/SB for ZZ SR in  $3b$ -tag region. Uncertainties are stat only.

TwoTag split	Sideband	Control	Signal
QCD Est	$16387.44 \pm 37.6$	$4827.76 \pm 19.86$	$3026.83 \pm 15.61$
$t\bar{t}$ Est.	$7671.95 \pm 69.14$	$1229.96 \pm 26.54$	$332.29 \pm 13.66$
$Z + jets$	$44.37 \pm 13.23$	$13.34 \pm 6.6$	$36.47 \pm 12.88$
Total Bkg Est	$24103.77 \pm 79.8$	$6071.07 \pm 33.8$	$3395.59 \pm 24.42$
Data	$24104.0 \pm 155.25$	$6261.0 \pm 79.13$	$3258.0 \pm 57.08$
$c = 1.0, m = 1.0 TeV$	$4.57 \pm 0.14$	$4.65 \pm 0.14$	$1.91 \pm 0.089$
$c = 1.0, m = 2.0 TeV$	$0.16 \pm 0.0038$	$0.26 \pm 0.0047$	$0.12 \pm 0.0032$
$c = 1.0, m = 3.0 TeV$	$0.012 \pm 0.00024$	$0.019 \pm 0.00029$	$0.0085 \pm 0.00019$

**Table 8.6:** Background prediction in SR/CR/SB for ZZ SR in  $2bs$ -tag region. Uncertainties are stat only.

ZZ Signal Region	Data	Prediction	(Predict - Data)/Data
FourTag	$37.0 \pm 6.08$	$27.37 \pm 1.16$	$-26.0\% \pm 15.3\%$
ThreeTag	$645.0 \pm 25.4$	$671.15 \pm 12.11$	$4.05\% \pm 5.97\%$
TwoTag split	$3258.0 \pm 57.08$	$3395.59 \pm 24.42$	$4.22\% \pm 2.58\%$

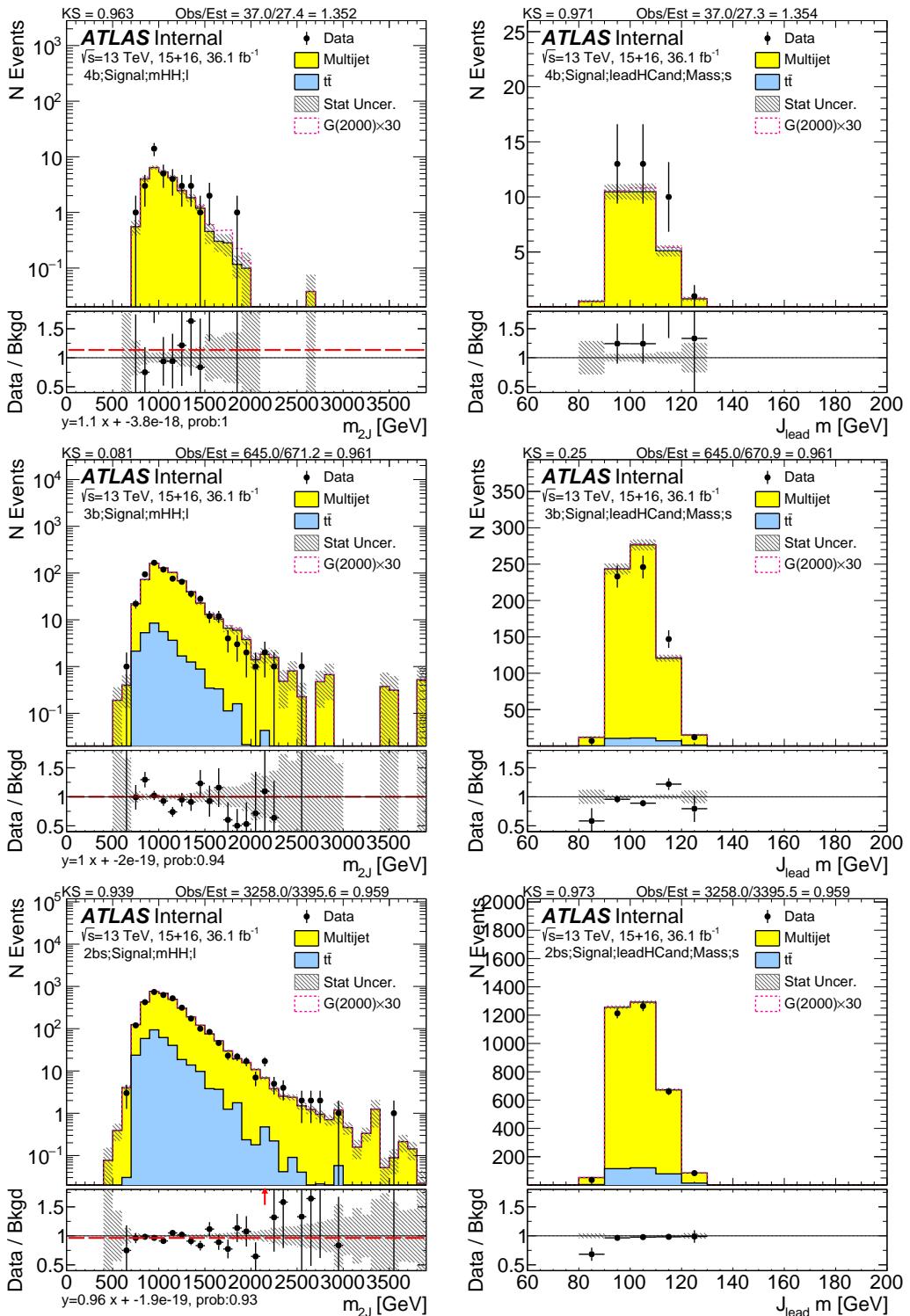
**Table 8.7:** Agreement between data and prediction in ZZ SR in  $4b$ ,  $3b$  and  $2bs$  regions.

FourTag	Sideband	Control	Signal
QCD Est	$152.28 \pm 2.72$	$63.47 \pm 1.77$	$28.6 \pm 1.21$
$t\bar{t}$ Est.	$19.86 \pm 0.22$	$7.45 \pm 0.15$	$15.02 \pm 0.2$
$Z + jets$	$0 \pm 0$	$6.18 \pm 5.12$	$0 \pm 0$
Total Bkg Est	$172.14 \pm 2.73$	$77.1 \pm 5.42$	$43.62 \pm 1.23$
Data	$172.0 \pm 13.11$	$81.0 \pm 9.0$	$46.0 \pm 6.78$
$c = 1.0, m = 1.0 TeV$	$2.38 \pm 0.097$	$5.4 \pm 0.15$	$0.15 \pm 0.024$
$c = 1.0, m = 2.0 TeV$	$0.033 \pm 0.0015$	$0.1 \pm 0.0026$	$0.0011 \pm 0.00027$
$c = 1.0, m = 3.0 TeV$	$0.00031 \pm 3.6e-05$	$0.0008 \pm 5.6e-05$	$1.5e-05 \pm 7.7e-06$

**Table 8.8:** Background prediction in SR/CR/SB for TT SR in  $4b$ -tag region. Uncertainties are stat only.

ThreeTag	Sideband	Control	Signal
QCD Est	$3106.11 \pm 25.79$	$1427.41 \pm 17.53$	$570.01 \pm 11.6$
$t\bar{t}$ Est.	$495.21 \pm 18.75$	$148.55 \pm 10.21$	$406.57 \pm 5.42$
$Z + jets$	$32.5 \pm 11.34$	$11.21 \pm 5.65$	$0.3 \pm 0.3$
Total Bkg Est	$3633.82 \pm 33.85$	$1587.17 \pm 21.05$	$976.88 \pm 12.81$
Data	$3633.0 \pm 60.27$	$1553.0 \pm 39.41$	$1017.0 \pm 31.89$
$c = 1.0, m = 1.0 TeV$	$7.57 \pm 0.18$	$12.58 \pm 0.23$	$0.32 \pm 0.037$
$c = 1.0, m = 2.0 TeV$	$0.15 \pm 0.0034$	$0.38 \pm 0.0054$	$0.0047 \pm 0.0006$
$c = 1.0, m = 3.0 TeV$	$0.0034 \pm 0.00012$	$0.0075 \pm 0.00018$	$0.00023 \pm 3.3e-05$

**Table 8.9:** Background prediction in SR/CR/SB for TT SR in  $3b$ -tag region. Uncertainties are stat only.



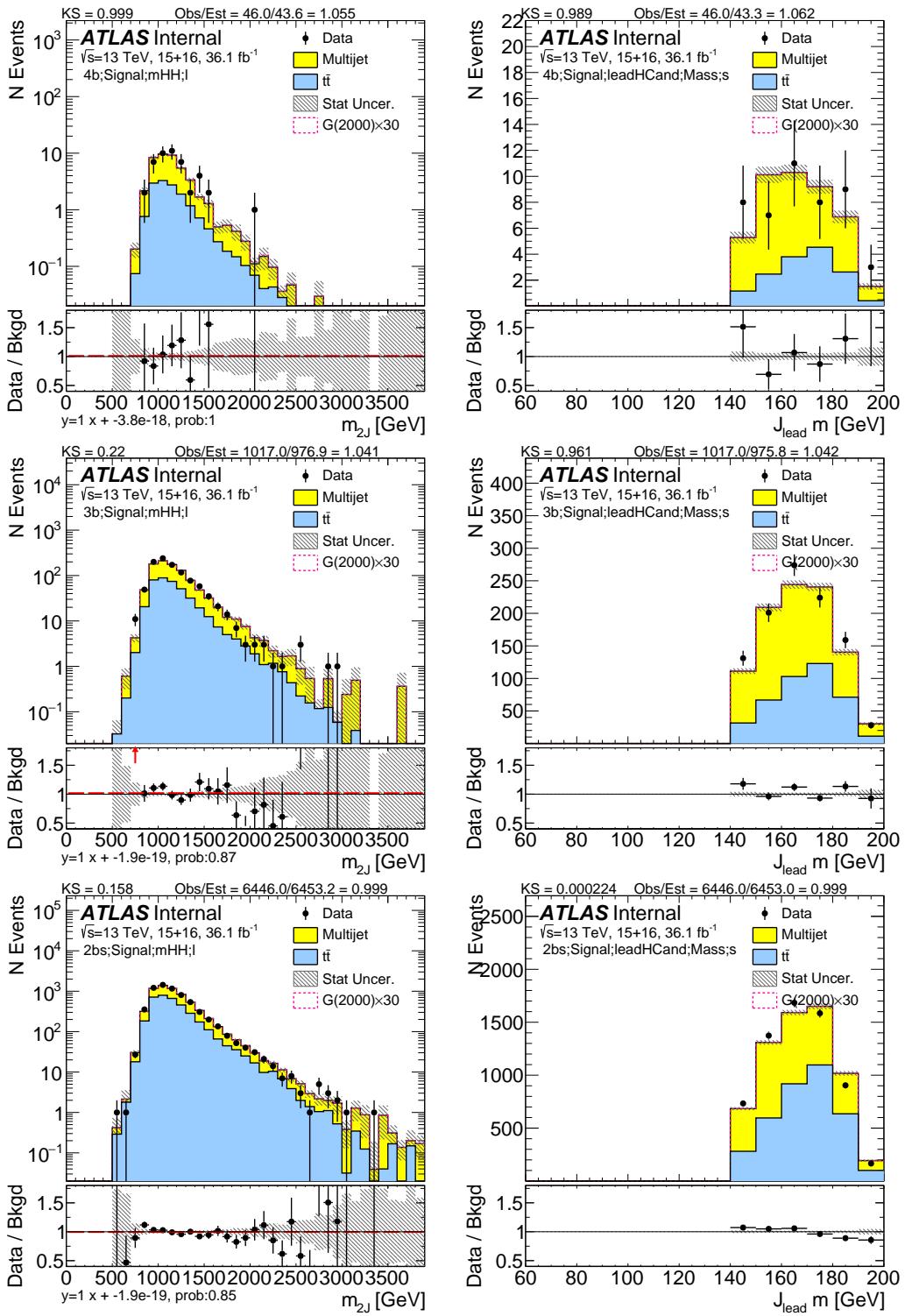
240  
**Figure 8.3:** ZZ signal region distribution of di-jet mass (left column) and leading large-R jet mass (right column) in low mass signal, for  $4b$  (top row),  $3b$  (middle row) and  $2b$  split (bottom row). The plots are with only statistical uncertainty.

TwoTag split	Sideband	Control	Signal
QCD Est	14980.05 ± 35.33	6803.06 ± 23.41	2817.92 ± 16.54
$t\bar{t}$ Est.	5170.92 ± 56.22	1468.85 ± 28.93	3628.91 ± 48.42
$Z + jets$	61.34 ± 16.04	26.44 ± 10.08	6.4 ± 5.05
Total Bkg Est	20212.31 ± 68.31	8298.34 ± 38.56	6453.23 ± 51.41
Data	20212.0 ± 142.17	8486.0 ± 92.12	6446.0 ± 80.29
$c = 1.0, m = 1.0 \text{ TeV}$	4.59 ± 0.14	6.33 ± 0.16	0.24 ± 0.033
$c = 1.0, m = 2.0 \text{ TeV}$	0.17 ± 0.0039	0.36 ± 0.0056	0.0066 ± 0.00077
$c = 1.0, m = 3.0 \text{ TeV}$	0.012 ± 0.00024	0.027 ± 0.00034	0.00089 ± 6.7e-05

**Table 8.10:** Background prediction in SR/CR/SB for TT SR in  $2bs$ -tag region. Uncertainties are stat only.

TT Signal Region	Data	Prediction	(Predict - Data)/Data
FourTag	46.0 ± 6.78	43.62 ± 1.23	-5.18 % ± 16.66 %
ThreeTag	1017.0 ± 31.89	976.88 ± 12.81	-3.95 % ± 4.27 %
TwoTag split	6446.0 ± 80.29	6453.23 ± 51.41	0.11 % ± 2.04 %

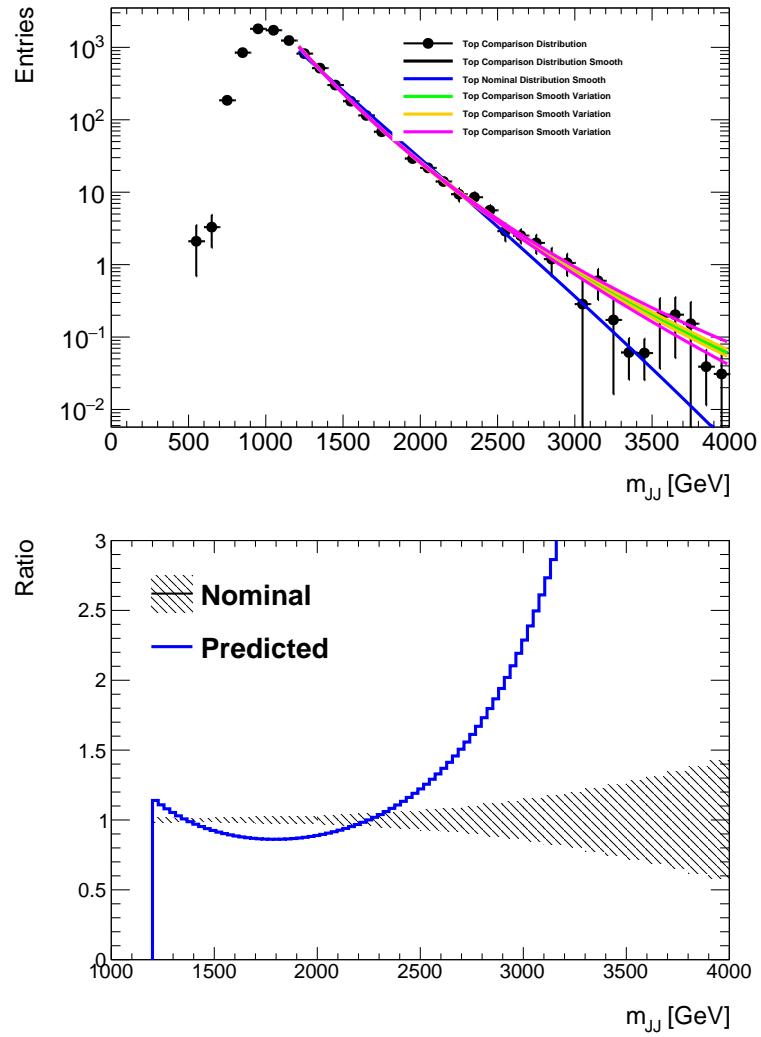
**Table 8.11:** Agreement between data and prediction in TT SR in  $4b$ ,  $3b$  and  $2bs$  regions.



242  
**Figure 8.4:** TT signal region distribution of di-jet mass (left column) and leading large-R jet mass (right column) in low mass signal, for 4 $b$  (top row), 3 $b$  (middle row) and 2 $b$  split (bottom row). The plots are with only statistical uncertainty.

## UNCERTAINTY ON THE SHAPE OF THE $t\bar{t}$ JET MASS IN THE $4/3b$ SIGNAL REGION

Because the  $4/3b$   $t\bar{t}$  MJJ distribution is extremely statistically limited, the  $4/3b$  shape is used to predict the final  $t\bar{t}$  background shape in the  $2bs$  signal region. In order to estimate the possible shape uncertainty, the  $2bs$  and  $3b$  sideband shapes are compared in Figure 8.5 (after being normalized to the same area). The  $2bs$  is used as there are not sufficient  $4b$  statistics to assess the comparison quality. In order to avoid large statistical uncertainties, the distributions of the  $3b$  and  $2b$  are smoothed. The ratio of the two smoothed distributions is taken as the shape systematic. We then use this function to apply a bin-by-bin scaling of the  $t\bar{t}$  background prediction in the signal region, maintaining the same normalization given by nominal  $t\bar{t}$  normalization prediction.

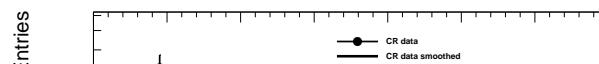
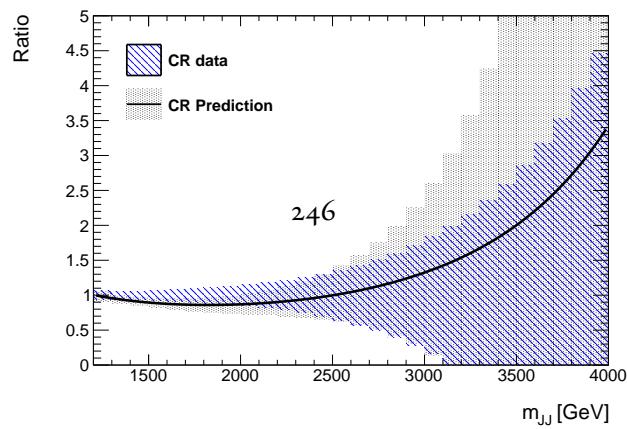
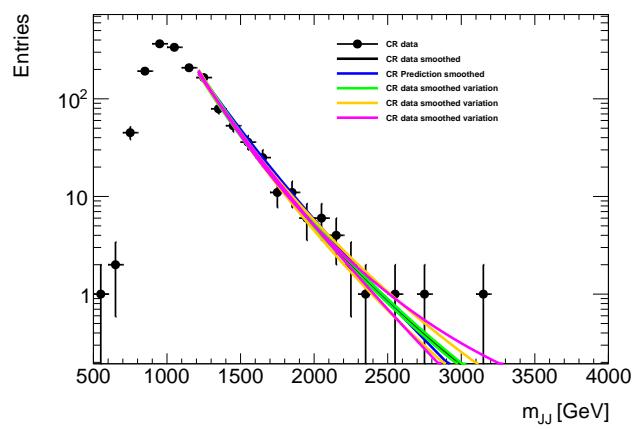
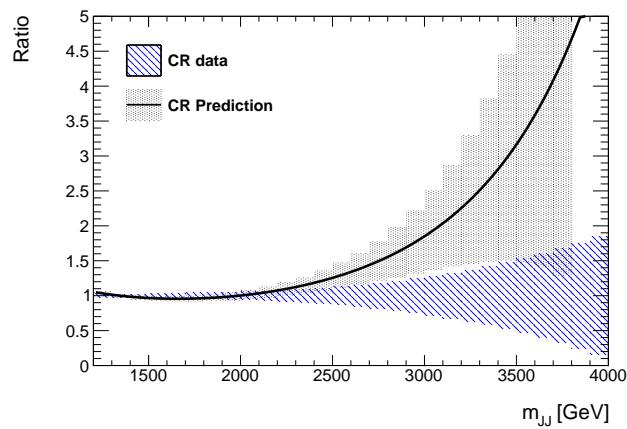
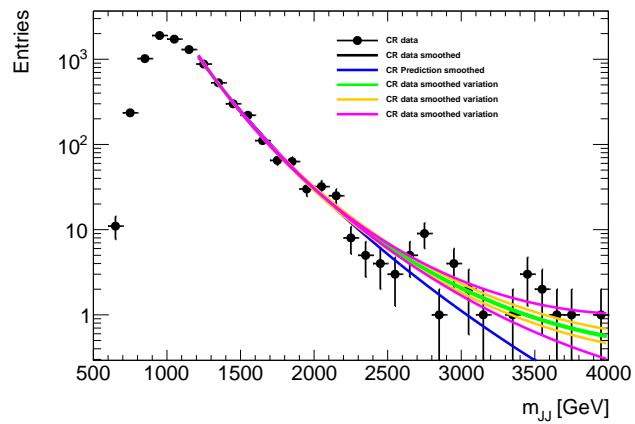


**Figure 8.5:** (left) Shape of the  $t\bar{t}$  di-large –  $R$ -jet mass in the sideband region, comparing the  $3b$  shape with that of the  $2b$ , in order to asses the systematic effect of additional  $b$ -tags changing the dijet mass distribution. The  $m_{JJ}$  distributions is shown on the left, and the ratio of  $3b$  to  $2b$  distributions on the right.

## UNCERTAINTY ON THE SHAPE OF THE $1/2b$ QCD DISTRIBUTION IN THE SIGNAL REGION

As shown in Figures 7.32 and 7.36, the shape distribution of the total predicted background using the scaled  $1/2b$  QCD sample was found to be in good agreement with the  $4b$ ,  $3b$ , and  $2bs$  data in the control region. However due to the low statistics in the data in the control region, the comparison is performed by first smoothing the  $1/2b$ , and the  $4/3/2bs$  distributions. The ratio of the smoothed  $1/2b$  distributions to that of the smoothed  $4/3/2bs$  distributions is taken as the shape systematic.

This function is then used to apply a bin-by-bin scaling of the QCD background prediction in the signal region, maintaining the same normalization given by  $\mu_{\text{multijet}}$ . The CR distributions and the smoothing fit ratios can be found in Figure 8.6. This systematics is further split into two parts: one below 2000 GeV and the other above 2000 GeV, to ensure the low and high mass shape variation post-fit pulls can vary independently. It should be noted, that this uncertainty is used for both the dijet mass, and the scaled dijet mass distribution, and the correction to scaling is expected to be small relative to the dijet mass.



## UNCERTAINTY ON QCD SMOOTHING FUNCTION IN THE SIGNAL REGION

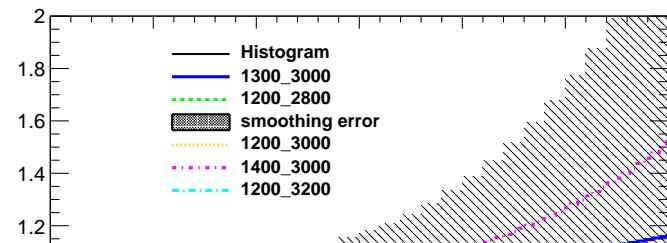
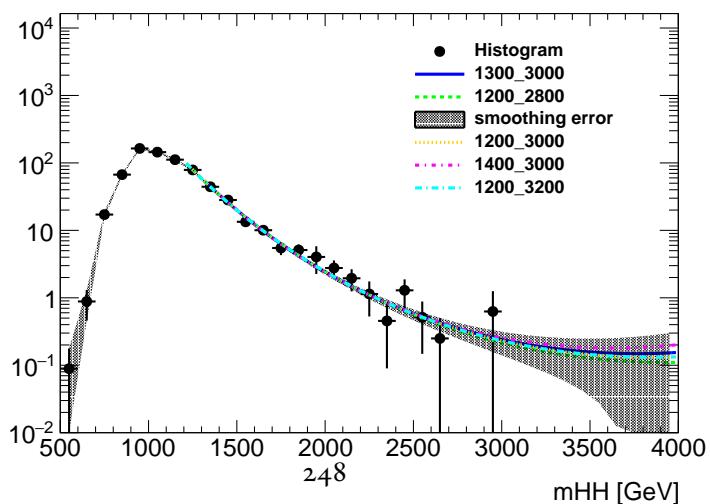
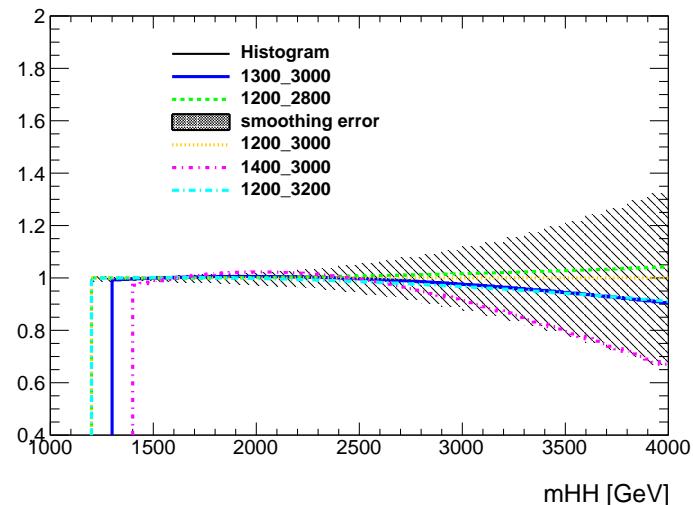
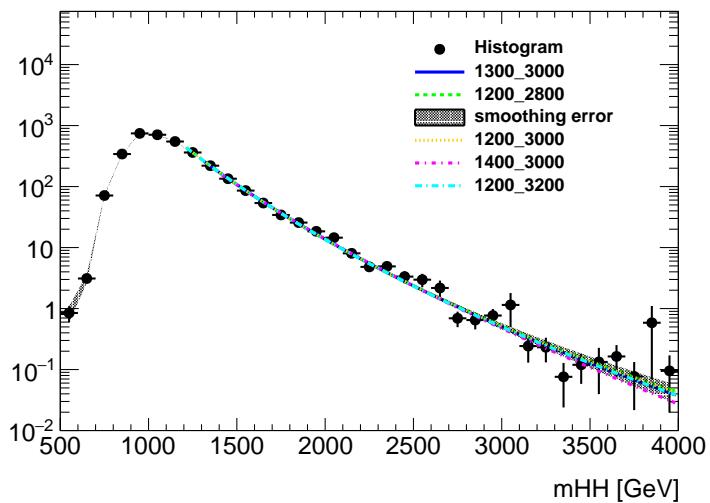
The MJ8 function has been used to fit the QCD background prediction in order to smooth the distribution and provide non-zero background estimates up to dijet masses beyond which we have  $1/2b$  statistics. While this distribution is observed to fit the  $1/2b$  data well, it does not have a concrete physical motivation, and in principle the high mass tail of the distribution could be larger than predicted by an exponential. Two checks are performed, changing the boundaries where the fit is performed, and changing the fit function.

To test the impact of the region in which the fit is performed, we varied the upper bound on the dijet fit region to be each of the values  $\{2800, 3000, 3200\}$  GeV and the starting value between  $\{1200, 1300, 1400\}$  GeV. The ratio of the fits for each upper bound, to that of the nominal (1200-3000 GeV) can be found in Figure 8.7, along with a hash band showing the statistical uncertainty of the nominal fit. The maximum deviation from the nominal fit, per bin, is taken as the shape systematic uncertainty. This is estimated separately for  $2b$ ,  $3b$ , and  $4b$  samples.

It should be noted that fits in which the fit  $\chi^2$  probability was less than 0.001%, or in which the fit integrals between 1500-2000 GeV, 2000-2500 GeV, or  $>2500$  GeV were not in agreement with the original  $1/2b$  distribution within a factor of 2 or 0.5, were not used to estimate the uncertainty. The aforementioned checks ensure that we do not use poor fits of the  $1/2b$  distribution to estimate the uncertainty.

As a second test, we fit the  $1/2b$  QCD prediction with a variety of other distributions which can show both power law behavior in the bulk of the distribution as well as longer tails. The set of additional functions examined (labelled MJ1-MJ7) can be found in Table 8.12, where  $x = m_{JJ}/\sqrt{s}$ .

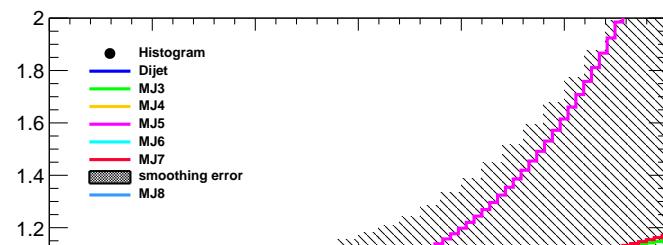
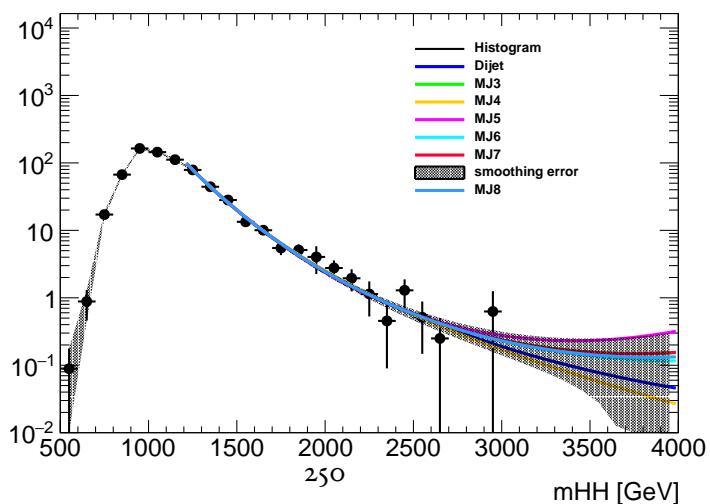
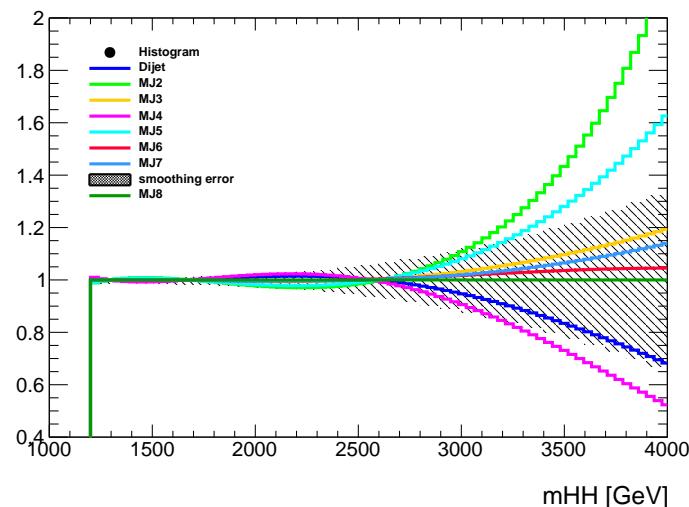
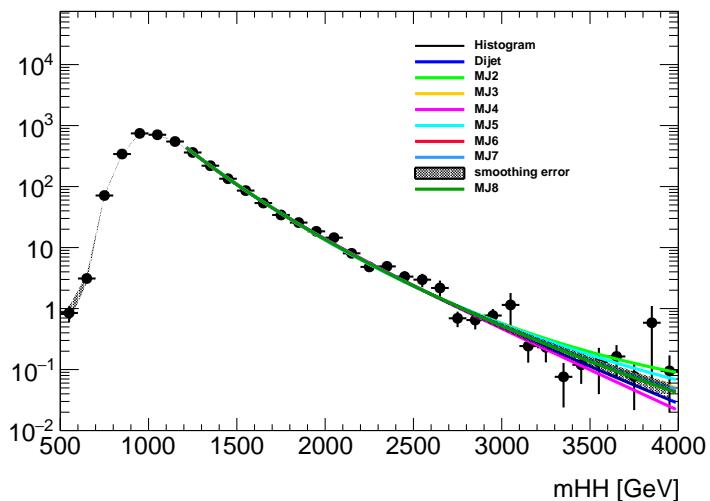
Figure 8.8 shows the fits to the QCD prediction in the  $4/3/2b$  signal regions, and the nominal dijet fit, as well as the ratios of the nominal fit to that of the additional functions. The maximum per bin deviation is taken as the shape systematic, separately for the  $4/3/2b$  SRs.



Name	Functional Form
MJ1 (Dijet)	$f_1(x) = p_o(1-x)^{p_1}x^{p_2}$
MJ2	$f_2(x) = p_o(1-x)^{p_1}e^{p_2 x^2}$
MJ3	$f_3(x) = p_o(1-x)^{p_1}x^{p_2 x}$
MJ4	$f_4(x) = p_o(1-x)^{p_1}x^{p_2 \ln x}$
MJ5	$f_5(x) = p_o(1-x)^{p_1}(1+x)^{p_2 x}$
MJ6	$f_6(x) = p_o(1-x)^{p_1}(1+x)^{p_2 \ln x}$
MJ7	$f_7(x) = \frac{p_o}{x}(1-x)^{p_1-p_2 \ln x}$
MJ8	$f_8(x) = \frac{p_o}{x^2}(1-x)^{p_1-p_2 \ln x}$

**Table 8.12:** Functions used to fit the QCD dijet mass distributions, where  $x = m_{jj}/\sqrt{s}$ .

As before, fits in which the fit  $\chi^2$  probability was less than 0.1%, or in which the fit integrals between 1500-2000 GeV, 2000-2500 GeV, or  $>2500$  GeV were not in agreement with the original ob distribution within a factor of 2 or 0.5, were not used to estimate the uncertainty. The aforementioned checks ensure that we do not use poor fits of the ob distribution to estimate the uncertainty.



#### 8.0.4 SUMMARY OF SYSTEMATICS

Table 8.13 shows the percent impact of systematics used in this analysis on the backgrounds yields and on the expected yields for RSG  $c = 1.0$  signals in the  $4b$  signal region. The correspondent values are shown for the  $3b$  signal region in Table 8.14, and are shown for the  $2bs$  signal region in Table 8.15.

A 3.2% luminosity uncertainty is also considered for the Z+jets background and the RSG signal predictions. The JER/JMR/JES/JMS/track jet b-tag scale factor uncertainties are applied to RSG and  $t\bar{t}$  samples.

The “ $t\bar{t}$  SB shape” is the background normalization uncertainty due to the mis-modeling of the shape of the  $t\bar{t}$  dijet mass distribution in the sideband region, taken from comparing the shape with  $3/4$   $b$ -tags to the shape with only a  $2bs$  requirement. The “Background Normalization Fit” uncertainty comes from summing in quadrature the independent uncertainty components calculated from the correlated statistical errors of  $\mu_{\text{multijet}}$  and  $\alpha_{t\bar{t}}$ . The “QCD Non-Closure in CR” systematic is derived as the maximum of (a) difference between the predicted and observed  $4b/3b/2bs$  QCD yields in the control region, (b) the fractional change in SR predictions from varying the CR and SB definitions. Both options gave similar sized uncertainties, but the uncertainty from the CR/SB variations was found to be larger. All these uncertainties are summed in quadrature and shown in the “Bkg Est” row in the table.

The remaining systematics not listed in this table, as they do not impact the acceptance, include uncertainties on the shape of the QCD and  $t\bar{t}$  backgrounds in the signal region, and the uncertainties from the smoothing / extrapolation procedure.

The size of the Monte Carlo modeling systematics on the RSG  $c=1.0$  signal yield as a function of the signal mass can be found in Figure 8.9. These uncertainties have a similar impact on the other signals. The largest uncertainty in the  $4b$  and  $2b$  signal region is from  $b$ -tagging, followed by the

JMR uncertainty. In the  $3b$  signal region, although  $b$ -tagging systematics is still one of the largest uncertainty, it has been much reduced compared to  $4b$  region, as discussed in Section 8.0.1. Then the jet mass scale and resolution are the largest uncertainties following  $b$ -tagging.

FourTag	totalbkg	qcd	ttbar	RSG <sub>1</sub> 1000	RSG <sub>1</sub> 2000	RSG <sub>1</sub> 3000
JER	0.45	0.27	3.98	2.44	1.07	0.67
JMR	7.9	10.35	39.95	12.33	13.16	15.08
Top	-	-	-	-	-	-
JES/JMS	1.32	1.49	24.36	5.18	3.72	5.62
Bkg Est	15.67	18.19	67.82	-	-	-
b-tag SF	1.11	0.79	18.85	18.34	28.11	27.73
Total Sys	17.64	21.0	84.62	22.83	31.28	32.07
Stat	3.13	3.29	2.47	1.97	1.63	4.9
Estimated Events	34.59	32.91	1.68	10.07	0.25	0.0016

**Table 8.13:** Percent impact of the dominant systematics on the background acceptance and on the signal acceptance of RS  $c = 1.0$  graviton predictions in the  $4b$  signal region.

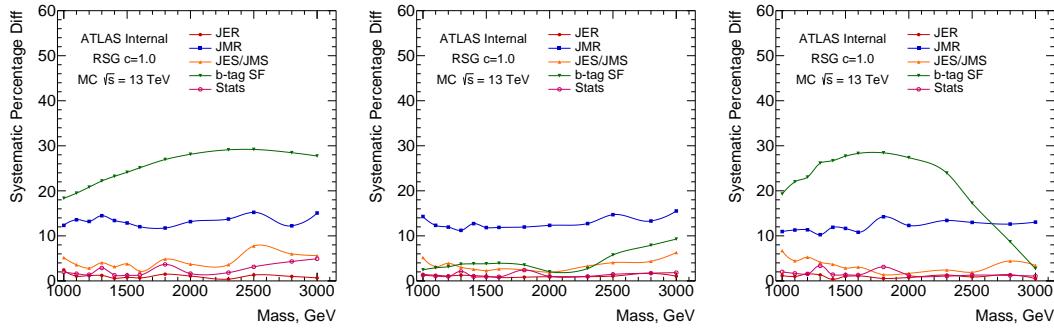
ThreeTag	totalbkg	qcd	ttbar	RSG <sub>1</sub> 1000	RSG <sub>1</sub> 2000	RSG <sub>1</sub> 3000
JER	1.38	3.52	17.5	1.41	0.93	1.08
JMR	1.35	4.26	24.38	14.3	12.33	15.53
Top	-	-	-	-	-	-
JES/JMS	2.03	1.26	26.22	5.19	1.94	6.35
Bkg Est	4.84	5.62	9.45	-	-	-
b-tag SF	0.47	0.53	8.45	2.45	2.01	9.27
Total Sys	5.61	8.0	41.82	15.47	12.68	19.2
Stat	1.32	1.44	2.47	1.26	1.0	1.83
Estimated Events	780.89	701.52	79.38	26.0	0.76	0.013

**Table 8.14:** Percent impact of the dominant systematics on the background acceptance and on the signal acceptance of RS  $c = 1.0$  graviton predictions in the  $3b$  signal region.

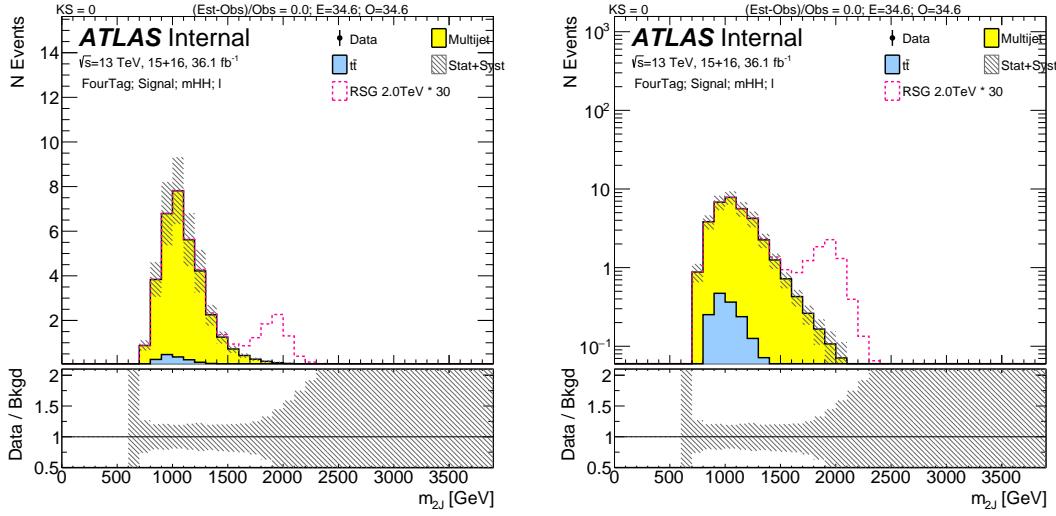
The final background prediction of MJJ along with total systematic uncertainties can be found in Figure 8.10, 8.11, and 8.12. The final background prediction of scaled MJJ along with total uncertainties can be found in Figure 8.13, 8.14, and 8.15.

TwoTag split	totalbkg	qcd	ttbar	RSG <sub>1</sub> 1000	RSG <sub>1</sub> 2000	RSG <sub>1</sub> 3000
JER	0.25	0.48	3.14	1.18	0.74	0.5
JMR	0.52	1.73	9.43	10.96	12.3	13.03
Top	-	-	-	-	-	-
JES/JMS	0.43	1.67	7.17	6.72	1.7	3.55
Bkg Est	2.7	3.32	2.4	-	-	-
b-tag SF	0.83	1.43	1.82	19.28	27.36	2.72
Total Sys	2.92	4.37	12.62	23.2	30.05	13.79
Stat	0.6	0.41	2.47	2.0	1.2	1.07
Estimated Events	4251.49	3392.79	858.7	10.87	0.6	0.039

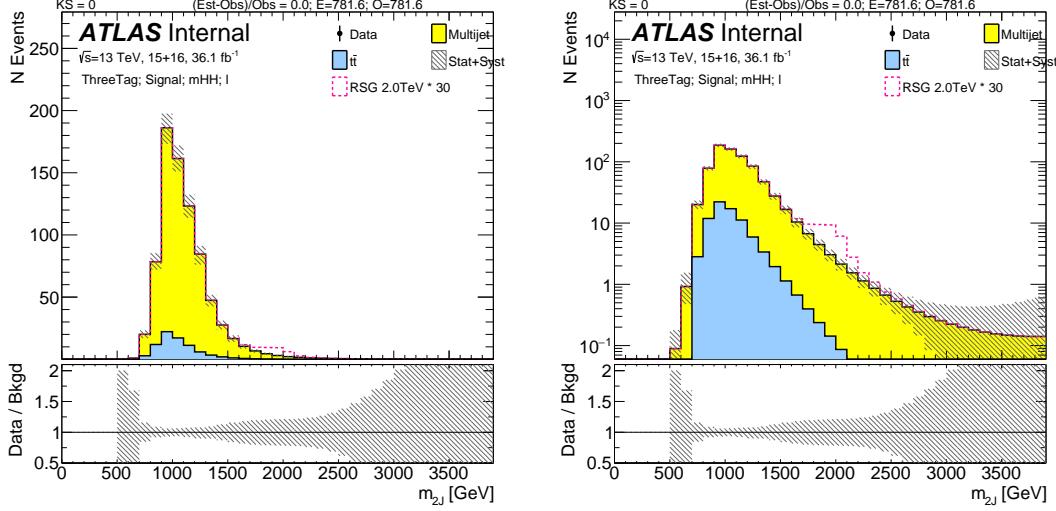
**Table 8.15:** Percent impact of the dominant systematics on the background acceptance and on the signal acceptance of RS  $c = 1.0$  graviton predictions in the  $2bs$  signal region.



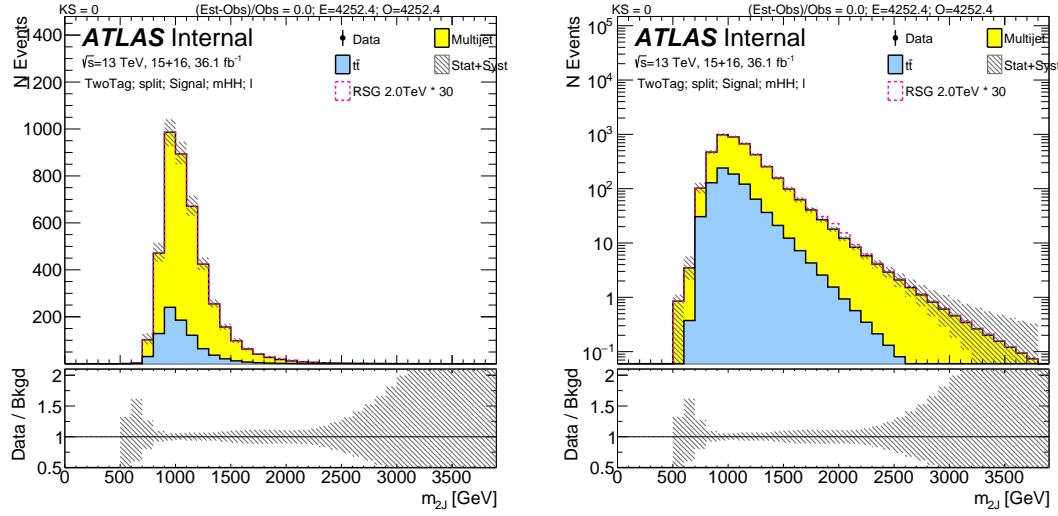
**Figure 8.9:** Impact of each systematic on the signal prediction as a function of the signal mass, in the  $4b$  (left) and  $3b$  (middle) and  $2bs$  signal regions.



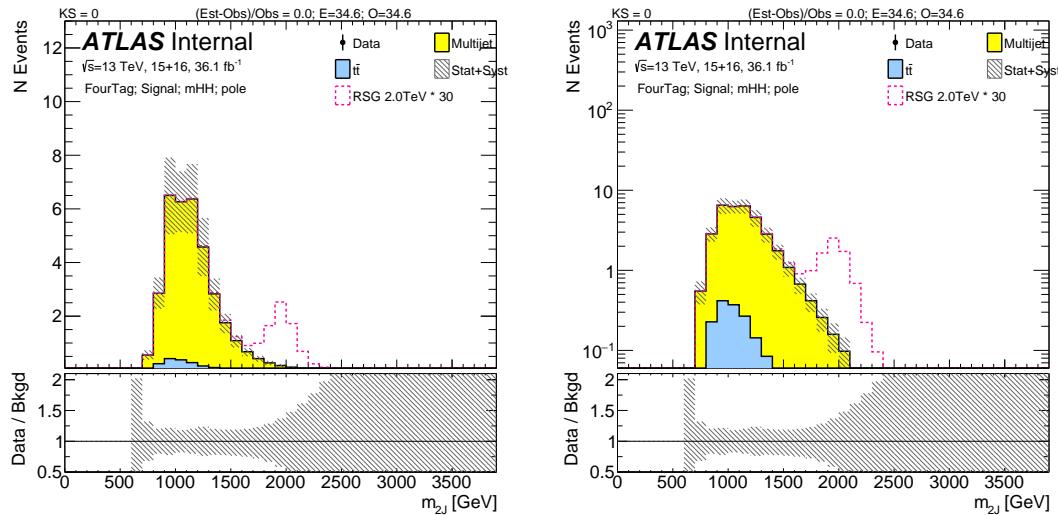
**Figure 8.10:** The total background estimation in  $4b$  signal region, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down.



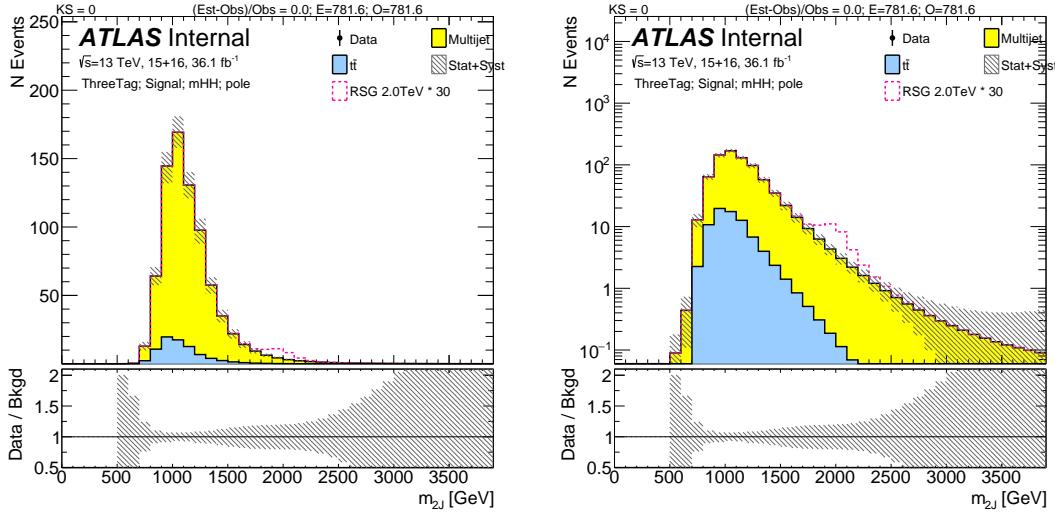
**Figure 8.11:** The total background estimation in  $3b$  signal region, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down.



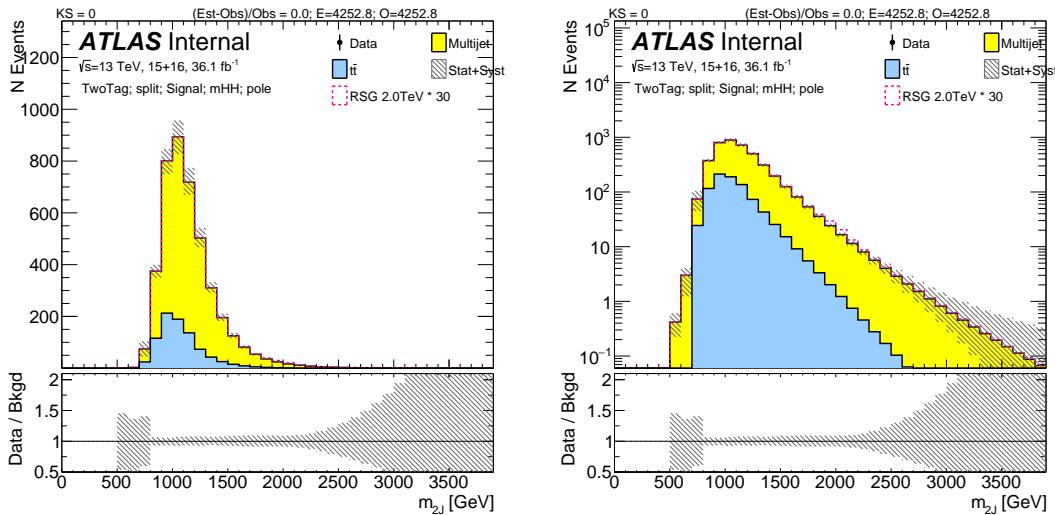
**Figure 8.12:** The total background estimation in  $2b$  signal region, with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down.



**Figure 8.13:** The total background estimation in  $4b$  signal region, scaled  $m_{JJ}$ , with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down.



**Figure 8.14:** The total background estimation in  $3b$  signal region, scaled  $m_{JJ}$ , with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down.



**Figure 8.15:** The total background estimation in  $2bs$  signal region, scaled  $m_{JJ}$ , with linear scale on the left and with log scale on the right, along with total uncertainties (stats.+systematic) variation up and down.

*“You must be ready to give up even the most attractive ideas when experiment shows them to be wrong.”*

Alessandro Volta

# 9

## Result

### 9.0.1 UNBLINDED RESULTS

The unblinded results are summarised in Table 9.1.

For reader’s interest, we integrate the background prediction from a certain mass point on and compare that with our unblinded observations. These are listed in Table 9.2, 9.3, 9.4. The unscaled  $2bs/3b/4bs$  dijet mass distributions are shown in Figures 9.3, 9.2, 9.1. No significant excess of number of events or in the dijet mass distribution is observed.

For the scaled dijet mass, the integral values are listed in Table 9.5, 9.6, 9.7. The scaled  $2bs/3b/4bs$  dijet mass distributions are shown in Figures 9.6, 9.5, 9.4. No significant excess of number of events or in the dijet mass distribution is observed as well.

Sample	FourTag	ThreeTag	TwoTag split
qcd	$32.92 \pm 7.07$	$702.16 \pm 63.12$	$3393.81 \pm 148.78$
ttbar	$1.68 \pm 1.43$	$79.41 \pm 33.12$	$859.03 \pm 107.86$
totalbkg	$34.6 \pm 6.28$	$781.56 \pm 52.42$	$4252.83 \pm 125.73$
Data	$31.0 \pm 5.57$	$801.0 \pm 28.3$	$4376.0 \pm 66.15$

**Table 9.1:** Unblinded Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown.

Mass Range	$>1000$	$>1500$	$>2000$	$>2500$	$>3000$
totalbkg	$23.09 \pm 1.59$	$1.94 \pm 0.15$	$0.26 \pm 0.072$	$0.061 \pm 0.058$	$0.021 \pm 0.047$
data	$21.0 \pm 4.58$	$3.0 \pm 1.73$	-	-	-

**Table 9.2:**  $4b$  unblinded Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals.

Mass Range	$>1000$	$>1500$	$>2000$	$>2500$	$>3000$
totalbkg	$495.92 \pm 12.34$	$51.72 \pm 2.46$	$10.42 \pm 0.95$	$4.07 \pm 0.85$	$2.21 \pm 0.79$
data	$499.0 \pm 22.34$	$42.0 \pm 6.48$	$3.0 \pm 1.73$	$1.0 \pm 1.0$	-

**Table 9.3:**  $3b$  unblinded Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals.

Mass Range	$>1000$	$>1500$	$>2000$	$>2500$	$>3000$
totalbkg	$2688.71 \pm 34.09$	$288.51 \pm 4.96$	$42.19 \pm 2.13$	$8.85 \pm 1.55$	$2.72 \pm 1.09$
data	$2755.0 \pm 52.49$	$287.0 \pm 16.94$	$38.0 \pm 6.16$	$4.0 \pm 2.0$	$1.0 \pm 1.0$

**Table 9.4:**  $2bs$  unblinded Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals.

Mass Range	$>1000$	$>1500$	$>2000$	$>2500$	$>3000$
totalbkg	$24.64 \pm 1.84$	$2.84 \pm 0.22$	$0.25 \pm 0.044$	$0.02 \pm 0.011$	$0.0014 \pm 0.0026$
data	$22.0 \pm 4.69$	$4.0 \pm 2.0$	$1.0 \pm 1.0$	-	-

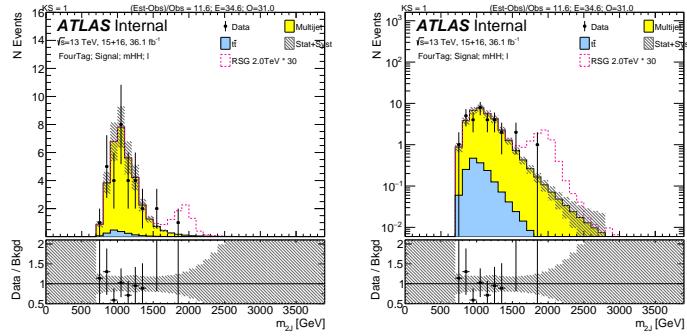
**Table 9.5:**  $4b$  unblinded Scaled dijet mass Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals.

Mass Range	>1000	>1500	>2000	>2500	>3000
totalbkg	559.38 ± 14.06	69.27 ± 3.22	13.35 ± 1.14	4.37 ± 0.96	2.0 ± 0.87
data	570.0 ± 23.87	59.0 ± 7.68	4.0 ± 2.0	1.0 ± 1.0	-

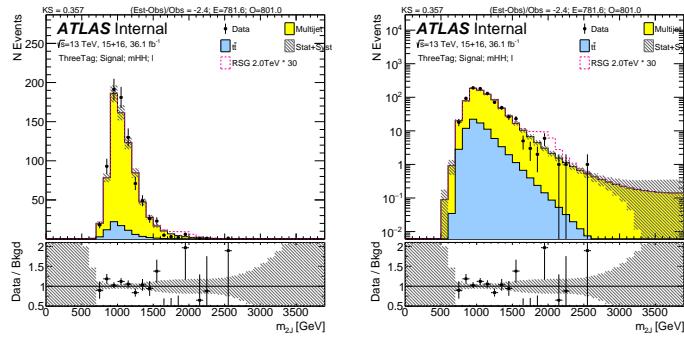
**Table 9.6:**  $3b$  unblinded Scaled dijet mass Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals.

Mass Range	>1000	>1500	>2000	>2500	>3000
totalbkg	2998.69 ± 40.31	377.8 ± 6.39	57.47 ± 2.88	11.78 ± 1.95	3.39 ± 1.27
data	3078.0 ± 55.48	379.0 ± 19.47	47.0 ± 6.86	6.0 ± 2.45	2.0 ± 1.41

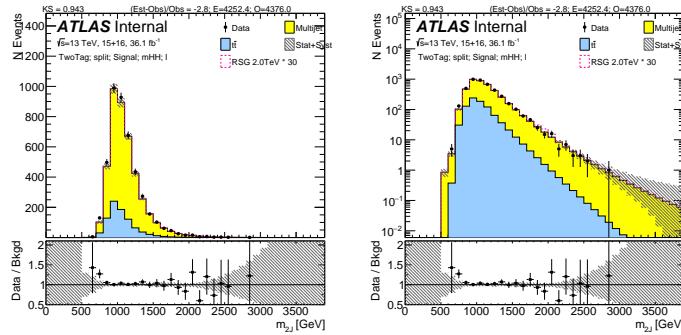
**Table 9.7:**  $2bs$  unblinded Scaled dijet mass Signal Region predictions and results. All systematic uncertainties included for backgrounds. For Data, the statistical uncertainty is shown. Mass range is broken into greater than 1 TeV, 1.5 TeV, 2 TeV, 2.5 TeV, and 3 TeV intervals.



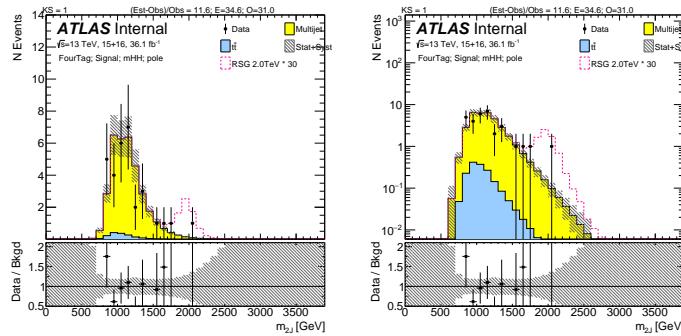
**Figure 9.1:** Unscaled dijet mass distribution in the  $4b$  Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic uncertainty are shown on the plot.



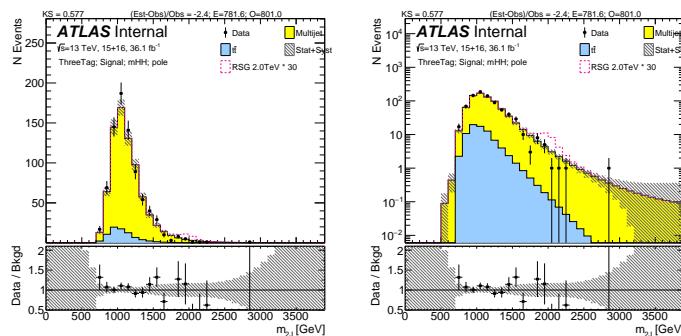
**Figure 9.2:** Unscaled dijet mass distribution in the  $3b$  Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic uncertainty are shown on the plot.



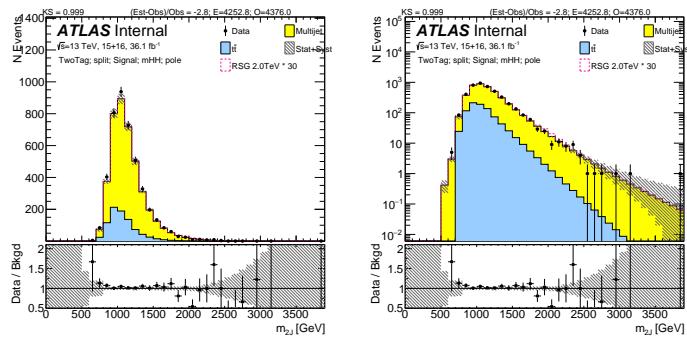
**Figure 9.3:** Unscaled dijet mass distribution in the  $2b$  Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucnertainties are shown on the plot.



**Figure 9.4:** Scaled dijet mass distribution in the  $4b$  Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucnertainties are shown on the plot.



**Figure 9.5:** Scaled dijet mass distribution in the  $3b$  Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucnertainties are shown on the plot.



**Figure 9.6:** Scaled dijet mass distribution in the  $2\bar{b}s$  Signal Region after unblinding. The left plot is on linear scale and the right plot is on log scale. Stat uncertainty and systematic ucnertainties are shown on the plot.

### 9.0.2 KINEMATIC DISTRIBUTIONS

This section shows unblinded comparisons of data with the prediction of QCD multi-jets and  $t\bar{t}$  in the signal region (SR). Plots shown are with stat uncertainty only.

Figures 9.7, 9.8, 9.9, and 9.10 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $4b$  selection.

Figures 9.11, 9.12, 9.13, and 9.14 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $3b$  selection.

Figures 9.15, 9.16, 9.17, and 9.18 show predictions of various kinematics of the large- $R$  jets and their associated track jets in the  $2b$  selection.

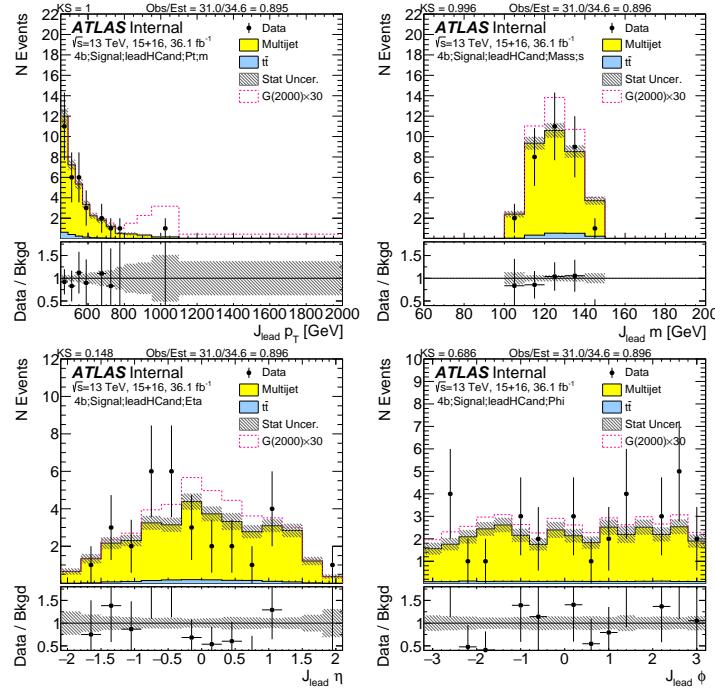


Figure 9.7: Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring  $4b$ -tags.

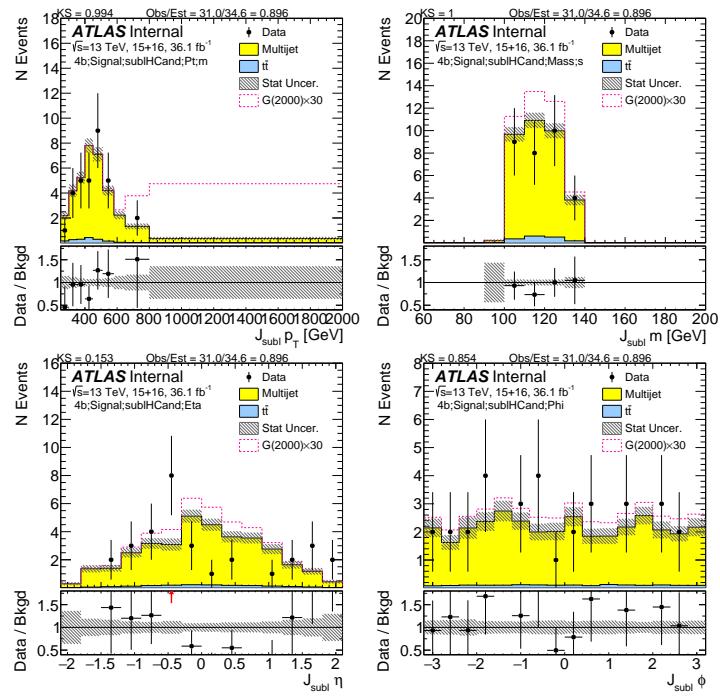
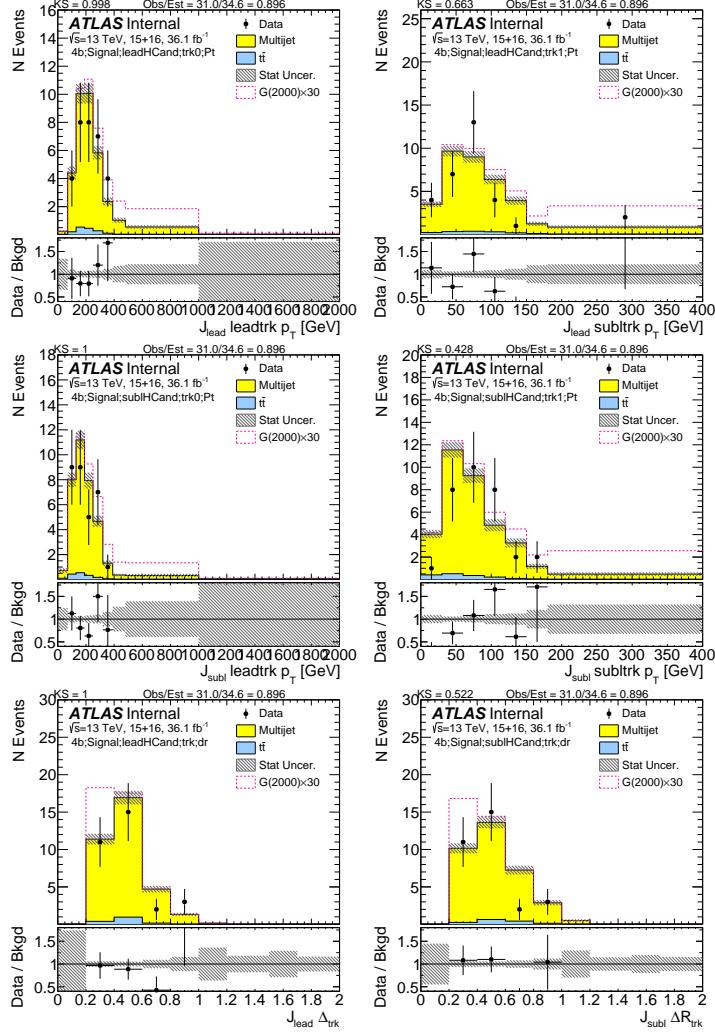


Figure 9.8: Kinematics of the sub-lead large- $R$  jet in data and prediction in the signal region after requiring 4  $b$ -tags.



**Figure 9.9:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 4  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.

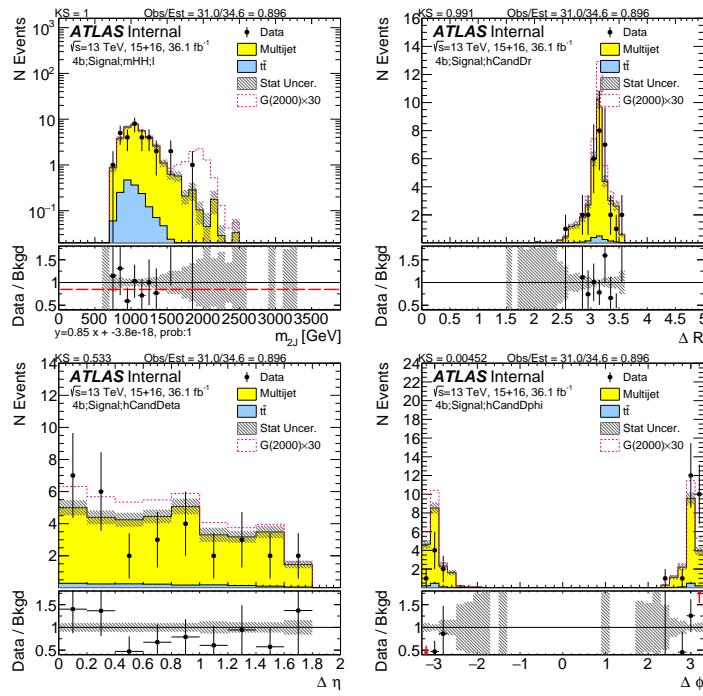


Figure 9.10: Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 4  $b$ -tags.

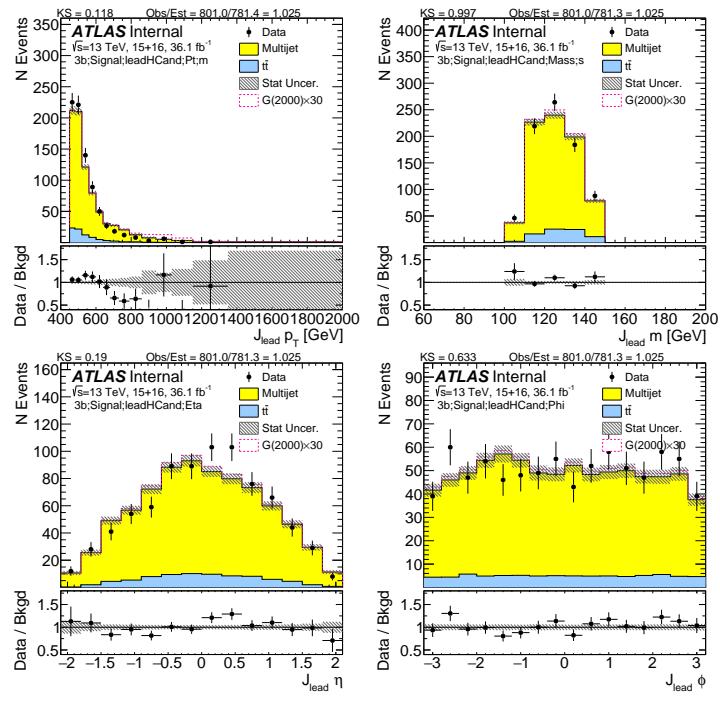
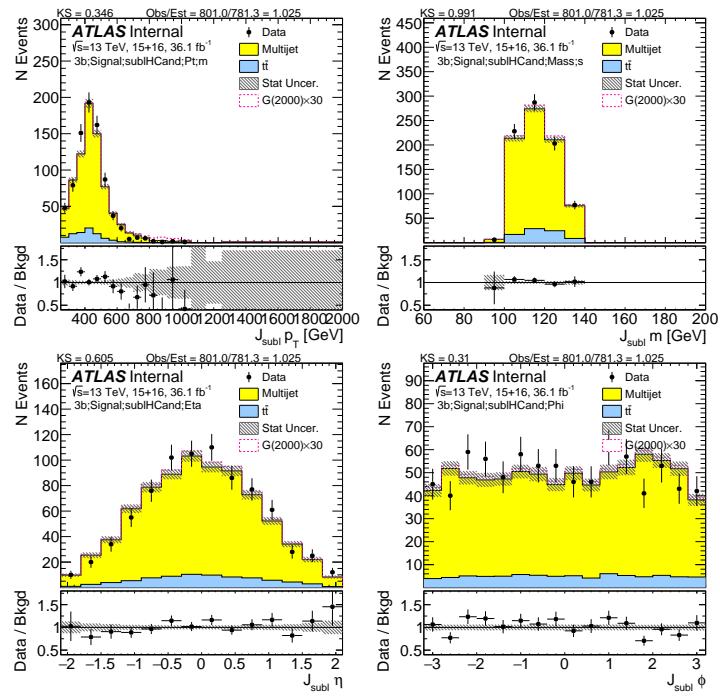
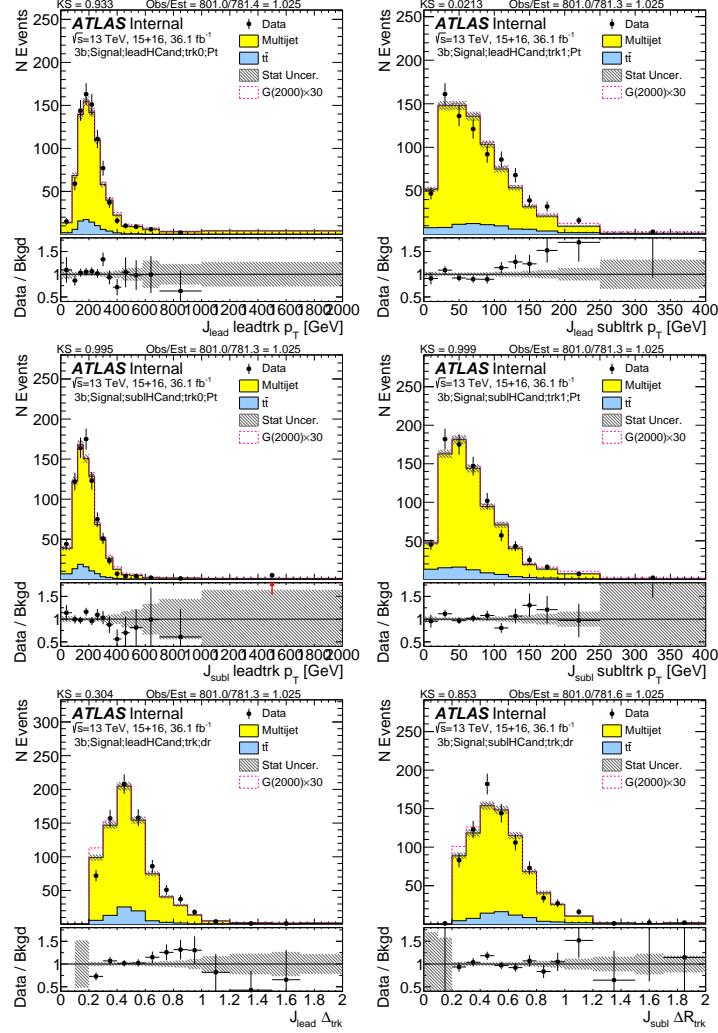


Figure 9.11: Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags.



**Figure 9.12:** Kinematics of the sub-leading large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags.



**Figure 9.13:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 3  $b$ -tags. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.

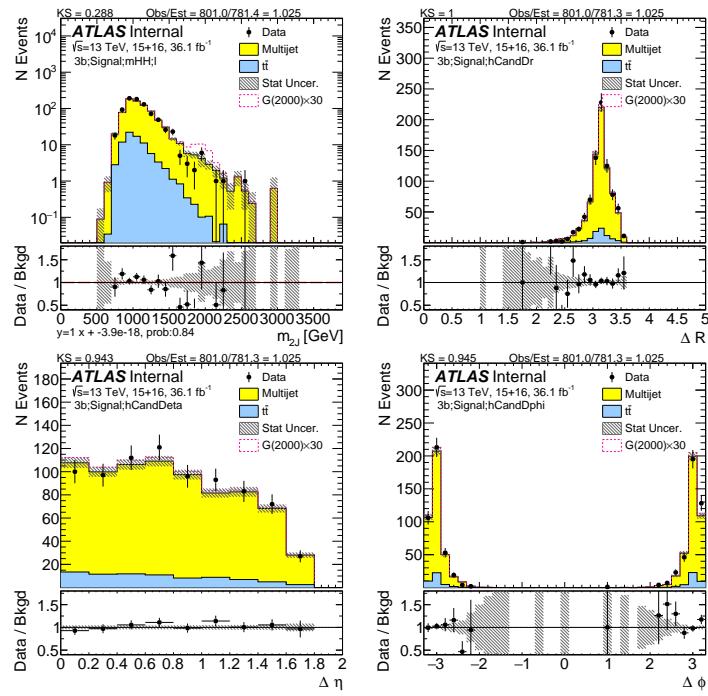
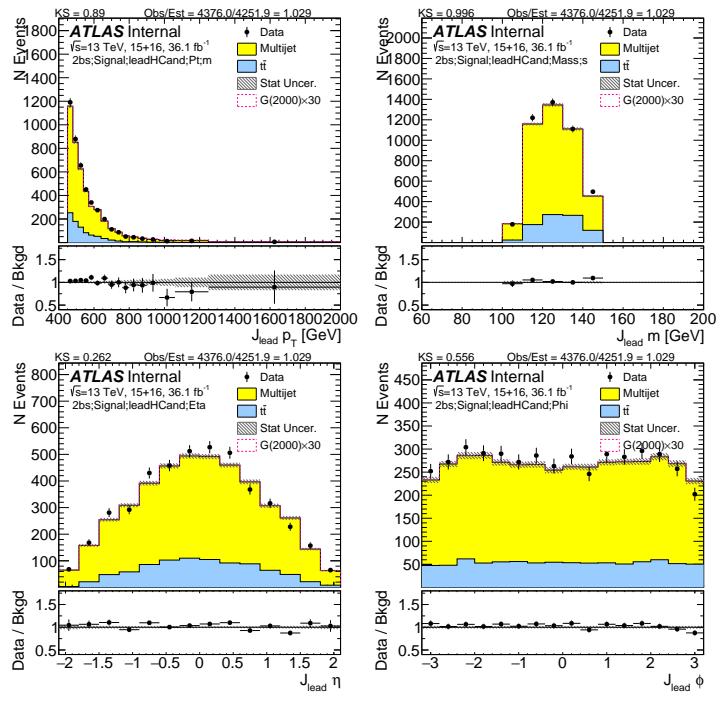
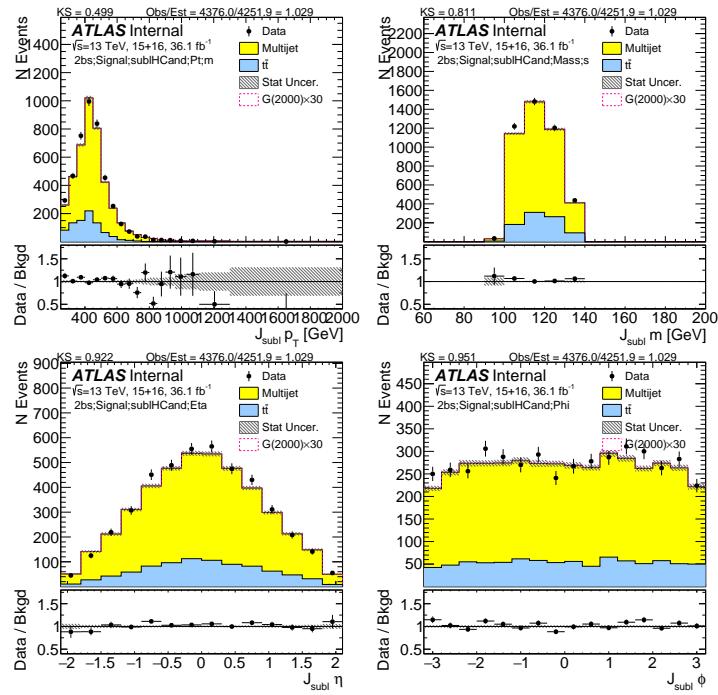


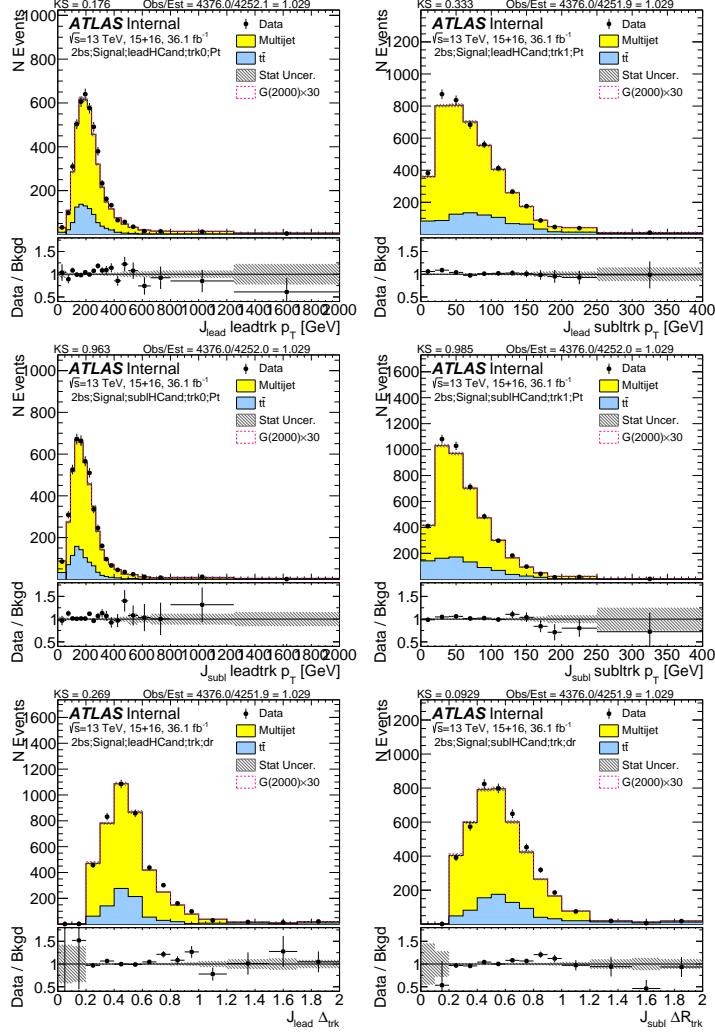
Figure 9.14: Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 3  $b$ -tags.



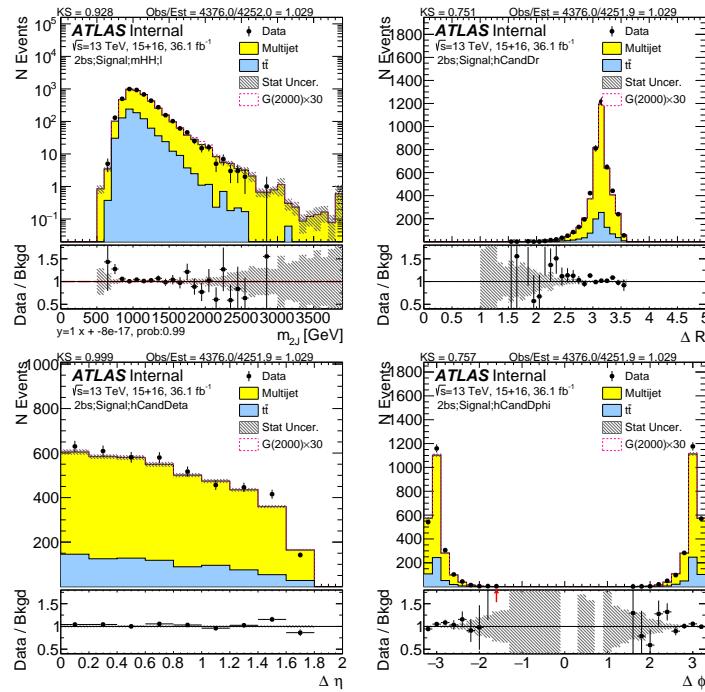
**Figure 9.15:** Kinematics of the lead large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split.



**Figure 9.16:** Kinematics of the sub-lead large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split.



**Figure 9.17:** First two rows show the kinematics of the lead (left) and sub-lead (right) small- $R$  track jets associated to the lead (first-row) and sub-lead (second-row) large- $R$  jet in data and prediction in the signal region after requiring 2  $b$ -tags split. Third row shows the  $\Delta R$  between two leading small- $R$  track-jets associated to the leading (left) and sub-leading (right) large- $R$  jet.



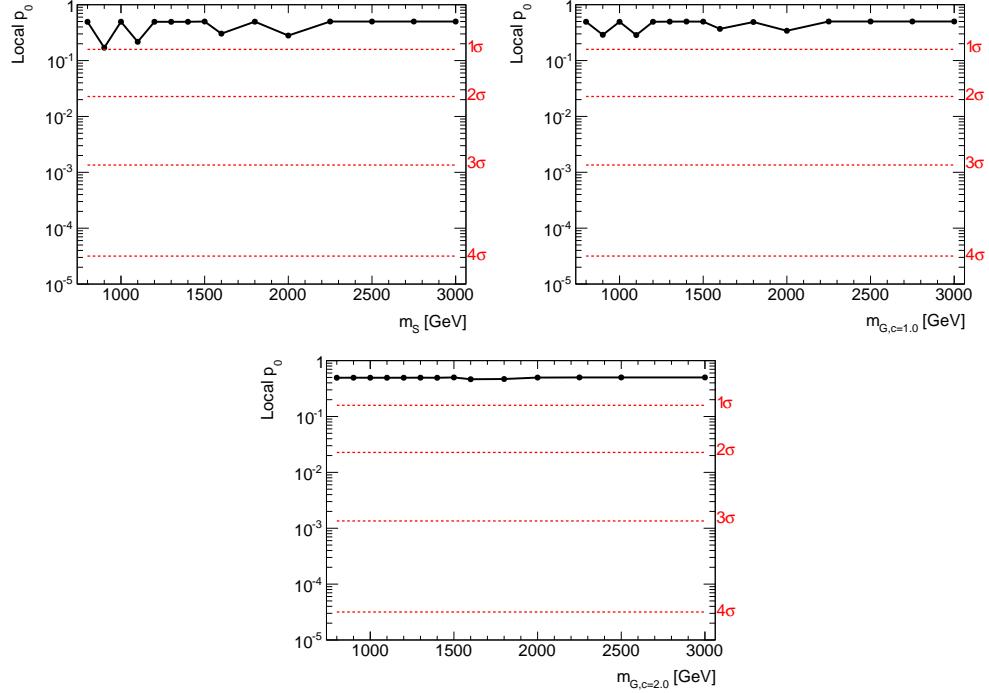
**Figure 9.18:** Kinematics of the large- $R$  jet system in data and prediction in the signal region after requiring 2  $b$ -tags split.

### 9.0.3 TEST OF THE BACKGROUND MODEL HYPOTHESIS

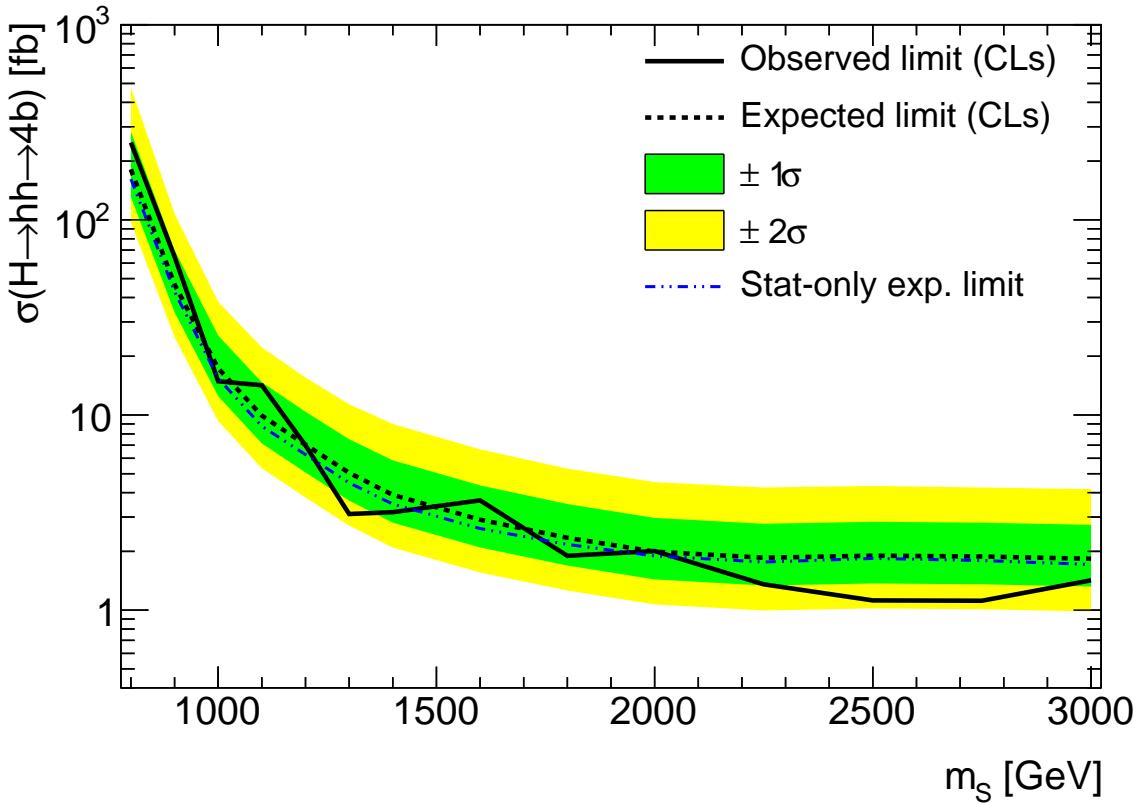
Here the results are displayed for the mass range that includes the boosted categories (ie. 800 GeV and above). In the range 800-1400 GeV the boosted categories are combined with the resolved categories. The full mass range results are collected in the resolved note.

A search for statistically significant deviation from the background model hypothesis is performed following the procedure described in Sec. ??, computing the local  $p_0$  value using the asymptotic approximation.

The background model is found to describe the data and no significant excess is observed. The smallest local  $p_0 = 0.175$  ( $1\sigma$ ) is found at 1100 GeV when fitting with the narrow scalar model. The local  $p_0$  values for the three signal models as a function of the resonance mass are shown in Fig. 9.19.



**Figure 9.19:** Local  $p_0$  of the (a) scalar, (b)  $c=1$  Graviton and (c)  $c=2$  Graviton.

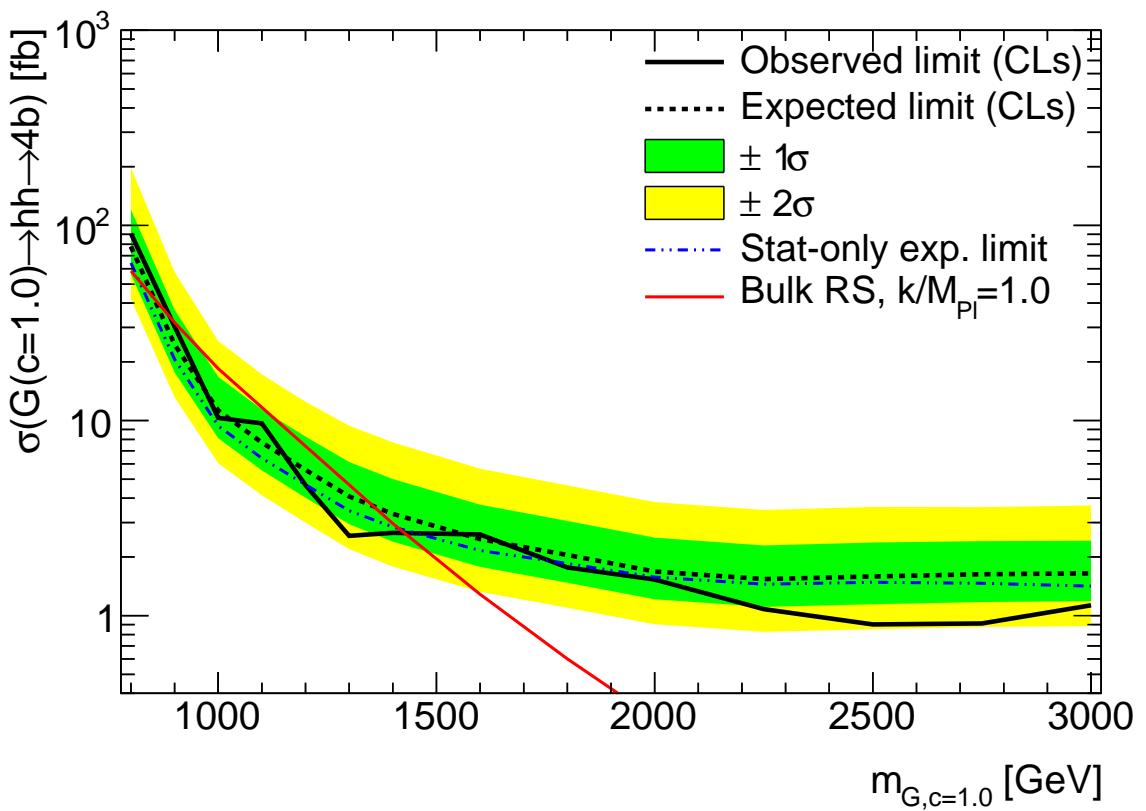


**Figure 9.20:** The expected and observed 95% C.L. upper exclusion limits for the boosted  $4b$  analysis calculated including all systematic uncertainties for the narrow scalar model. The dot-dashed line shows the expected limit when only statistical uncertainties are included. The limits are derived within the asymptotic approximation.

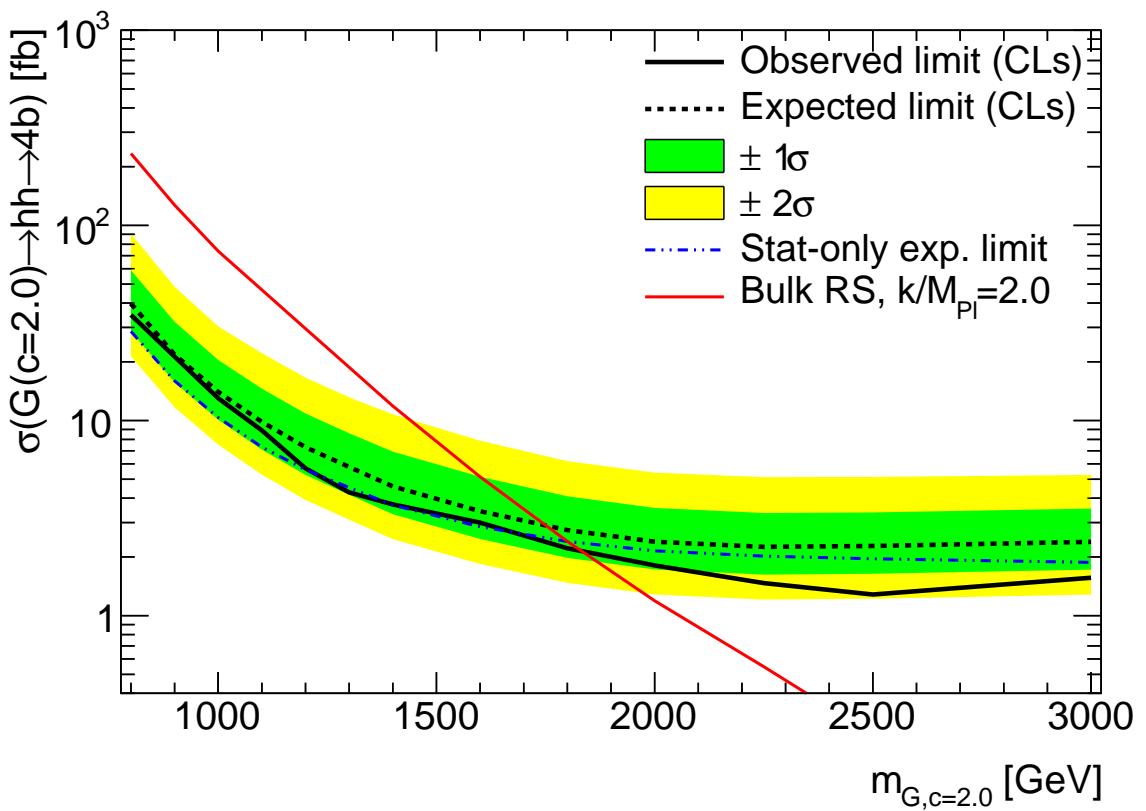
#### 9.0.4 OBSERVED LIMITS

The observed limit for the narrow scalar is shown in Fig. 9.22. The stat-only limit is also shown. The impact of systematic uncertainties is small. The observed limits for the Graviton models is shown in Fig. ?? for  $c=1$  and in Fig. ?? for  $c=2$ . These limits do not contain any of the resolved categories.

Figure 9.23 shows the pulls of the systematic uncertainty nuisance parameters and their correlations for the 2000 GeV mass point. One nuisance parameter (QCD\_ShapeCRHigh) in both the



**Figure 9.21:** The expected and observed 95% C.L. upper exclusion limits for the boosted  $4b$  analysis calculated including all systematic uncertainties for the  $c=1.0$  Graviton. The dot-dashed line shows the expected limit when only statistical uncertainties are included. The limits are derived within the asymptotic approximation.



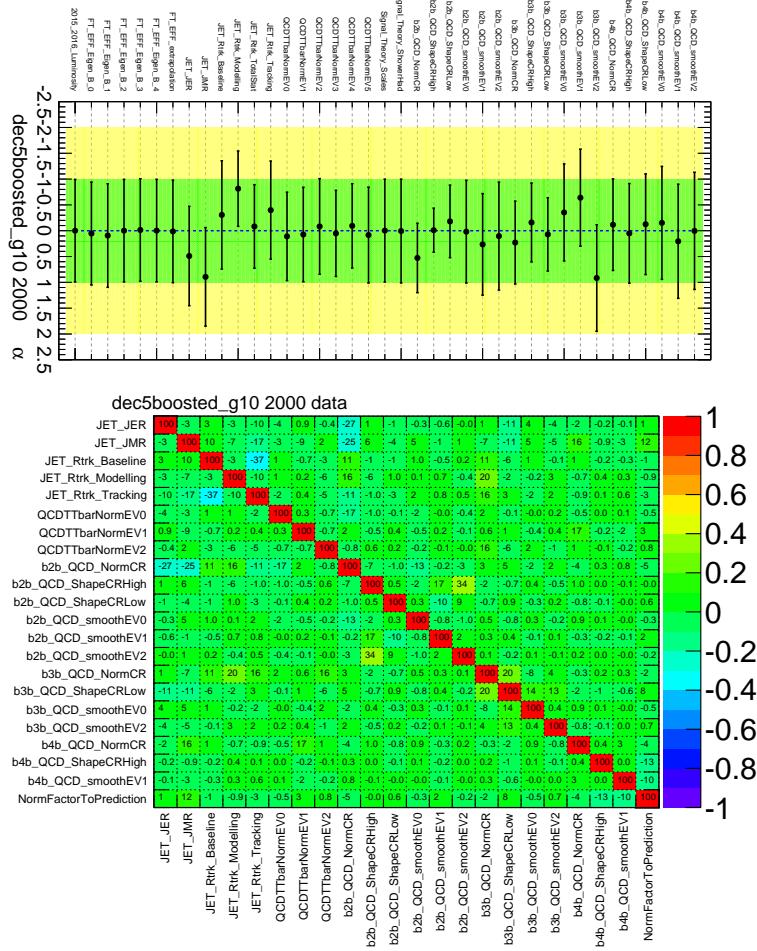
**Figure 9.22:** The expected and observed 95% C.L. upper exclusion limits for the boosted  $4b$  analysis calculated including all systematic uncertainties for the  $c=2.0$  Graviton. The dot-dashed line shows the expected limit when only statistical uncertainties are included. The limits are derived within the asymptotic approximation.

$2b$  and  $3b$  samples shows a significant constraint coming from the signal region data. This nuisance parameter corresponds to the shape uncertainty on the QCD background derived from the  $2b$  and  $3b$  control regions, as explained in Section 8.0.3. The prior probability distributions for this nuisance parameter is very broad, with the relative uncertainty on the background prediction reaching 15000% at high  $m_{bb}$ . This is because there is very little data in the control region at high mass to constrain the uncertainty. In the signal region however, the two events found suffice to constrain it: very tightly in comparison to the extremely loose prior constraint.

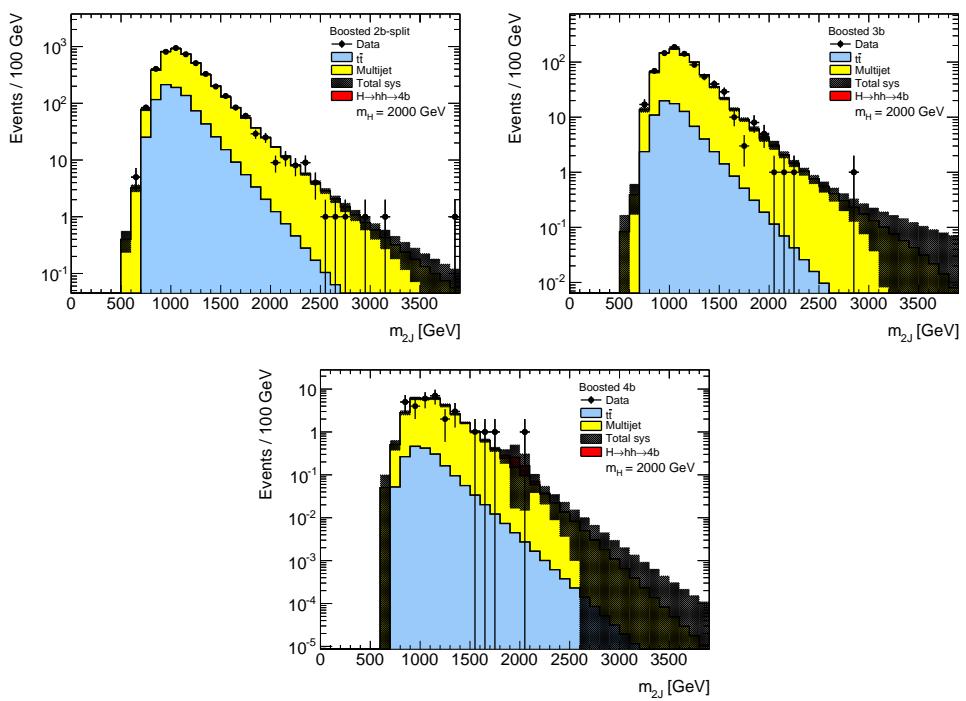
Examples of the fit used to set these limits are shown in Figures 9.24 and 9.25, where a narrow scalar is used for the signal model. In the case the best-fit is negative, the fit is repeated with  $\mu$  bounded to zero. This happens at several mass points, for example at 1.5, 2.5 and 3 TeV. At 2 TeV, the fitted signal is positive ( $\mu = 0.1 \pm 0.25$ ) though well consistent with the background-only hypothesis. In all fits, good agreement is seen between data and the background model.

The impact of the uncertainties on the fitted signal cross section is displayed in Fig. 9.26 for the three signal models at 2000 GeV. The parameters are ranked by their postfit impact. Only the leading 30 nuisance parameters are displayed.

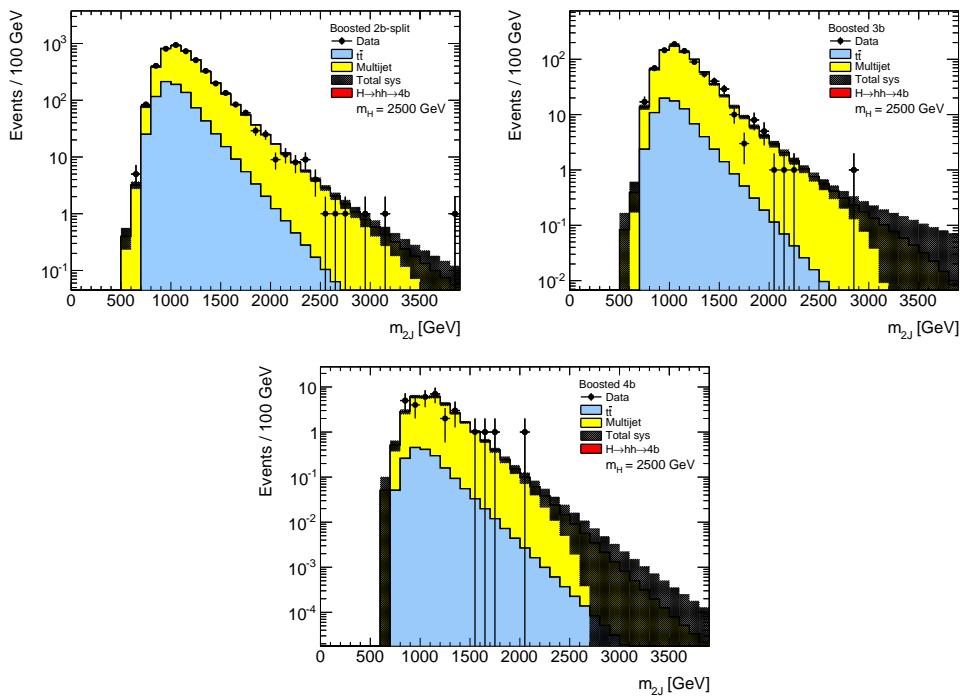
Not yet.



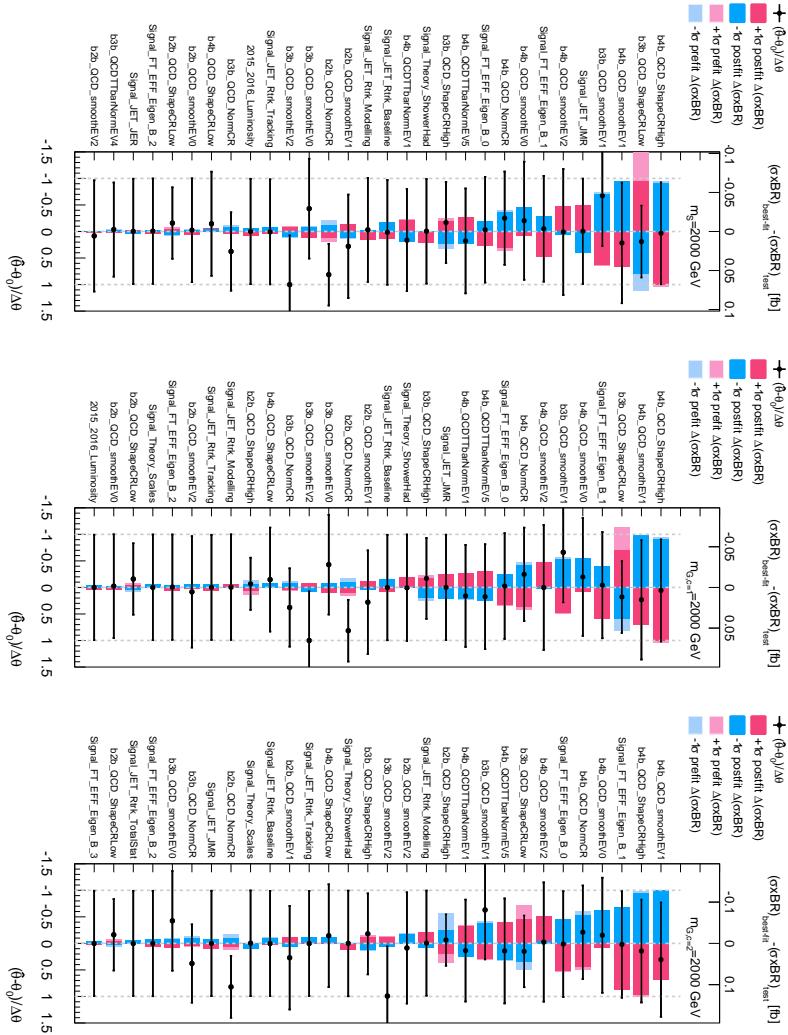
**Figure 9.23:** Nuisance parameters associated with the background modelling, after the conditional likelihood fit for a bulk RS graviton signal with  $m_{G_{KK}^*} = 2$  TeV and  $k/\bar{M}_{Pl} = 1.0$ . The tight constraints of  $2b\text{-QCD\_CRShape}$  and  $3b\text{-QCD\_CRShape}$  are a result of the nuisance parameter prior being unconstrained due to a lack of control region data at high mass.



**Figure 9.24:** Postfit distributions after fitting the data with the 2000 GeV signal hypothesis. The signal strength is slightly positive.



**Figure 9.25:** Postfit distributions after fitting the data with the 2500 GeV signal hypothesis. The signal strength is zero.



**Figure 9.26:** The impact of nuisance parameters on the fitted cross section, ranked by their postfit impact. The signal mass used in this fits is 2000 GeV, and the signal model is (a) narrow scalar, (b)  $c=1$  Graviton and (c)  $c=2$  Graviton.

*The aim of science is not to open the door to infinite wisdom, but to set a limit to infinite error.*

Bertolt Brecht

# 10

## Intepretation

In order to avoid the **Oops-Leon** cases, commonly accepted standard for announcing the discovery of a particle is that the number of observed events is 5 standard deviations ( $\sigma$ ) above the expected level of the background.

With no excess observed, a limit needs to be set on the cross section of the signal <sup>?</sup>.

One method to parametrise signal is called **singal morphing**. Specifically, in momentum morphing, not only the signal strength is scaled, but also the width is modified by a non-linear transformation.

An estimator should satisfy three criterias:

- Consistency: the value of the estimator should converge to the truth value if the sample size goes to infinity
- Efficiency: theory limits the variance of the true value, given a sample size N (Minimum Variance bound, or MVB). If the variance of the estimator is equal to the MVB the estimator is called efficient.
- Unbiased: the estimator should have no difference from the true value, otherwise it is biased.

Maximum likelihood estimator is a good estimator, based on these criterias.

For measurements, most of the time <https://arxiv.org/abs/1611.01927> is done to compare with generator level distributions. This accounts for detector effects, statistical fluctuations and background mis-identification. Since the analysis is a search, this  $bb \rightarrow b\bar{b}b\bar{b}$  is a search, unfolding is less applicable in this case.

*It is a far, far better thing that I do, than I have ever  
done; it is a far, far better rest that I go to than I have  
ever known.*

Charles Dickens

# 11

## Conclusion

Di-Higgs search has a short history, but will have a long future. This thesis presents a search for both resonant and non-resonant production of pairs of Standard Model Higgs bosons has been carried out in the dominant  $b\bar{b}b\bar{b}$  channel, using 27.5–36.1  $\text{fb}^{-1}$  of LHC  $pp$  collision data at  $\sqrt{s} = 13 \text{ TeV}$  collected by ATLAS in 2015 and 2016. The search sensitivity of this analysis exceeds that of the previous analysis of the  $\sqrt{s} = 13 \text{ TeV}$  2015 dataset<sup>3</sup> for non-resonant signal and also across the entire mass range of 260–3000 GeV for the resonance search, with significantly improvement in the high mass resonance sensitivities. The resolved analysis has each  $h \rightarrow b\bar{b}$  reconstructed as two separate  $b$ -tagged jets, and the boosted analysis has each  $h \rightarrow b\bar{b}$  reconstructed as a single large-radius jet associated with at least one small-radius  $b$ -tagged track-jet. The estimated background consists mainly of multi-jet and  $t\bar{t}$  events.

No significant excess is observed in the data. Upper limits on the production cross section times branching ratio to the  $b\bar{b}b\bar{b}$  final state are set for a narrow-width spin-0 scalar and for wider spin-2 resonances. The bulk RS model with  $k/\bar{M}_{\text{Pl}} = 1$  is excluded for masses between 313 and 1362 GeV, and the bulk RS model with  $k/\bar{M}_{\text{Pl}} = 2$  is excluded for masses below 1744 GeV. The 95% CL upper limit on the non-resonant production is 147 fb, which corresponds to 13.0 times the SM expectation.

This result confirms the great success of the Standard Model. The Higgs potential shape couldn't be very different from the SM predictions at TeV energy scale. Without any significant excess, the phase space for Beyond the Standard Model physics is further constrained.

Improvement with future Run 2 data could come  $b$ -tagging, especially efficiency increase in high  $p_T$  region. Other aspects include advanced trigger technologies, which can increase the signal event rate, and improved jet energy and mass resolution, which can increase the purity in selection. Together with the larger dataset, it is possible to double the current resonance search sensitivity. For non-resonance search, combined  $|\frac{\lambda}{\lambda_{\text{SM}}}| \sim 10$  is possible at the end of Run 2 in 2020.

For longer perspectives, di-Higgs searches and measurements will continue to be one of the most important analysis. It can constrain or hint the physics Beyond the Standard Model. In 2030 to 2040, the High Luminosity LHC will be able to constrain  $|\frac{\lambda}{\lambda_{\text{SM}}}| \sim 1$  with all the different channels from both ATLAS and CMS combined.

For even longer future developments in high energy experiments, all aspects—accelerator, detector, computation and theory—must advance together to answer the questions the Standard Model cannot answer, or find questions the Standard Model didn't ask<sup>70,71</sup>. Life is short for mankind, and the understanding of the universe is an endless journey. I am deeply honored to be a small part of this odyssey.

# References

- [1] C. Patrignani et al. Review of Particle Physics. *Chin. Phys.*, C40(10):100001, 2016. doi: 10.1088/1674-1137/40/10/100001.
- [2] W.J. Stirling.  $7/8$  and  $13/8$  TeV LHC luminosity ratios. 2013. URL [http://www.hep.physics.cern.ac.uk/~wstirlin/plots/lhclumi7813\\_2013\\_v0.pdf](http://www.hep.physics.cern.ac.uk/~wstirlin/plots/lhclumi7813_2013_v0.pdf).
- [3] Lyndon Evans. The Large Hadron Collider. *Annual Review of Nuclear and Particle Science*, 61(1):435–466, 2011. doi: 10.1146/annurev-nucl-102010-130438.
- [4] ATLAS Collaboration. ATLAS Luminosity Public Results, Run 2. 2015. URL <https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2>.
- [5] ATLAS Collaboration. Jet mass reconstruction with the ATLAS Detector in early Run 2 data. ATLAS-CONF-2016-035, 2016. URL <https://cds.cern.ch/record/2200211>.
- [6] ATLAS Collaboration. Optimisation of the ATLAS  $b$ -tagging performance for the 2016 LHC Run. *ATL-PHYS-PUB-2016-012*, (ATL-PHYS-PUB-2016-012), Jun 2016. URL <https://cds.cern.ch/record/2160731>.
- [7] Kaustubh Agashe, Hooman Davoudiasl, Gilad Perez, and Amarjit Soni. Warped gravitons at the CERN LHC and beyond. *Phys. Rev. D*, 76:036006, 2007. doi: 10.1103/PhysRevD.76.036006.
- [8] Liam Fitzpatrick, Jared Kaplan, Lisa Randall, and Lian-Tao Wang. Searching for the Kaluza-Klein graviton in bulk RS models. *JHEP*, 09:013, 2007. doi: 10.1088/1126-6708/2007/09/013.
- [9] T. D. Lee. A theory of spontaneous  $t$  violation. *Phys. Rev. D*, 8:1226–1239, Aug 1973. doi: 10.1103/PhysRevD.8.1226.
- [10] G.C. Branco et al. Theory and phenomenology of two-Higgs-doublet models. *Phys. Rept.*, 516:1, 2012. doi: 10.1016/j.physrep.2012.02.002.
- [11] Graham D. Kribs and Adam Martin. Enhanced di-higgs production through light colored scalars. *Phys. Rev. D*, 86:095023, 2012. doi: 10.1103/PhysRevD.86.095023.

- [12] R. Gröber and M. Mühlleitner. Composite Higgs boson pair production at the LHC. *JHEP*, 06:020, 2011. doi: [10.1007/JHEP06\(2011\)020](https://doi.org/10.1007/JHEP06(2011)020).
- [13] Roberto Contino et al. Anomalous couplings in double Higgs production. *JHEP*, 08:154, 2012. doi: [10.1007/JHEP08\(2012\)154](https://doi.org/10.1007/JHEP08(2012)154).
- [14] ATLAS Collaboration. Search for pair production of Higgs bosons in the  $b\bar{b}b\bar{b}$  final state using proton–proton collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector. *Phys. Rev. D*, 94: 052002, 2016. doi: [10.1103/PhysRevD.94.052002](https://doi.org/10.1103/PhysRevD.94.052002).
- [15] ATLAS Collaboration. Search for Higgs boson pair production in the  $b\bar{b}b\bar{b}$  final state from pp collisions at  $\sqrt{s} = 8$  TeV with the ATLAS detector. *Eur. Phys. J. C*, 75:412, 2015. doi: [10.1140/epjc/s10052-015-3628-x](https://doi.org/10.1140/epjc/s10052-015-3628-x).
- [16] ATLAS Collaboration. Search for Higgs Boson Pair Production in the  $\gamma\gamma b\bar{b}$  Final State Using  $pp$  Collision Data at  $\sqrt{s} = 8$  TeV from the ATLAS Detector. *Phys. Rev. Lett.*, 114: 081802, 2015. doi: [10.1103/PhysRevLett.114.081802](https://doi.org/10.1103/PhysRevLett.114.081802).
- [17] David Griffiths. *Introduction to elementary particles*. 2008.
- [18] Christopher G. Tully. *Elementary particle physics in a nutshell*. 2011.
- [19] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 2014. ISBN 1107034736, 9781107034730.
- [20] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1, 2012. doi: [10.1016/j.physletb.2012.08.020](https://doi.org/10.1016/j.physletb.2012.08.020).
- [21] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, 716:30, 2012. doi: [10.1016/j.physletb.2012.08.021](https://doi.org/10.1016/j.physletb.2012.08.021).
- [22] R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, P. Torrielli, E. Vryonidou, and M. Zaro. Higgs pair production at the LHC with NLO and parton-shower effects. *Phys. Lett.*, B732:142–149, 2014. doi: [10.1016/j.physletb.2014.03.026](https://doi.org/10.1016/j.physletb.2014.03.026).
- [23] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016. doi: [10.23731/CYRM-2017-002](https://doi.org/10.23731/CYRM-2017-002).

- [24] Graham D. Kribs, Andreas Maier, Heidi Rzezak, Michael Spannowsky, and Philip Waite. Electroweak oblique parameters as a probe of the trilinear Higgs boson self-interaction. *Phys. Rev.*, D95(9):093004, 2017. doi: 10.1103/PhysRevD.95.093004.
- [25] Viviana Cavaliere and Gabriel Facini. Summary of Limits on BSM Models from Diboson Searches. Technical Report ATL-COM-PHYS-2016-1071, CERN, Geneva, Sep 2016. URL <https://cds.cern.ch/record/2203605>.
- [26] Georges Aad et al. Searches for Higgs boson pair production in the  $hh \rightarrow bb\tau\tau, \gamma\gamma WW^*, \gamma\gamma bb, bbbb$  channels with the ATLAS detector. *Phys. Rev.*, D92:092004, 2015. doi: 10.1103/PhysRevD.92.092004.
- [27] Lyndon R Evans and Philip Bryant. LHC Machine. *J. Instrum.*, 3:So8001. 164 p, 2008. URL <https://cds.cern.ch/record/1129806>. This report is an abridged version of the LHC Design Report (CERN-2004-003).
- [28] ATLAS Collaboration. The ATLAS experiment at the CERN Large Hadron Collider. *JINST*, 3:So8003, 2008. doi: 10.1088/1748-0221/3/08/So8003.
- [29] CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):So8004, 2008. URL <http://stacks.iop.org/1748-0221/3/i=08/a=So8004>.
- [30] LHCb Collaoration. The LHCb Detector at the LHC. *JINST*, 3:So8005, 2008. doi: 10.1088/1748-0221/3/08/So8005.
- [31] ALICE Collaboration. The ALICE experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):So8002, 2008. URL <http://stacks.iop.org/1748-0221/3/i=08/a=So8002>.
- [32] ATLAS Collaboration. Luminosity Determination in  $pp$  Collisions at  $\sqrt{s} = 7$  TeV Using the ATLAS Detector at the LHC. *Eur. Phys. J.*, C 71:1630, 2011. doi: 10.1140/epjc/s10052-011-1630-5.
- [33] Paul Collier for the LHC team. LHC Machine Status. CERN Resource Review Board, 2015. URL <https://cds.cern.ch/record/2063924/files/CERN-RRB-2015-119.PDF>.
- [34] Giulia Papotti for the LHC team. LHC Machine Status Report. CERN Resource Review Board, 2016. URL [https://indico.cern.ch/event/563488/contributions/2277292/attachments/1340292/2019570/20160921\\_LHCC.pdf](https://indico.cern.ch/event/563488/contributions/2277292/attachments/1340292/2019570/20160921_LHCC.pdf).

- [35] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008. doi: 10.1088/1748-0221/3/08/S08003.
- [36] ATLAS Collaboration. ATLAS Insertable B-Layer Technical Design Report. CERN-LHCC-2010-013, ATLAS-TDR-19, Sep 2010. URL <https://cds.cern.ch/record/1291633>.
- [37] ATLAS Collaboration. Expected performance of the ATLAS  $b$ -tagging algorithms in Run-2. ATL-PHYS-PUB-2015-022, 2015. URL <https://cds.cern.ch/record/2037697>.
- [38] Morad Aaboud et al. Study of the material of the ATLAS inner detector for Run 2 of the LHC. *JINST*, 12(12):P12009, 2017. doi: 10.1088/1748-0221/12/12/P12009.
- [39] ATLAS Collaboration. Performance of the ATLAS Trigger System in 2015. *Eur. Phys. J. C*, 77:317, 2017. doi: 10.1140/epjc/s10052-017-4852-3.
- [40] M. Aaboud et al. Performance of the ATLAS Track Reconstruction Algorithms in Dense Environments in LHC Run 2. *Eur. Phys. J.*, C77(10):673, 2017. doi: 10.1140/epjc/s10052-017-5225-7.
- [41] ATLAS Collaboration. Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1. *Eur. Phys. J. C*, 77:490, 2017. doi: 10.1140/epjc/s10052-017-5004-5.
- [42] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The catchment area of jets. *JHEP*, 04:005, 2008. doi: 10.1088/JHEP04(2008)05.
- [43] ATLAS Collaboration. Jet energy measurement with the ATLAS detector in proton-proton collisions at  $\sqrt{s} = 7$  TeV. *Eur. Phys. J. C*, 73:2304, 2013. doi: 10.1140/epjc/s10052-013-2304-2.
- [44] ATLAS Collaboration. Jet Calibration and Systematic Uncertainties for Jets Reconstructed in the ATLAS Detector at  $\sqrt{s} = 13$  TeV. ATL-PHYS-PUB-2015-015, 2015. URL <https://cds.cern.ch/record/2037613>.
- [45] D. Krohn, J. Thaler, and L.-T. Wang. Jet trimming. *JHEP*, 02:084, 2010. doi: 10.1007/JHEP02(2010)084.
- [46] Stephen D. Ellis and Davison E. Soper. Successive combination jet algorithm for hadron collisions. *Phys. Rev. D*, 48:3160, 1993. doi: 10.1103/PhysRevD.48.3160.

- [47] ATLAS Collaboration. Performance of jet substructure techniques for large- $R$  jets in proton–proton collisions at  $\sqrt{s} = 7$  TeV using the ATLAS detector. *JHEP*, 09:076, 2013. doi: 10.1007/JHEP09(2013)076.
- [48] ATLAS Collaboration. Performance of  $b$ -jet identification in the ATLAS experiment. *JINST*, 11:P04008, 2016. doi: 10.1088/1748-0221/11/04/P04008.
- [49] Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton–proton collision data. Technical Report ATLAS-CONF-2016-024, CERN, Geneva, Jun 2016. URL <http://cds.cern.ch/record/2157687>.
- [50] Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using  $p\bar{p}$  collisions at  $\sqrt{s} = 13$  TeV. Technical Report ATLAS-CONF-2017-029, CERN, Geneva, May 2017. URL <http://cds.cern.ch/record/2261772>.
- [51] Georges Aad et al. Muon reconstruction performance of the ATLAS detector in proton–proton collision data at  $\sqrt{s} = 13$  TeV. *Eur. Phys. J.*, C76(5):292, 2016. doi: 10.1140/epjc/s10052-016-4120-y.
- [52] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. PYTHIA 6.4 physics and manual. *JHEP*, 05:026, 2006. doi: 10.1088/1126-6708/2006/05/026.
- [53] ATLAS Collaboration. Summary of ATLAS Pythia 8 tunes. 2012. URL <http://cdsweb.cern.ch/record/1474107>.
- [54] S. Agostinelli et al. GEANT4: a simulation toolkit. *Nucl. Instrum. Meth. A*, 506:250–303, 2003. doi: 10.1016/S0168-9002(03)01368-8.
- [55] ATLAS Collaboration. The ATLAS Simulation Infrastructure. *Eur. Phys. J. C*, 70:823–874, 2010. doi: 10.1140/epjc/s10052-010-1429-9.
- [56] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 06:043, 2010. doi: 10.1007/JHEP06(2010)043.
- [57] Peter Zeiler Skands. Tuning Monte Carlo generators: The Perugia tunes. *Phys. Rev. D*, 82:074018, 2010. doi: 10.1103/PhysRevD.82.074018.

- [58] Michał Czakon and Alexander Mitov. Top++: A program for the calculation of the top-pair cross-section at hadron colliders. *Comput. Phys. Commun.*, 185:2930, 2014. ISSN 0010-4655. doi: [10.1016/j.cpc.2014.06.021](https://doi.org/10.1016/j.cpc.2014.06.021).
- [59] J. Alwall et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014. doi: [10.1007/JHEP07\(2014\)079](https://doi.org/10.1007/JHEP07(2014)079).
- [60] Richard D. Ball et al. Parton distributions with LHC data. *Nucl. Phys. B*, 867:244, 2013. doi: [10.1016/j.nuclphysb.2012.10.003](https://doi.org/10.1016/j.nuclphysb.2012.10.003).
- [61] A. Carvalho. Gravity particles from Warped Extra Dimensions, predictions for LHC. 2017.
- [62] M. Bahr et al. Herwig++ physics and manual. *Eur. Phys. J. C*, 58:639–707, 2008. doi: [10.1140/epjc/s10052-008-0798-9](https://doi.org/10.1140/epjc/s10052-008-0798-9).
- [63] Hung-Liang Lai, Marco Guzzi, Joey Huston, Zhao Li, Pavel M. Nadolsky, Jon Pumplin, and C. P. Yuan. New parton distributions for collider physics. *Phys. Rev. D*, 82:074024, 2010. doi: [10.1103/PhysRevD.82.074024](https://doi.org/10.1103/PhysRevD.82.074024).
- [64] Pavel M. Nadolsky et al. Implications of cteq global analysis for collider observables. *Phys. Rev. D*, 78:013004, 2008. doi: [10.1103/PhysRevD.78.013004](https://doi.org/10.1103/PhysRevD.78.013004).
- [65] Michael H. Seymour and Andrzej Siodmok. Constraining MPI models using  $\sigma_{\text{eff}}$  and recent Tevatron and LHC Underlying Event data. *JHEP*, 10:113, 2013. doi: [10.1007/JHEP10\(2013\)113](https://doi.org/10.1007/JHEP10(2013)113).
- [66] S. Dawson, S. Dittmaier, and M. Spira. Neutral Higgs boson pair production at hadron colliders: QCD corrections. *Phys. Rev. D*, 58:115012, 1998. doi: [10.1103/PhysRevD.58.115012](https://doi.org/10.1103/PhysRevD.58.115012).
- [67] T. Plehn, M. Spira, and P.M. Zerwas. Pair production of neutral Higgs particles in gluon-gluon collisions. *Nucl. Phys. B*, 479:46, 1996. doi: [10.1016/0550-3213\(96\)00418-X](https://doi.org/10.1016/0550-3213(96)00418-X). [Erratum: *Nucl. Phys. B* 531 (1998) 655].
- [68] S. Borowka, N. Greiner, G. Heinrich, S. P. Jones, M. Kerner, J. Schlenk, U. Schubert, and T. Zirke. Higgs boson pair production in gluon fusion at NLO with full top-quark mass dependence. *Phys. Rev. Lett.*, 117(1):012001, 2016. doi: [10.1103/PhysRevLett.117.012001](https://doi.org/10.1103/PhysRevLett.117.012001).
- [69] S. Borowka, N. Greiner, G. Heinrich, S. P. Jones, M. Kerner, J. Schlenk, and T. Zirke. Full top quark mass dependence in Higgs boson pair production at NLO. *JHEP*, 10:107, 2016. doi: [10.1007/JHEP10\(2016\)107](https://doi.org/10.1007/JHEP10(2016)107).

- [70] Burton Richter. High Energy Colliding Beams; What Is Their Future? *Rev. Accel. Sci. Tech.*, 7:1–8, 2014. doi: 10.1142/9789814651493\_0001,10.1142/S1793626814300011.
- [71] Stephen Hawking and Gordon Kane. Should China build the Great Collider? 2018.