Measurement of boosted VH($b\bar{b}$) using MVA method at ATLAS

Jingyi Han

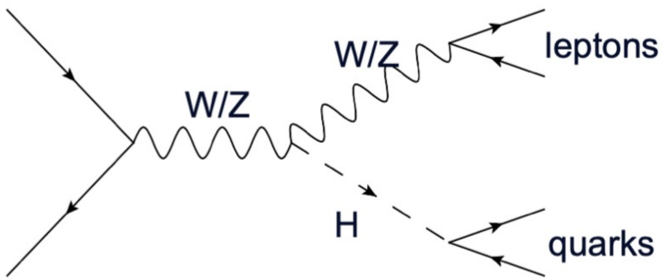
Shandong University

China LHC Physics Workshop, 27 Nov. 2021

Overview

➢ **Three VH($q\bar{q}$) stand-alone analyses** with full Run 2 dataset:

- VH($b\bar{b}$) *resolved* - Eur. Phys. J. C 81 (2021) 178
- VH($b\bar{b}$) *boosted* - Phys. Lett. B 816 (2021) 136204
- VH($c\bar{c}$) - ATLAS-CONF-2021-021, paper draft



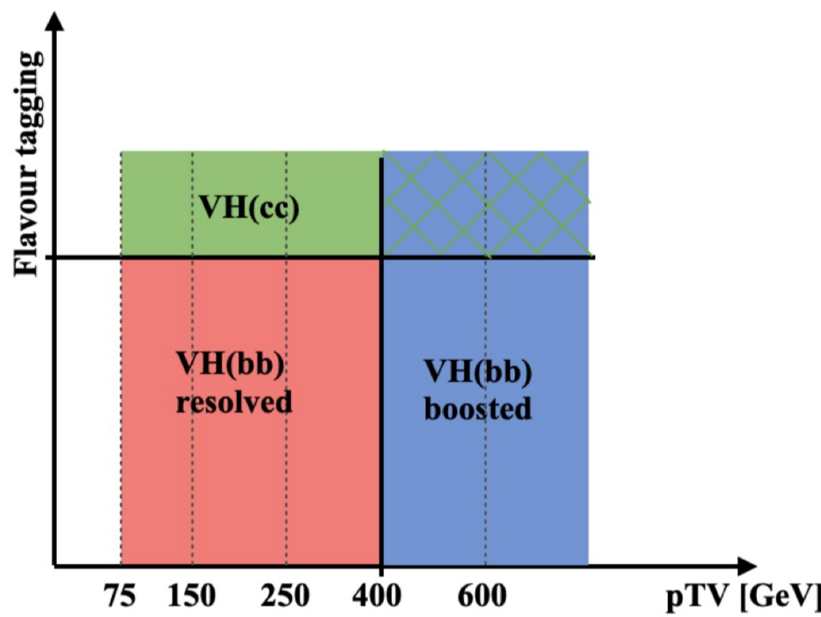
V decays:

- $Z \rightarrow \nu\bar{\nu}$: "0-lepton";
- $W \rightarrow l\nu$ $(l = e, \mu)$: "1-lepton";
- $Z \rightarrow ll$ $(l = e, \mu)$: "2-lepton";

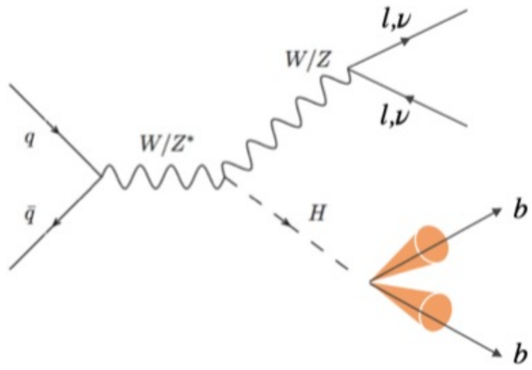
➢ **Goal of VH**($b\bar{b}/c\bar{c}$) **"legacy analysis"**: design one **coherent**, **harmonized** approach to extract VH($b\bar{b}/c\bar{c}$) results

Analysis strategy

- No overlap between VH(bb) resolved and VH(cc) analysis
- Possible overlap between VH(bb) boosted and VH(cc) analysis
- Simple combination strategy between VH(bb) resolved and VH(bb) boosted tested at the moment:
  - Use resolved analysis in p_T^V **< 400 GeV** region
  - Use boosted analysis in p_T^V **> 400 GeV** region
    - Add a split at **600 GeV**

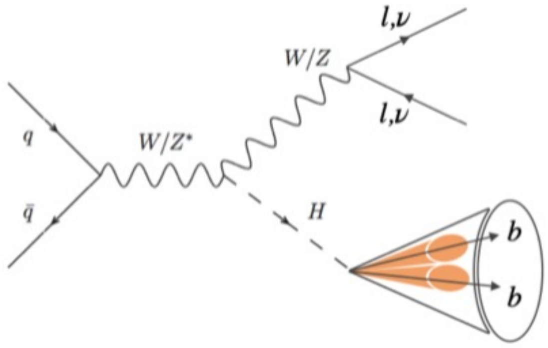


Event selection



VH($b\bar{b}$) resolved

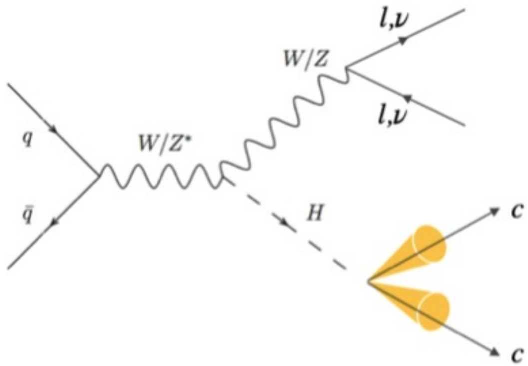
- $\geq$ 2 pFlow jets with exactly **2 b-tags**;
- **150 GeV $< P_T^V <$ 400 GeV (75GeV $< P_T^V <$ 400 GeV in 2L)**



VH($b\bar{b}$) boosted

- $\geq$ 1 large-R calo jet;
- Leading large-R associated with at least 2 VR jets;
- 2 b-tags in the leading 3 associated VR jets;
- **$P_T^V >$ 400 GeV**

Final state discussed in following slides.



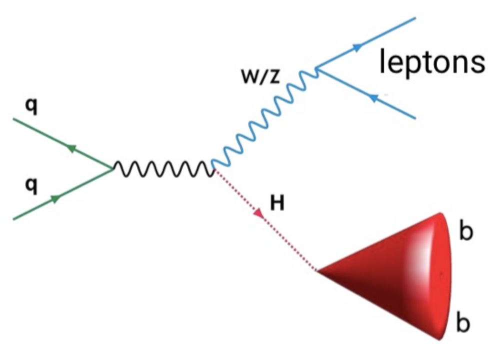
VH($c\bar{c}$) analysis

- $\geq$ 2 pFlow jets;
- 2 leading jets **c-tagged + not b-tagged**
- $P_T^V >$ 150 GeV ($P_T^V >$ 75 GeV in 2L)

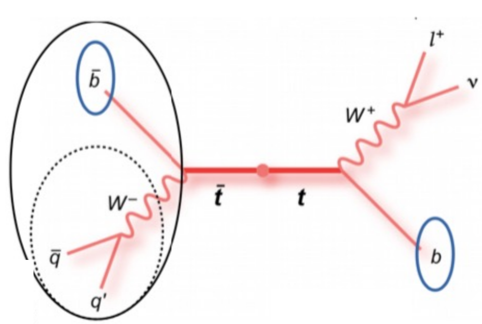
Event categorization

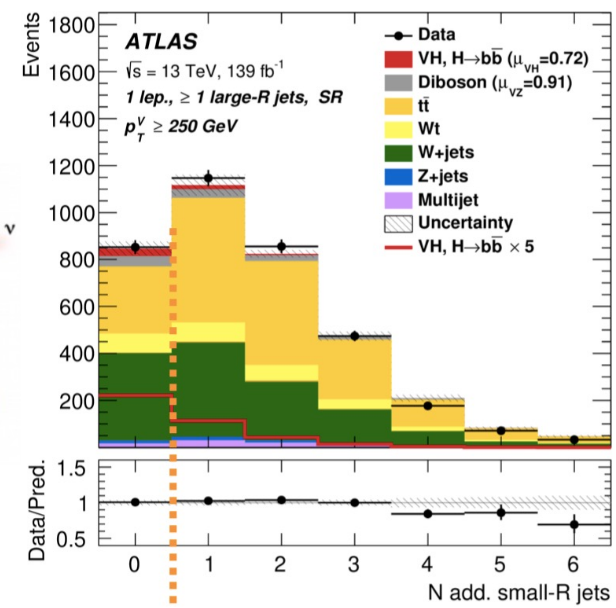
Channel	Categories					
	400< p_T^V <600GeV			p_T^V > 600GeV		
	0 additional b-tagged track jets: **Signal region**		≥ 1 add. b-tagged track jets	0 additional b-tagged track jets: **Signal region**		≥ 1 add. b-tagged track jets
	0 add. small-R jets	≥ 1 add. small-R jets	**top control region**	0 add. small-R jets	≥ 1 add. small-R jets	**top control region**
0-lepton	high purity	low purity		high purity	low purity	
1-lepton	high purity	low purity		high purity	low purity	
2-lepton	SR			SR		

Signal:



ttbar:



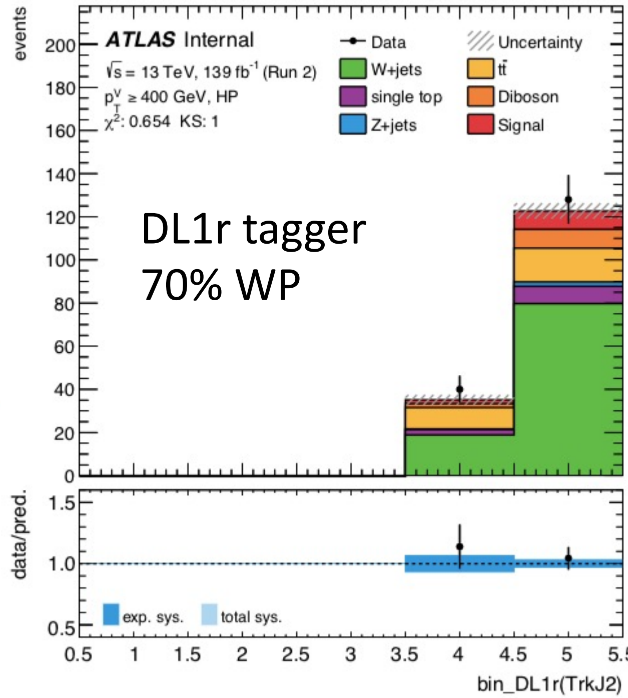


(Feynman diagrams at leading order)

Multivariate analysis

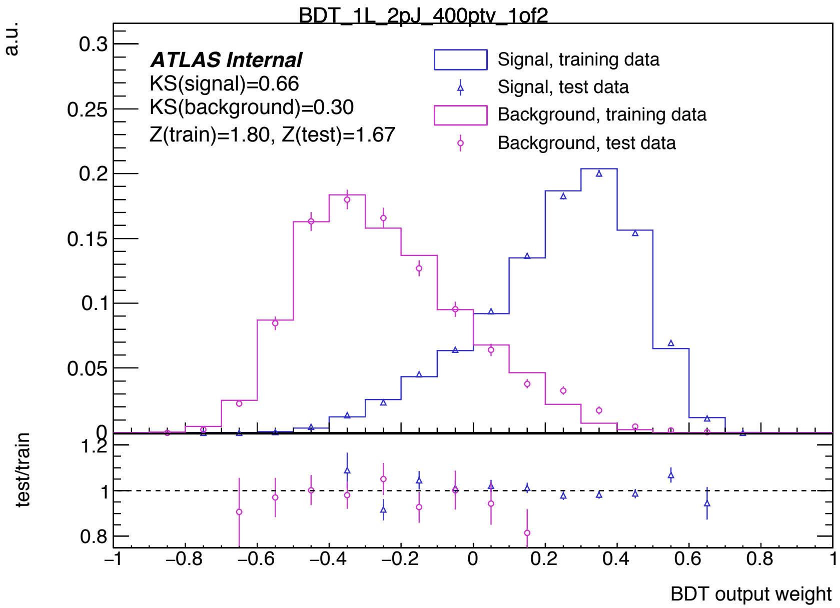
- **Multivariate analysis** used since long time in VH($b\bar{b}$) resolved analysis
  - **New: Introduce also in VH($b\bar{b}$) boosted analysis and in VH($c\bar{c}$) analysis**
- Baseline: Boosted Decision Tree (**BDT**)
- Input variables in VH($b\bar{b}$) boosted are inspired by information used in resolved MVA

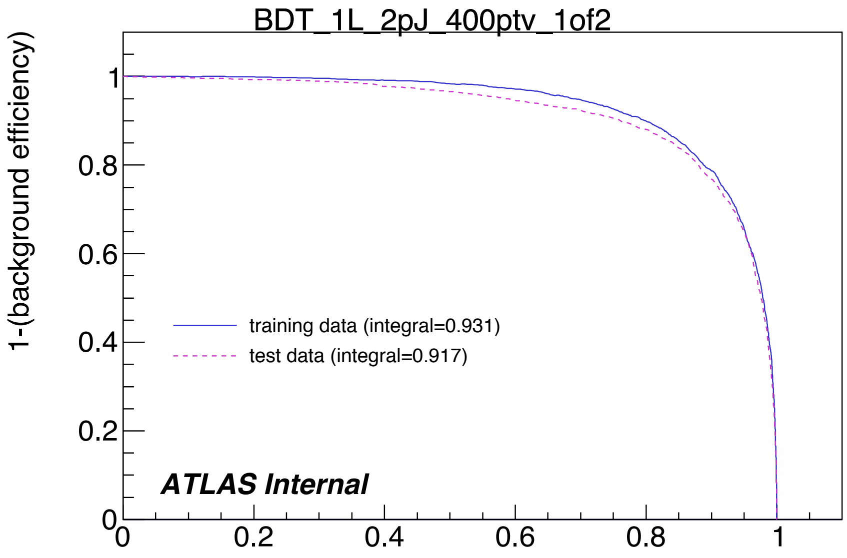
	0L	1L	2L
mJ			
dR(TrkJ1,TrkJ2)			
pT(TrkJ1)			
pT(TrkJ2)			
pTV			
dPhi(V,J)			
bin_DL1r(TrkJ1)			
bin_DL1r(TrkJ2)			
MEff			
dY(V,J)	n/a		
Lepton pT imbalance / cosTheta(l,Z)	n/a	n/a	
N(add. calo jets)			
N(trk jets in large-R jet)			



MVA baseline settings: 1L as an example

	$p_T^V > 400$ GeV
MiniNodeSize	5
MaxDepth	3
NTrees	200
nCuts	60
AdaBoostedBeta	0.35





Training and test agreement looks ok.

Boosted MVA fit

- First results of the MVA fit in the boosted regime
- Fit performed in $p_T^V > 400$ GeV region
  - Split at $p_T^V = 600$ GeV
  - HP SR and LP SR merged in $p_T^V > 600$ GeV
  - MVA training performed in $p_T^V > 400$ GeV, but evaluated in 400 GeV $< p_T^V < 600$ GeV and $p_T^V > 600$ GeV
- Preliminary studies considering the statistic only

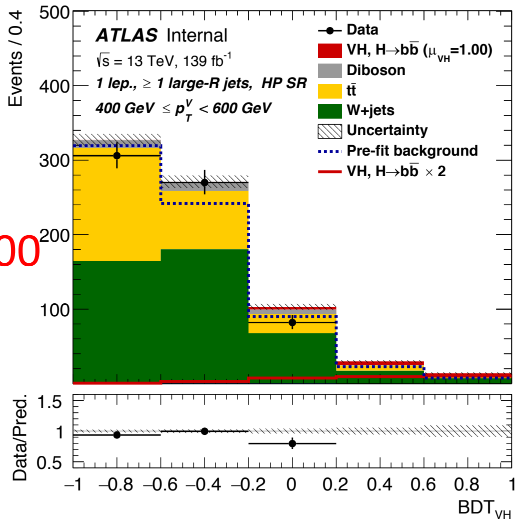
<table>
<tr><td colspan="2"></td><td colspan="2">0L</td><td colspan="2">1L</td><td colspan="2">2L</td></tr>
<tr><td colspan="2"></td><td colspan="2"># add small-R jets</td><td colspan="2"># add small-R jets</td><td colspan="2"># add small-R jets</td></tr>
<tr><td colspan="2"></td><td>0</td><td>>=1</td><td>0</td><td>>=1</td><td>0</td><td>>=1</td></tr>
<tr><td rowspan="2">400-600 GeV</td><td>no add b-tagged jets</td><td>HP SR</td><td>LP SR</td><td>HP SR</td><td>LP SR</td><td colspan="2" rowspan="2">SR</td></tr>
<tr><td>add b-tagged jets</td><td colspan="2">CR</td><td colspan="2">CR</td></tr>
<tr><td rowspan="2">>600 GeV</td><td>no add b-tagged jets</td><td colspan="2">(HP + LP) SR</td><td colspan="2">(HP + LP) SR</td><td colspan="2" rowspan="2">SR</td></tr>
<tr><td>add b-tagged jets</td><td colspan="2">CR</td><td colspan="2">CR</td></tr>
</table>

Fit in 8 SRs and 4 CRs

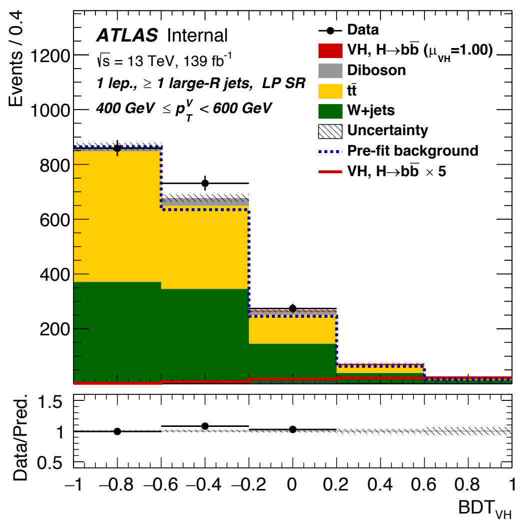
Post-fit plots: 1L

400_600 GeV

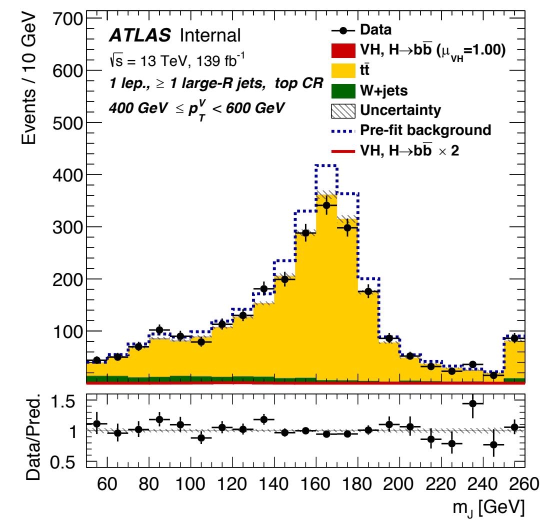
HP SR



LP SR

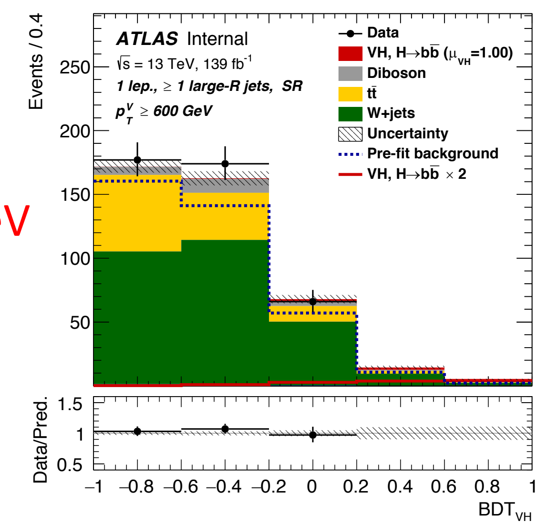


CR

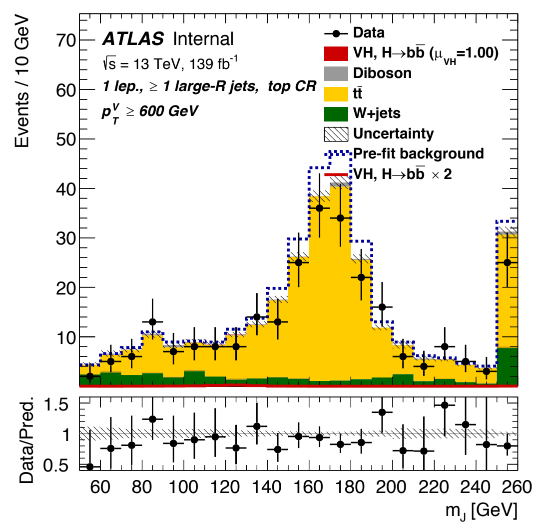


P_T^V > 600 GeV

SR



CR



Cut-based analysis

- Cut based strategy used in the standard alone boosted VH($b\bar{b}$) analysis

Selection	0 lepton channel	1 lepton channel		2 lepton channel	
		e sub-channel	μ sub-channel	e sub-channel	μ sub-channel
Trigger	$E_{\mathrm{T}}^{\mathrm{miss}}$	Single lepton	$E_{\mathrm{T}}^{\mathrm{miss}}$	Single lepton	$E_{\mathrm{T}}^{\mathrm{miss}}$
Leptons	0 VH-*loose* lepton	1 WH-*signal* lepton no second VH-*loose* lepton		$\geq$ 1 ZH-*signal* lepton 2 VH-*loose* leptons	
$E_{\mathrm{T}}^{\mathrm{miss}}$	> 250 GeV	> 50GeV	-	-	
p_{T}^{V}	p_{T}^{V} > 250 GeV				
Large-R jet	at least one large-R jet, $p_{\mathrm{T}} > 250$ GeV, $\lvert\eta\rvert < 2$				
Track-Jets	at least two track-jets, $p_{\mathrm{T}} > 10$ GeV, $\lvert\eta\rvert < 2.5$, matched to the leading large-R jet				
b-jets	leading two track-jets matched to the leading large-R must be b-tagged				
m_{J}	> 50 GeV				
min[$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}$,jets)]	> 30°	-			
$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, H_{\mathrm{cand}})$	> 120°	-			
$\Delta\phi$ ($E_{\mathrm{T}}^{\mathrm{miss}}$, $E_{\mathrm{T,\,trk}}^{\mathrm{miss}}$)	< 90°	-			
$\lvert\Delta Y(V,H)\rvert$	-	$\lvert\Delta Y(V,H)\rvert < 1.4$			
m_{ll}	-			66 GeV< m_{ll} < 116 GeV	
lepton p_{T} imbalance	-			$(p_{\mathrm{T}}^{l_1} - p_{\mathrm{T}}^{l_2})/p_{\mathrm{T}}^{Z} < 0.8$	
lepton flavor	-			two lepton same flavour	
lepton charge	-			opposite sign muons	

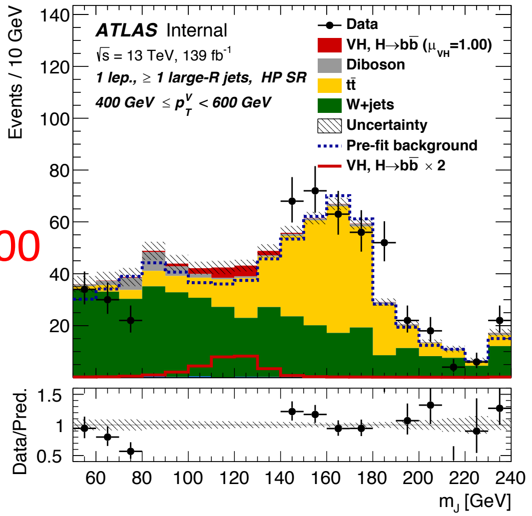
Same large-R jet cuts as MVA analysis

- Final discriminating variable is **mJ**, the leading large-R jet mass

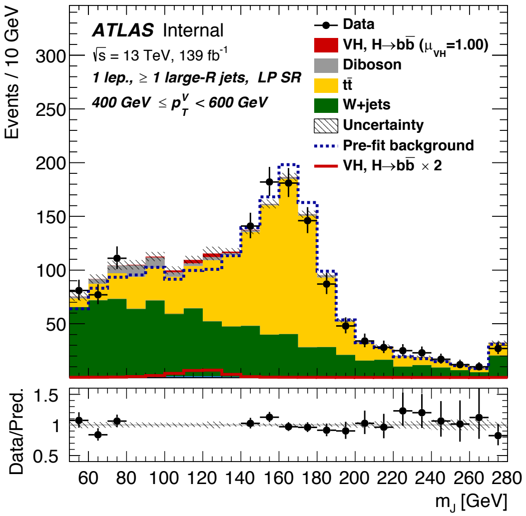
post-fit plots: 1L (cut based)

400_600 GeV

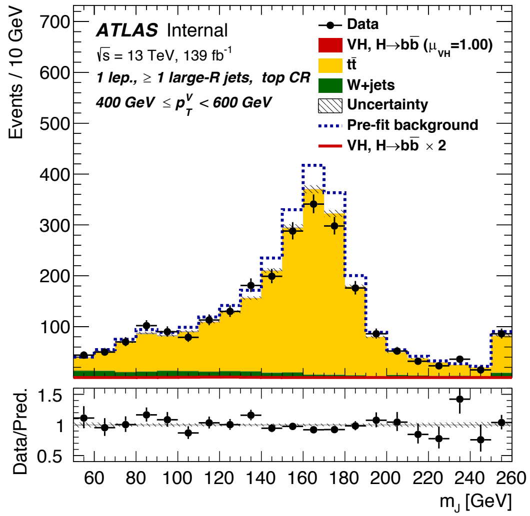
HP SR



LP SR

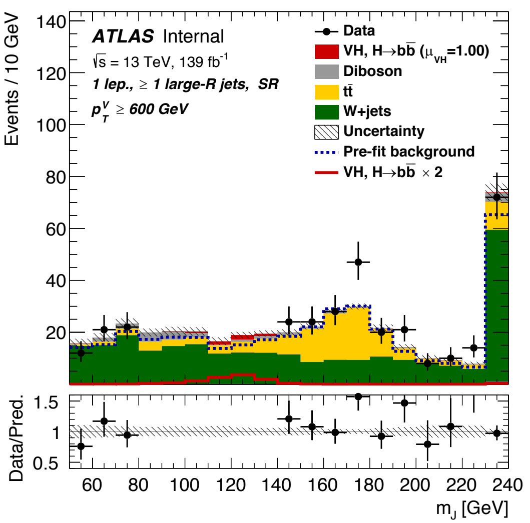


CR

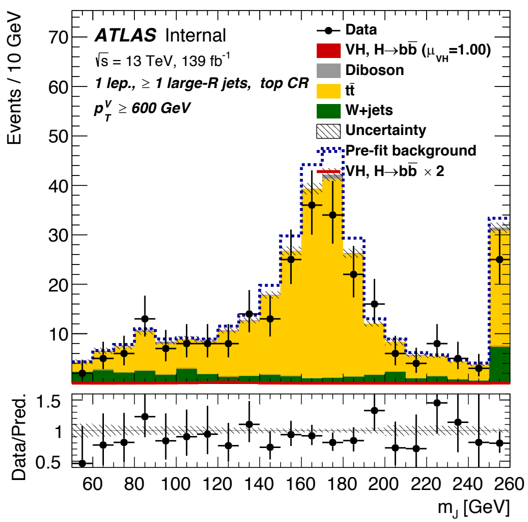


P_T^V > 600 GeV

SR



CR



Significance: MVA vs. cut based

	cut based	MVA	MVA improvement
0-lepton fit	1.04σ	2.10σ	+102%
1-lepton fit	1.06σ	2.35σ	+122%
2-lepton fit	0.70σ	1.46σ	+109%
combined fit	1.64σ	3.48σ	+112%

(Significance are from statistic only maximum-profile-likelihood fit)

Summary

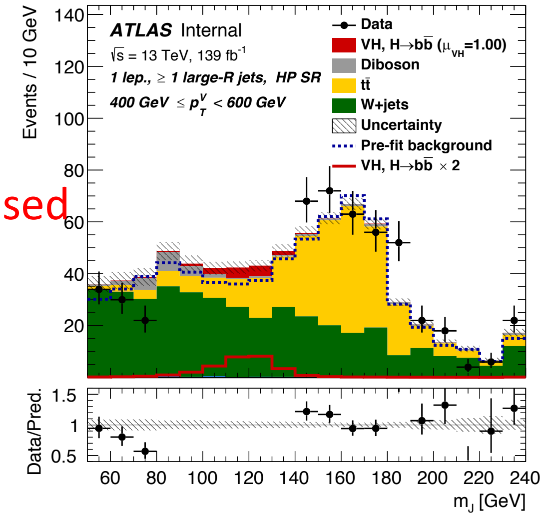
- Introduced the **VH($b\bar{b}$/$c\bar{c}$) "legacy" analysis**
- Showed preliminary results for MVA analysis in VH(bb) boosted regime
- MVA analysis significance improves 112% compared with cut based analysis
- Target date for publication: **Moriond 2023**

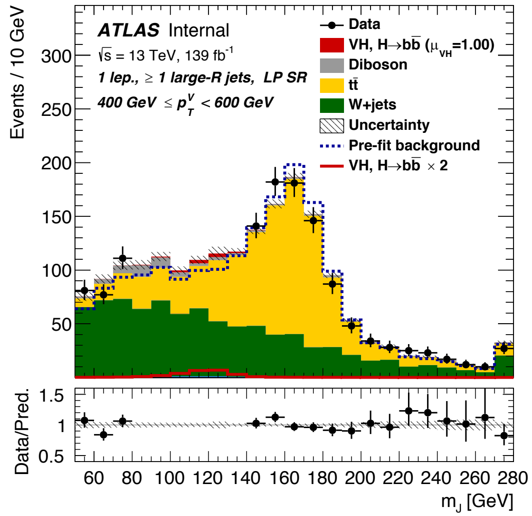
Back up

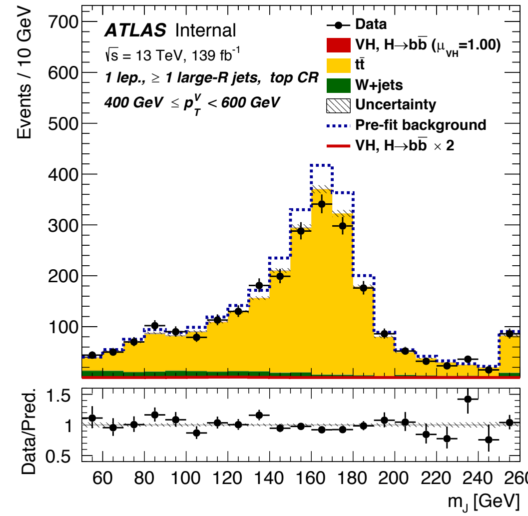
Post-fit plots: mJ vs. BDT (400_600GeV)

HP SR | LP SR | CR

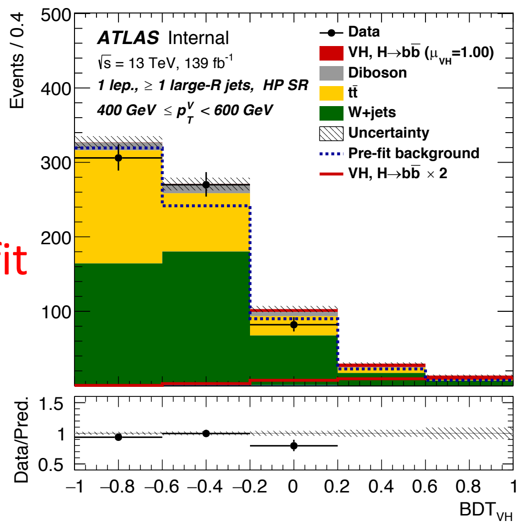
Cut-based fit

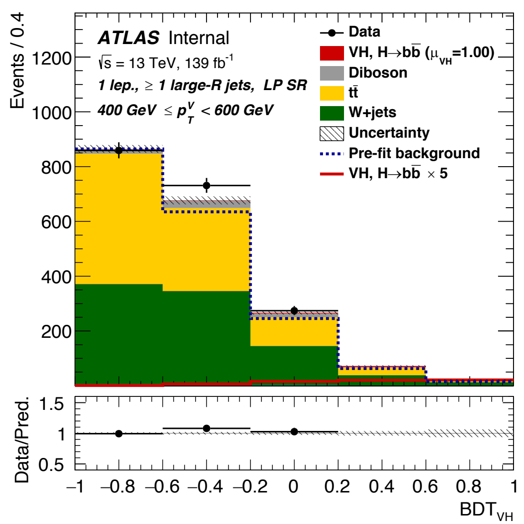


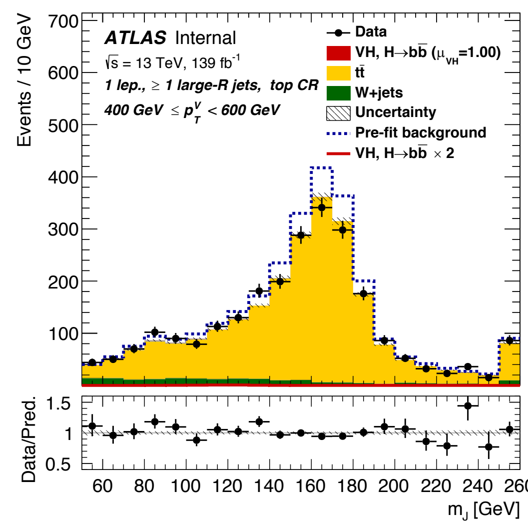




MVA fit





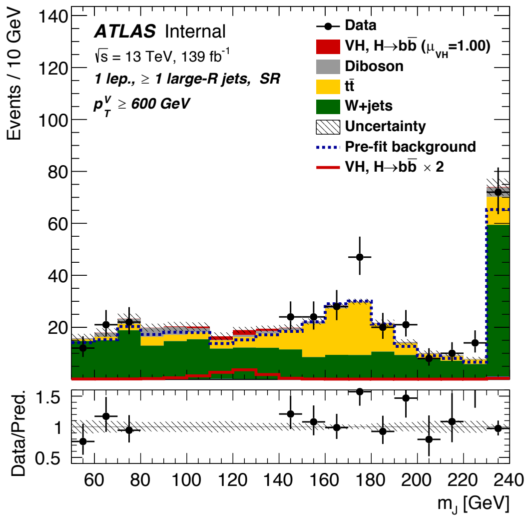


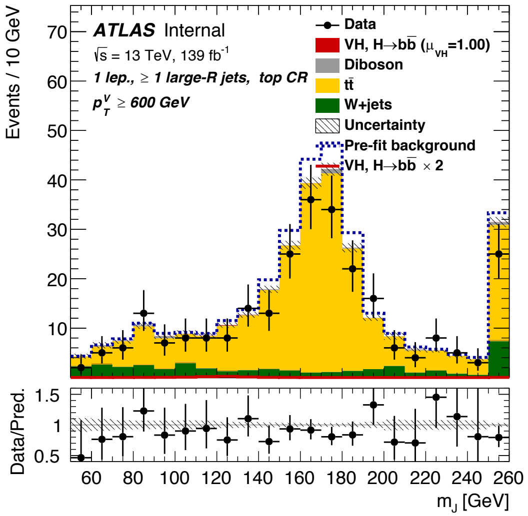
Post-fit plots: mJ vs. BDT ($p_T^V > 600$GeV)

SR

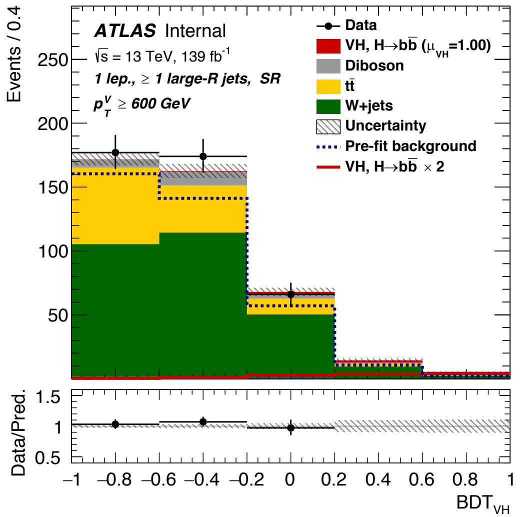
CR

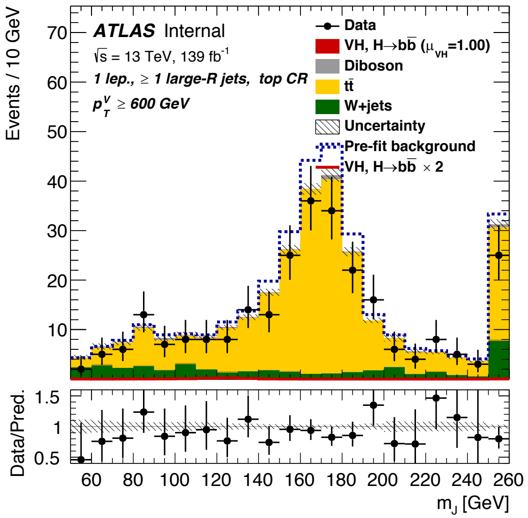
Cut-based fit





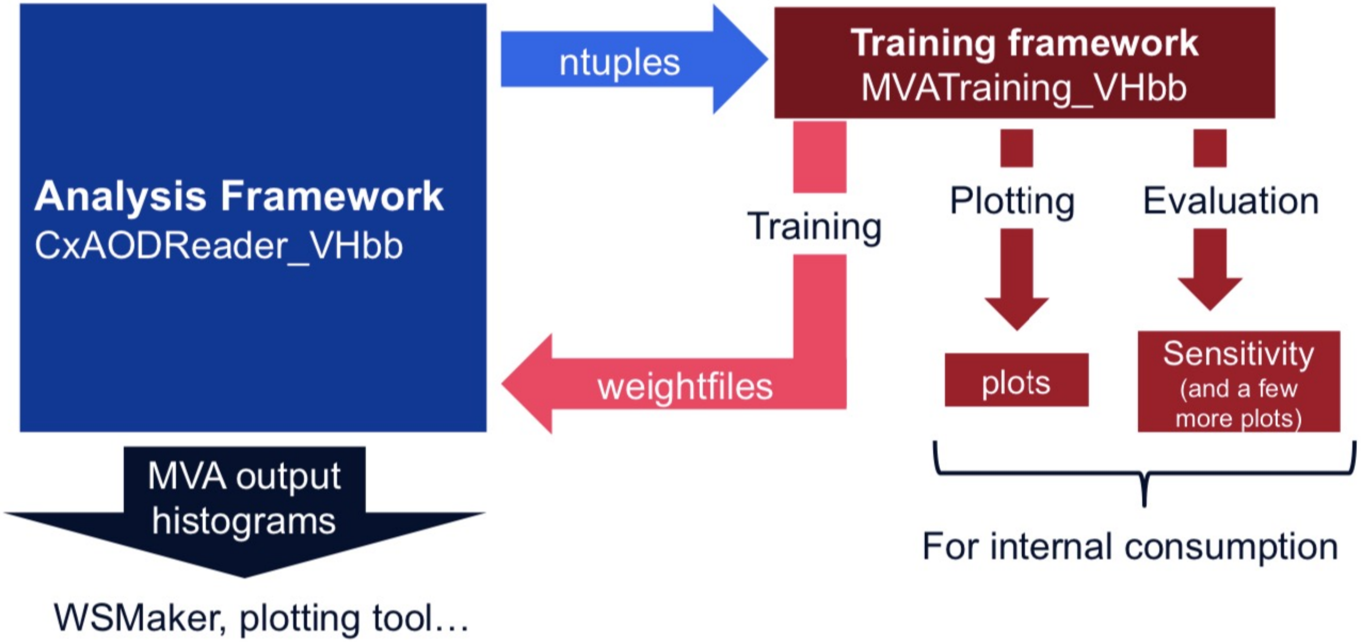
MVA fit





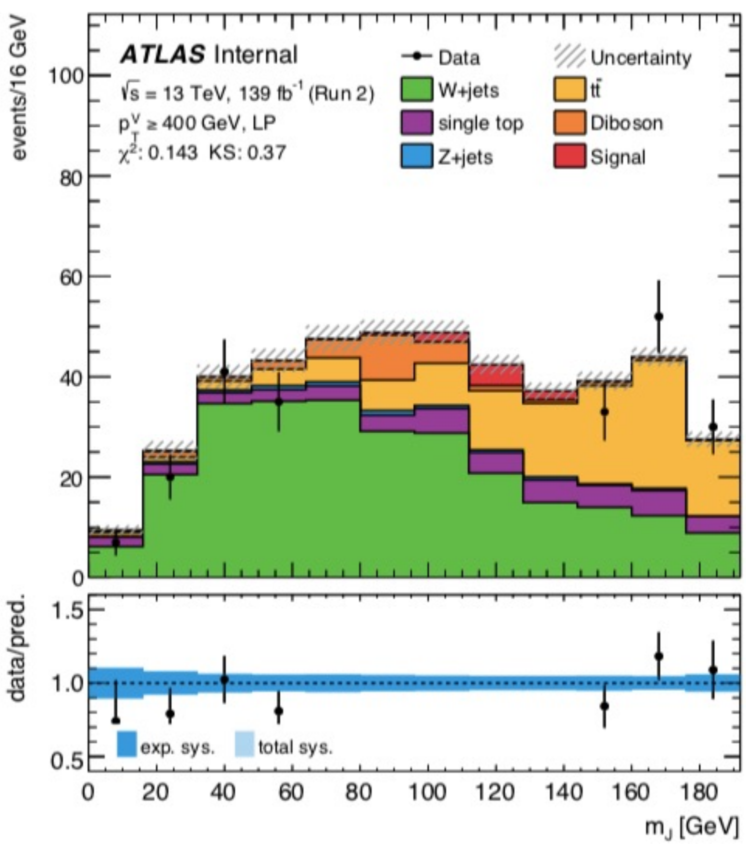
Multivariate analysis

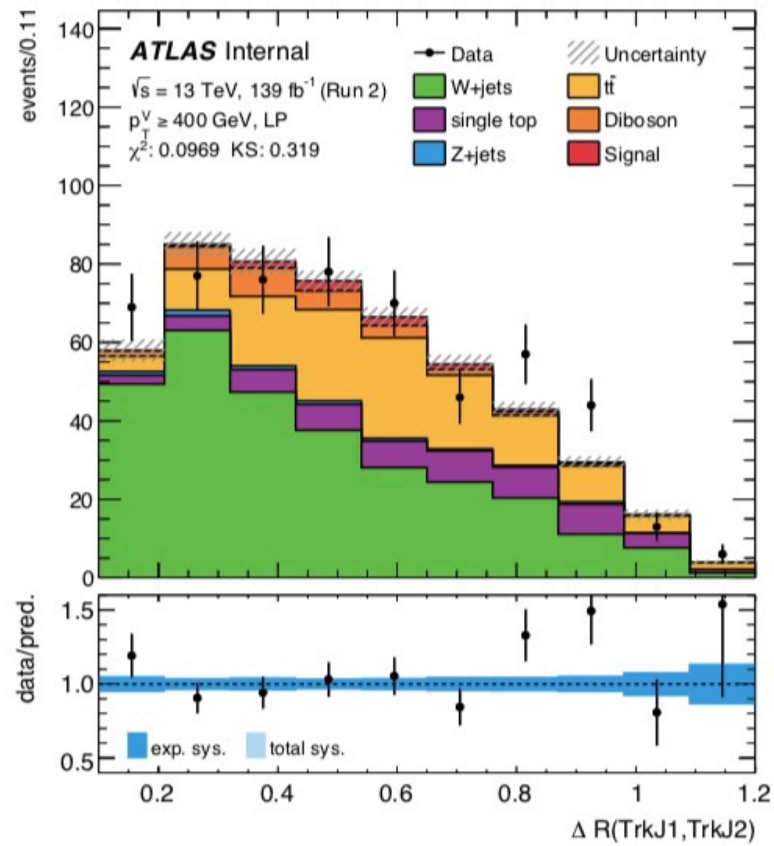
- **Multivariate analysis** used since long time in VH($b\bar{b}$) resolved analysis
  - **New: Introduce also in VH($b\bar{b}$) boosted analysis and in VH($c\bar{c}$) analysis**



- Baseline: Boosted Decision Tree (**BDT**)

Input variables data/MC (blinded)





mJ and **deltaR(b1,b2)** are the most important two variables in the training.

More plots here: link

Input variable importance ranking

Rank	Variable	Variable Importance
1	mJ	1.789e-01
2	deltaRbTrkJbTrkJ	1.447e-01
3	pTBTrkJ2	1.007e-01
4	deltaYVJ	9.876e-02
5	NAdditionalCaloJets	9.132e-02
6	pTBTrkJ1	8.874e-02
7	pTV	7.915e-02
8	absdeltaPhiVJ	6.894e-02
9	NMatchedTrackJetLeadFatJet	6.516e-02
10	bin_bTagBTrkJ2	4.597e-02
11	bin_bTagBTrkJ1	3.766e-02

MVA fit: Transformation parameter optimization

Transformation of the BDT output distribution: optimise the expected sensitivity and reduce the number of bins

$$Z = z_s \frac{n_s}{N_s} + z_b \frac{n_b}{N_b}$$

Default：
Z_s=3， Z_b=2

- N_s (N_b): total number of signal (background) events in the histogram;
- n_s (n_b): number of signal (background) events in interval $I[k, l]$ (interval between bin k and l);

The starting point for the transformation is the BDT output distribution with 500 equidistant bins between -1 and 1. The re-binning is then conducted using the following procedure:

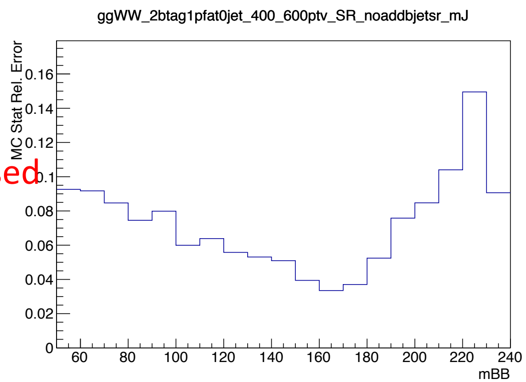
1. Starting from the last bin on the right of the original histogram, increase the range of the interval $I(k, last)$ by adding one after the other, the bins from the right to the left;
2. Calculate the value of Z at each step;
3. Once $Z(I[k_0, last]) > 1$, rebin all the bins in the interval $I(k_0, last)$ into a single bin;
4. Repeat steps 1-3, starting this time from the last bin on the right, not included in the previous remap (the new last is $k_0 - 1$), until k_0 in the first bin.

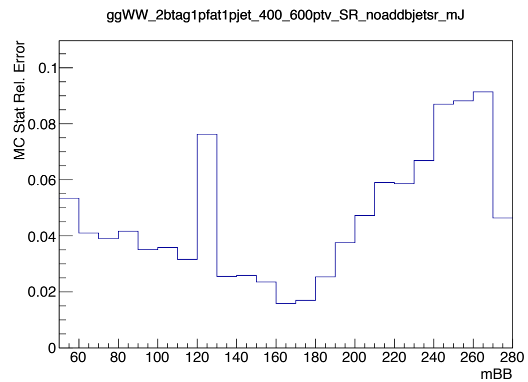
If the statistical uncertainty of the newly formed bin is larger than 20% step 2 is extended until the statistical uncertainty is below 20%.

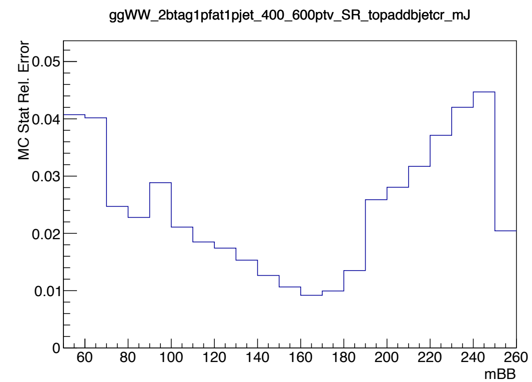
mJ fit vs. MVA fit: MC stat relative error: 400_600 GeV (1L)

HP SR | LP SR | CR

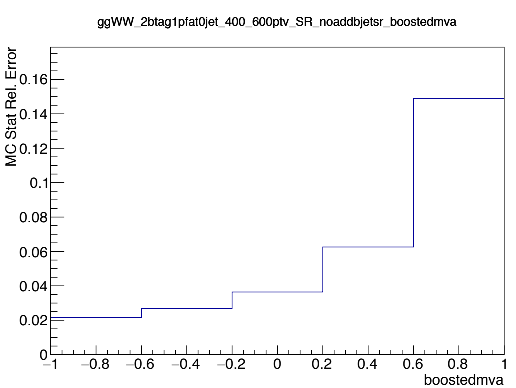
Cut-based fit

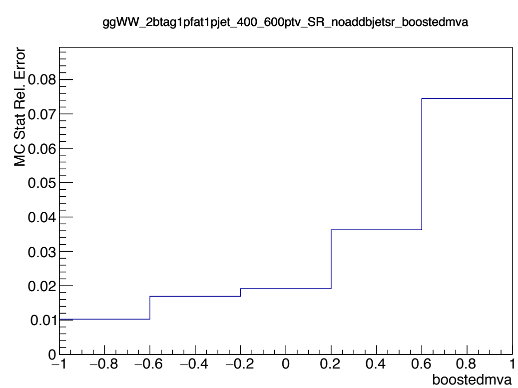


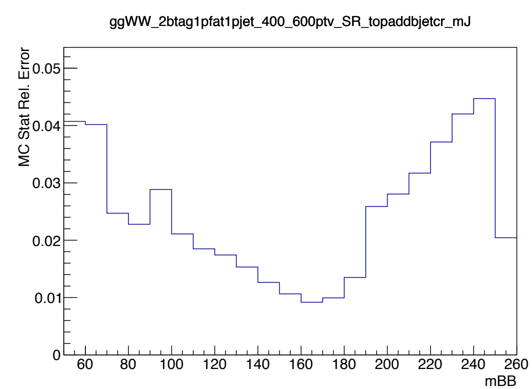




MVA





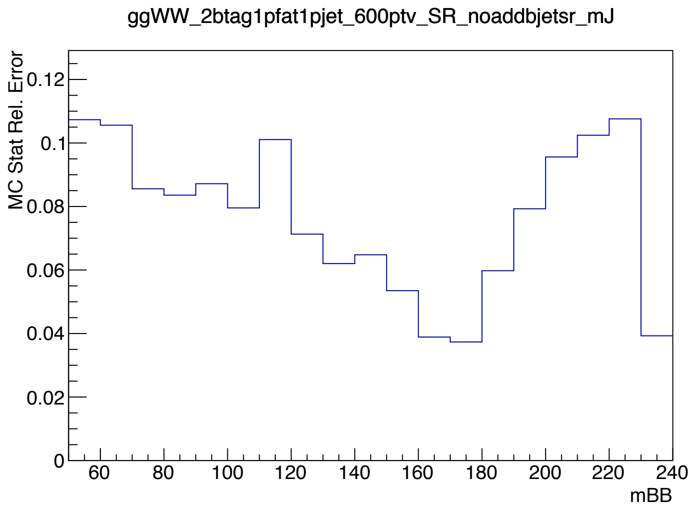


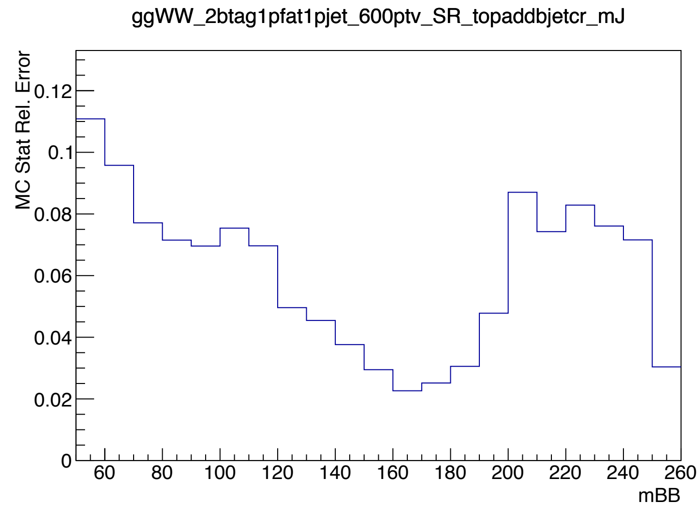
mJ fit vs. MVA fit: MC stat relative error: p_T^V>600 GeV (1L)

SR

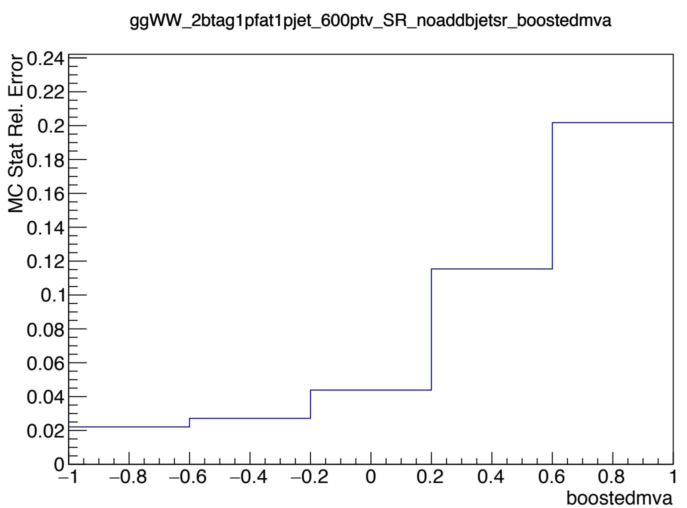
CR

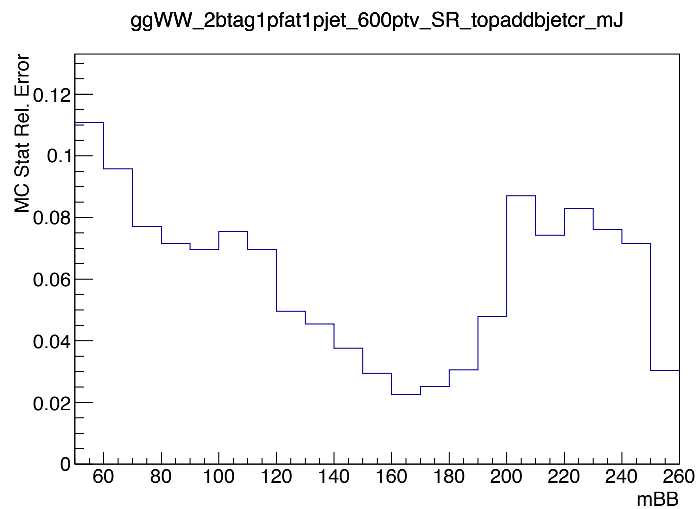
mJ fit





MVA fit



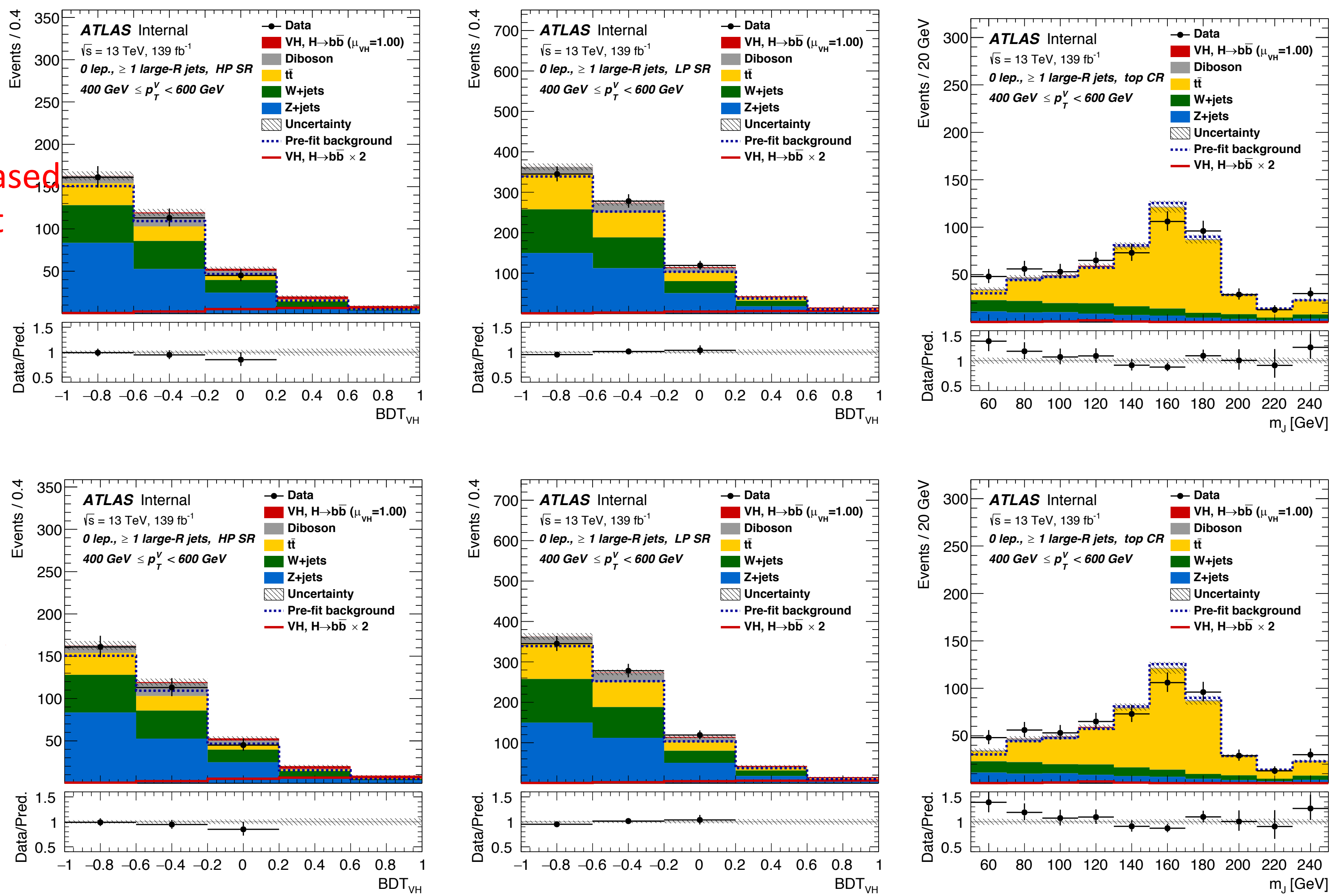


0L post-fit plots: mJ vs. BDT (400_600GeV)

HP SR | LP SR | CR

Cut-based fit

MVA

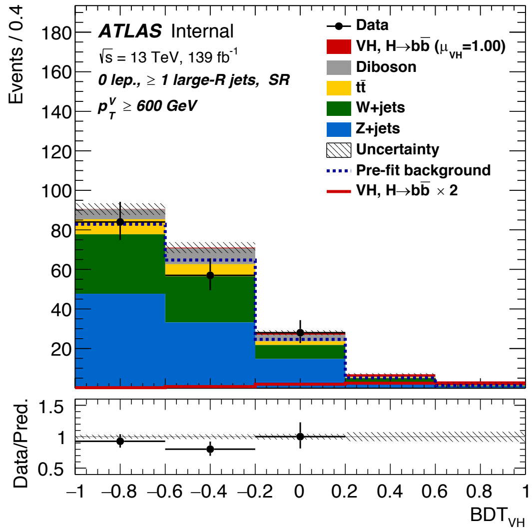


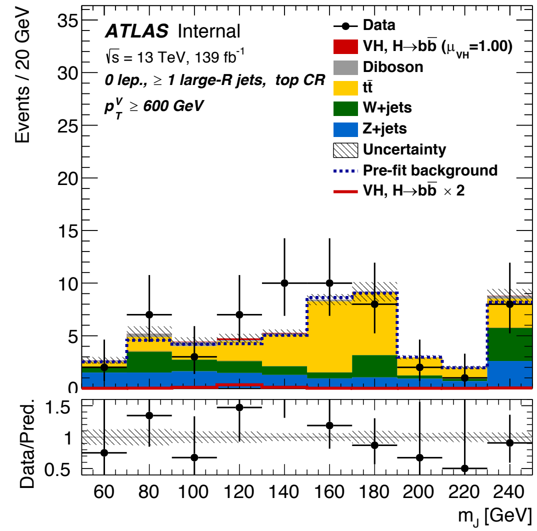
0L post-fit plots: mJ vs. BDT ($p_T^V > 600$GeV)

SR

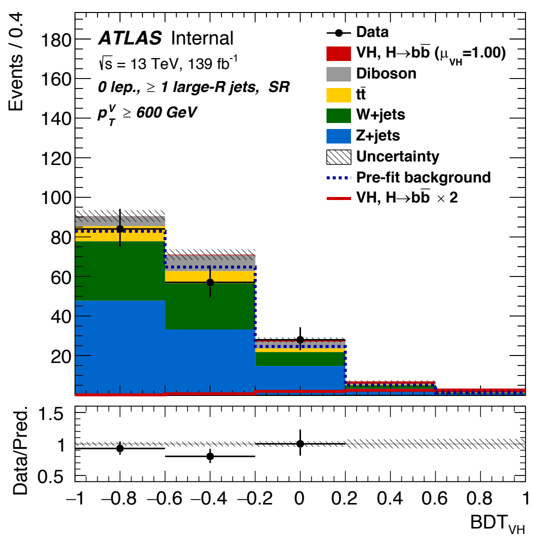
CR

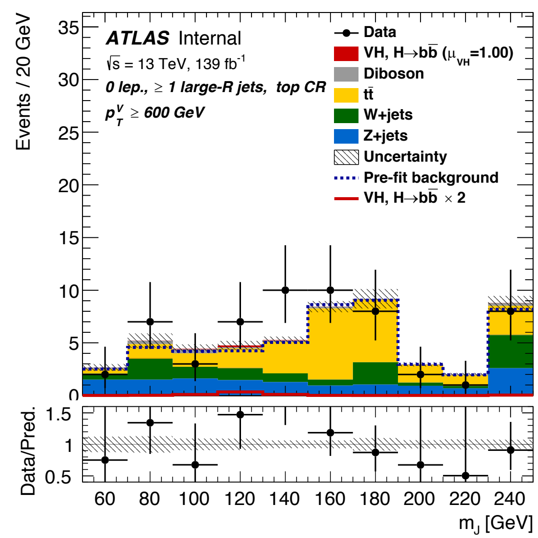
Cut-based fit





MVA fit

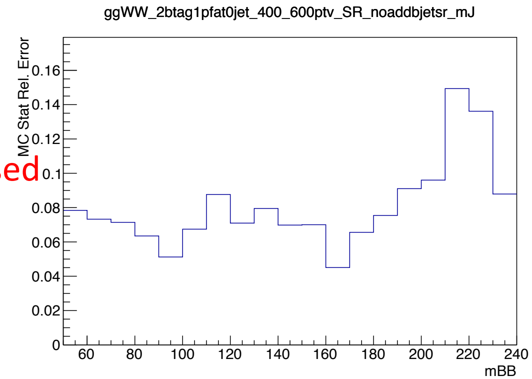


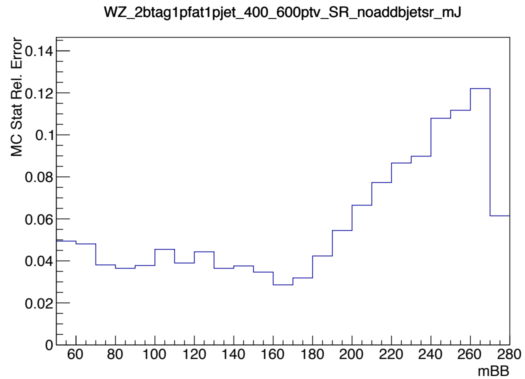


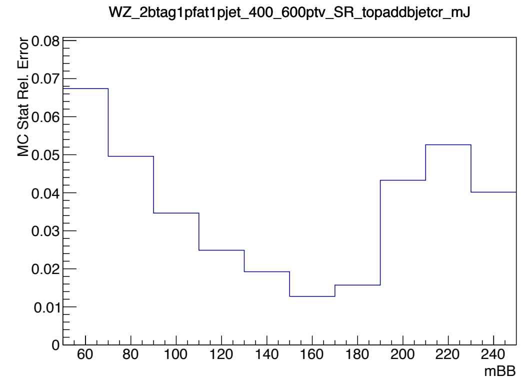
mJ fit vs. MVA fit: MC stat relative error: 400_600 GeV (0L)

HP SR | LP SR | CR

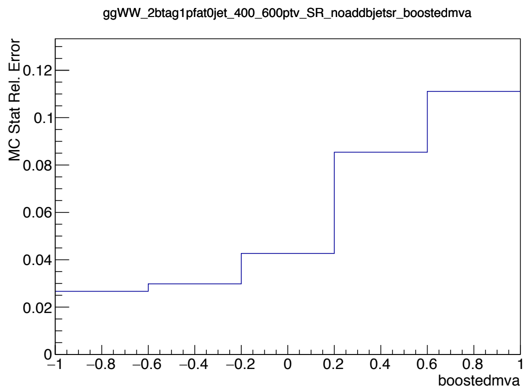
Cut-based fit

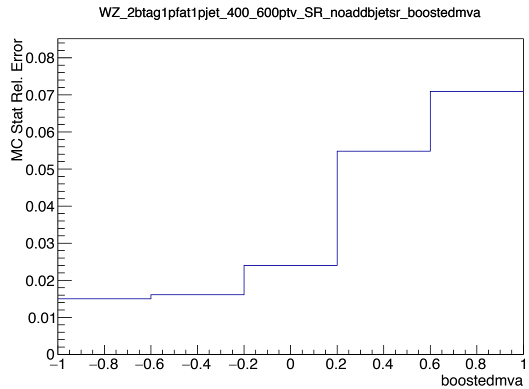


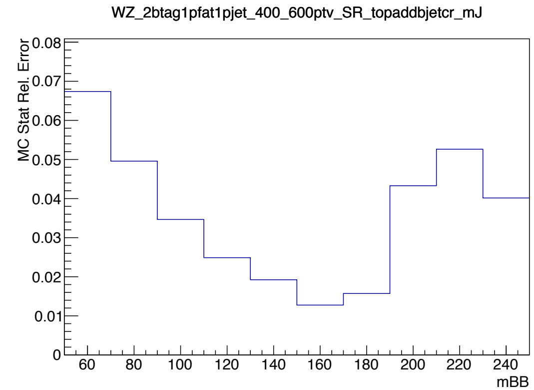




MVA





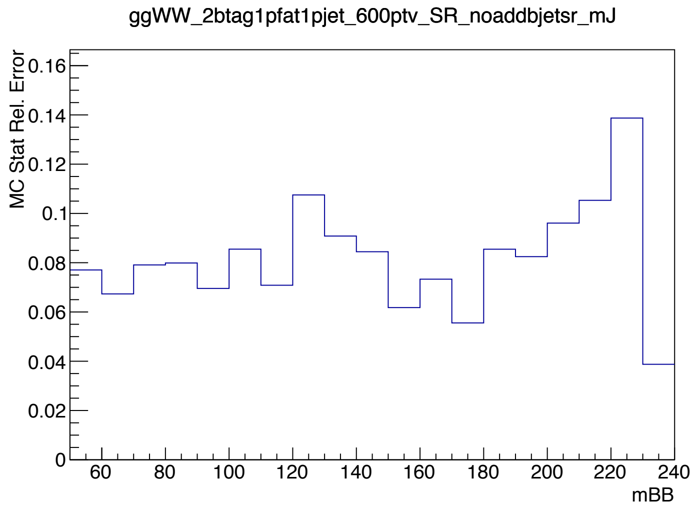


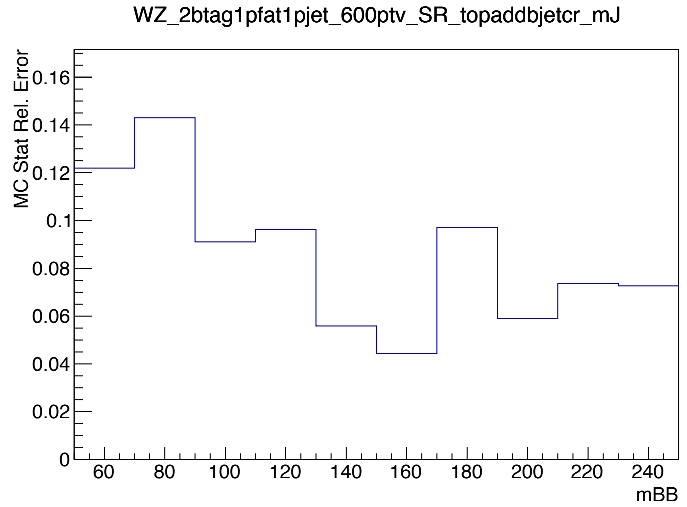
mJ fit vs. MVA fit: MC stat relative error: p_T^V>600 GeV (0L)

SR

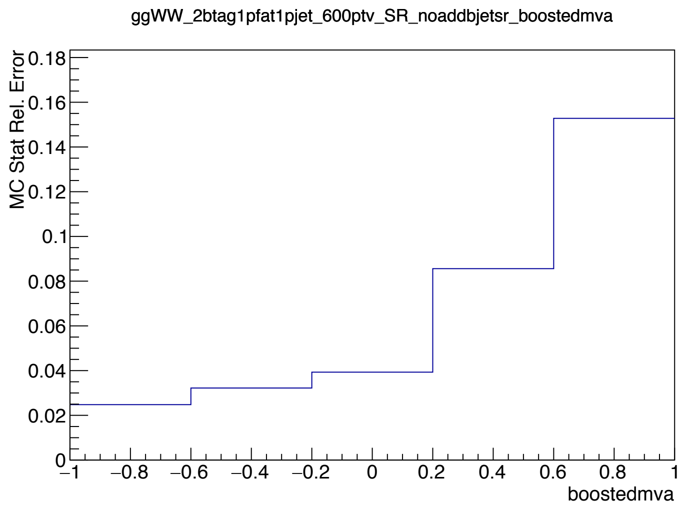
CR

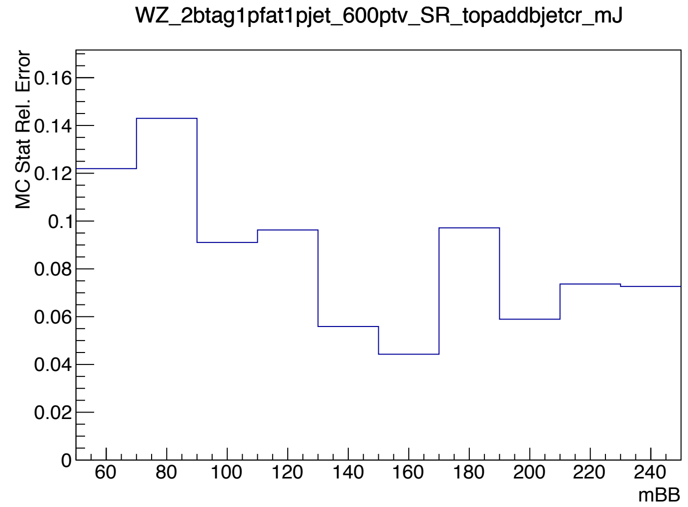
Cut-based fit





MVA fit



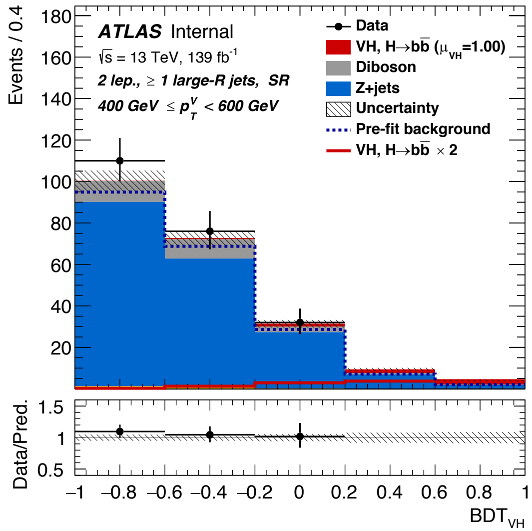


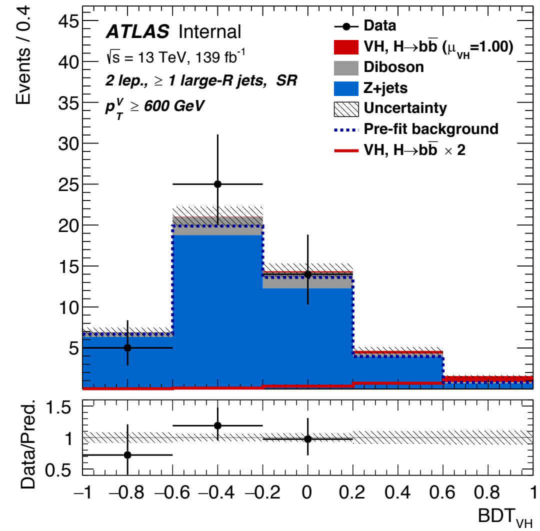
2L post-fit plots: mJ vs. BDT ($p_T^V > 600$GeV)

400_600 GeV

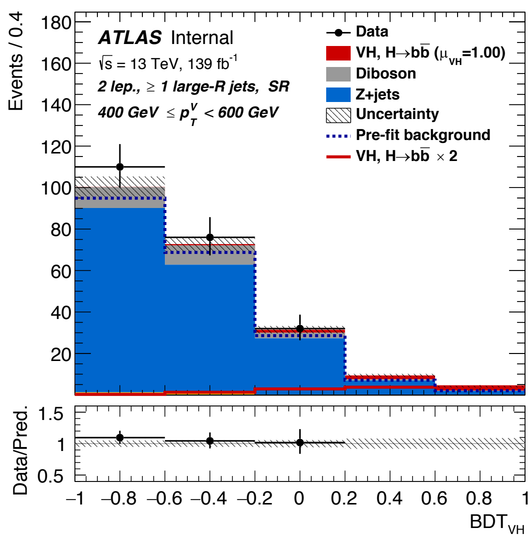
p_T^V>600 GeV

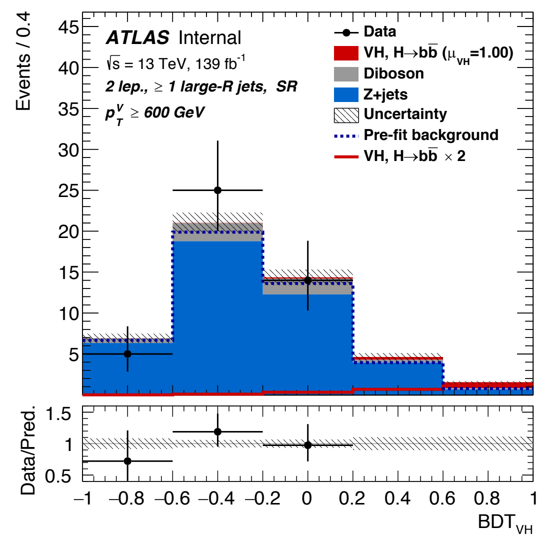
Cut-based fit





MVA fit



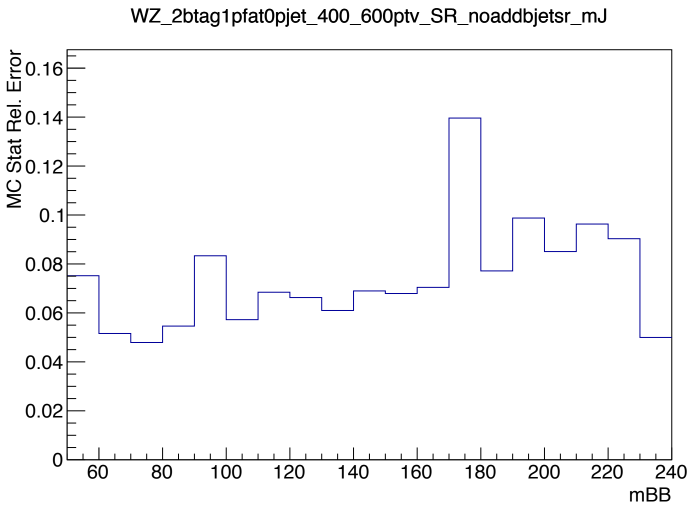


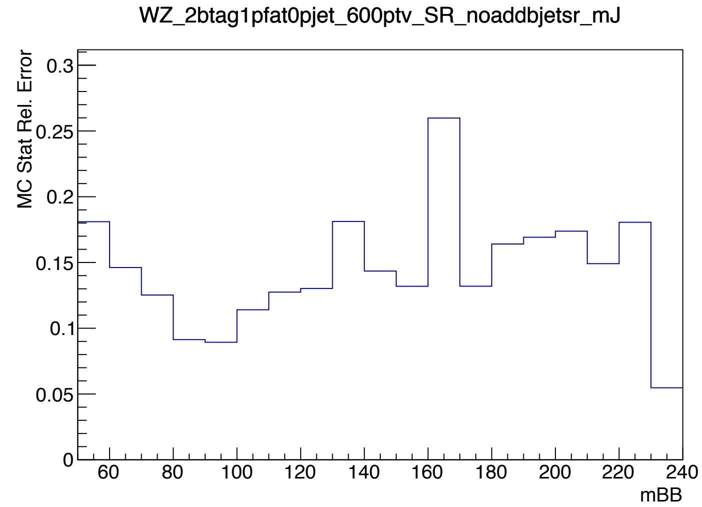
mJ fit vs. MVA fit: MC stat relative error (2L)

400_600 GeV

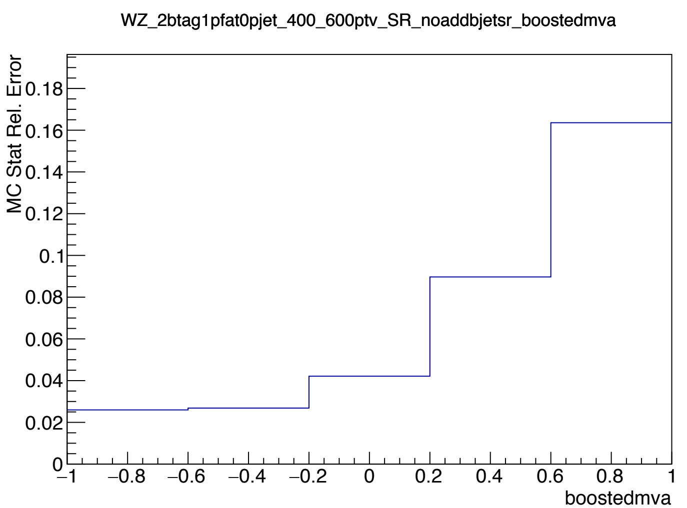
p_T^V>600 GeV

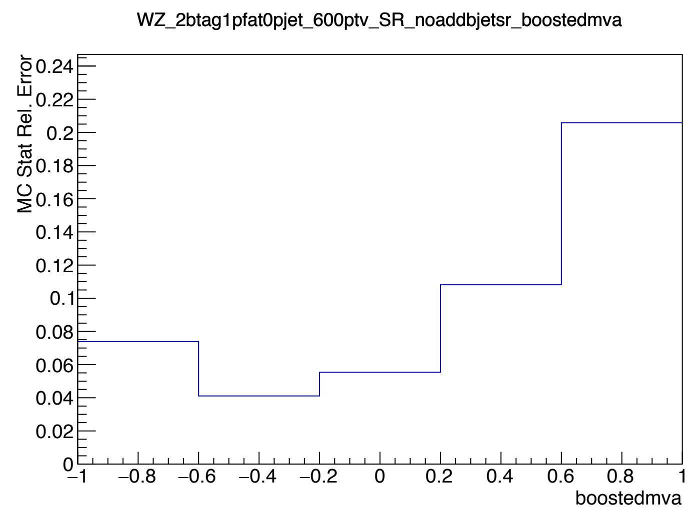
Cut-based fit





MVA fit



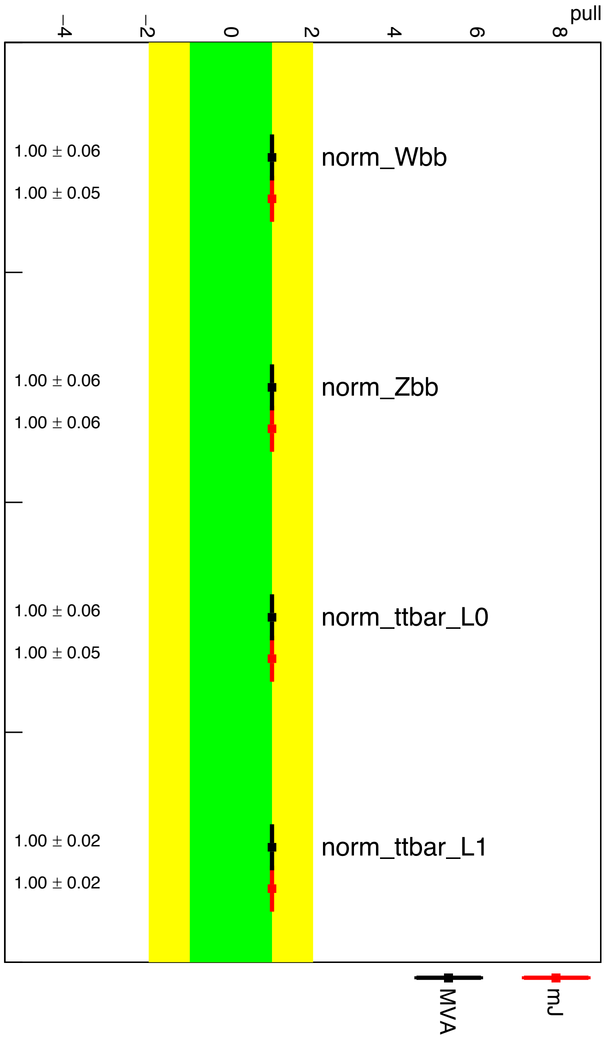


Breakdown

Set of nuisance parameters	Impact on error					
Total	+	0.628	-	0.615	+-	0.621
DataStat	+	0.601	-	0.585	+-	0.593
FullSyst	+	0.183	-	0.189	+-	0.186
Data stat only	+	0.505	-	0.490	+-	0.498
Floating normalizations	+	0.338	-	0.336	+-	0.337
Modelling: VH	+	0.000	-	0.000	+-	0.000
Modelling: Background	+	0.183	-	0.189	+-	0.186
Multi Jet	+	0.000	-	0.000	+-	0.000
Modelling: single top	+	0.000	-	0.000	+-	0.000
Modelling: ttbar	+	0.000	-	0.000	+-	0.000
Modelling: W+jets	+	0.000	-	0.000	+-	0.000
Modelling: Z+jets	+	0.000	-	0.000	+-	0.000
Modelling: Diboson	+	0.000	-	0.000	+-	0.000
MC stat	+	0.183	-	0.189	+-	0.186
Experimental Syst	+	0.000	-	0.000	+-	0.000
Detector: lepton	+	0.000	-	0.000	+-	0.000
Detector: MET	+	0.000	-	0.000	+-	0.000
Detector: JET	+	0.000	-	0.000	+-	0.000
Detector: FATJET	+	0.000	-	0.000	+-	0.000
Detector: FTAG (b-jet)	+	0.000	-	0.000	+-	0.000
Detector: FTAG (c-jet)	+	0.000	-	0.000	+-	0.000
Detector: FTAG (l-jet)	+	0.000	-	0.000	+-	0.000
Detector: FTAG (extrap)	+	0.000	-	0.000	+-	0.000
Detector: PU	+	0.000	-	0.000	+-	0.000
Lumi	+	0.000	-	0.000	+-	0.000

mJ fit vs. MVA fit: Floating normalization

Asimov fit



Conditional data fit