Chapter 5
Event Selection and Categorization

This chapter describes the selection of the $VH(b\bar{b})$ events in the boosted analysis. The first part discusses the criteria used to identify physics objects. The second part explains the selection of both the Higgs boson and vector boson candidates. In the final part all the details of the event categorization are illustrated.

5.1 Object Identification

The definition of the physics objects as well as the strategy of the *overlap removal* are provided in Sect. 4. In the following the list of all the objects, with their specific requirements, used in the $VH(b\bar{b})$ *boosted* analysis is presented. The description of the $E_{\mathrm{T}}^{\mathrm{miss}}$ object is skipped since the analysis does not apply any specific cut. In addition the b-tagging tool used to reconstruct the Higgs decay products is already shown in Sect. 4.3.7.

Leptons

The leptons used in the $VH(b\bar{b})$ analysis are electrons and muons.Two type of leptons are used in the analysis: *loose* leptons and *signal* leptons. The *loose* leptons are used to define the three main channels requiring exactly zero, one and two leptons. The *signal* leptons are a subset of the *loose* leptons with tighter identification and isolation requirements. In 1-lepton and 2-lepton channels at least one *signal* lepton is required to suppress the multi-jet background.

Loose electrons are required to have $p_{\mathrm{T}} > 7$ GeV and $|\eta| < 2.47$, to pass *Loose* identification requirement and *Fixed-Cut Loose* isolation (see Sect. 4.2). They have also to satisfy $|d_0/\sigma(d_0)| < 5$ and $|z_0 \sin(\theta)| < 0.5$ mm, where d_0 and z_0 are the transverse and longitudinal impact parameters defined relative to the primary vertex position and $\sigma(d_0)$ is the d_0 uncertainty. *Signal* electron in the 2-lepton channel are *loose* electron with a high p_{T} cut, $p_{\mathrm{T}} > 27$ GeV. Differently *signal* electrons in 1-lepton channel have tighter identification and isolation requirements with respect to

G. Di Gregorio, *A Beauty-ful Boson*, Springer Theses,
https://doi.org/10.1007/978-3-031-20013-7_5

Table 5.1 Electron selection requirements used in the $VH(b\bar{b})$ *boosted* analysis

Electron Selection	p_T	$\|\eta\|$	ID	$\|d_0/\sigma(d_0)\|$	$\|z_0 \sin(\theta)\|$	Isolation
Loose electron	>7 GeV	<2.47	*Loose*	<5	<0.5 mm	*Fixed-Cut Loose*
Signal electron in 2-lepton channel	>27 GeV	<2.47	*Loose*	<5	<0.5 mm	*Fixed-Cut Loose*
Signal electron in 1-lepton channel	>27 GeV	<2.47	*Tight*	<5	<0.5 mm	*Fixed-Cut High Pt Calo Only*

Table 5.2 Muon selection requirements used in the $VH(b\bar{b})$ *boosted* analysis.

Muon selection	p_T	$\|\eta\|$	ID	$\|d_0/\sigma(d_0)\|$	$\|z_0 \sin(\theta)\|$	Isolation
Loose muon	>7 GeV	<2.7	*Loose*	<3	<0.5 mm	*Fixed-Cut Loose*
Signal muon in 2-lepton channel	>27 GeV	<2.5	*Loose*	<3	<0.5 mm	*Fixed-Cut Loose*
Signal muon in 1-lepton channel	>25 GeV	<2.5	*Medium*	<3	<0.5 mm	*Fixed-Cut High Pt Track Only*

the *signal* electrons in 2-lepton channel in order to reject the multi-jet background. Table 5.1 summarizes the electron definitions used in the analysis.

Loose muons are required to have $p_T > 7$ GeV, $|\eta| < 2.7$, $|d_0/\sigma(d_0)| < 3$ and $|z_0 \sin(\theta)| < 0.5$. They have to pass *Loose* identification and *Fixed-Cut Loose* isolation requirements. *Signal* muons in the 2-lepton channel have the same requirements on the transverse and longitudinal impact parameters, together with the isolation and identification criteria of the *loose* muons. These muons must have $p_T > 27$ GeV and $|\eta| < 2.5$. *Signal* muons in 1-lepton channel have $p_T > 25$ GeV, $|\eta| < 2.5$, $|d_0/\sigma(d_0)| < 3$ and $|z_0 \sin(\theta)| < 0.5$. In addition they satisfy the *Medium* identification and the *Fixed-Cut High Pt Track Only* isolation criteria. Table 5.2 summarizes the muon definitions used in the analysis.

Jets

In the $VH(b\bar{b})$ analysis three different anti-k_t jet collections have been used: large-R jets (R=1.0), small-R jet ($R = 0.4$) and VR track-jets. The first two types of jets are reconstructed starting from the energy deposition in the calorimeter, while the track-jets are reconstructed from inner detector tracks.

The large-R jets are used to reconstruct the Higgs boson candidate in the high energy regime. Only large-R jets with $p_{\mathrm{T}} > 250$ GeV and $|\eta| < 2.0$ are considered. Due to the rule of thumb $\Delta R(jet_1, jet_2) \sim 2m/p_T$ where m and p_T are the mass and the transverse momentum of the large-R jet and ΔR is the angular separation between the decay products, the transverse momentum cut marks the point where the two b-quark jets are geometrically separated by $\Delta R = 1.0$.

To exploit the optimal large-R jet mass resolution over the full p_{T} range, the combined jet mass definition is used. A cut on the large-R jets mass m_J is applied, $m_J > 50$ GeV. The ATLAS Collaboration does not support the calibration in the low mass region due to the difference between data and simulation.

In the analysis two corrections are applied to the large-R jet to better set the scale and to improve the resolution of their energy and mass measurement. The large-R jets are first corrected to take into account the presence of muons from the b- or c-hadron decays. The muons are not included in the jet resulting in losses in the large-R jet energy. Therefore, a correction called *muon-in-jet*, which adds back the four-momentum of the muon associated to a large-R jet, is applied. When more than one muon is found, the one closest to the VR track-jet ghost-associated to the large-R jets is chosen.

The second correction called *Kinematic Fit* (KF) is applied only in the 2-lepton channel to improve the large-R jet energy resolution. The aim of this correction is to constrain the $ZH \rightarrow llb\bar{b}$ system to be balanced in the transverse plain. The KF includes also a constraint on the dilepton mass to the Z boson mass. This correction uses the energy resolution of the electrons and muons, which is typically 1%, to improve the large-R jet energy resolution, which is typically 10%.

To evaluate the improvement in resolution, the large-R jet mass m_J distributions before and after the corrections are fit with a Bukin function [1]. The resolution values σ correspond to the width of the fitted function. For the event selection of this analysis, the large-R jet mass resolution improves by 5% to 10% after the muon-in-jet correction, depending on the lepton channel. The KF brings an additional improvement in the 2-lepton channels of up to 40%. Figure 5.1 shows a comparison of the large-R jet mass when the additional corrections are applied to the jet energy scale in the 2-lepton channel.

In the $VH(b\bar{b})$ *boosted* analysis the small-R jets are used to build the $E_{\mathrm{T}}^{\mathrm{miss}}$, for the multi-jet estimate and for the event categorization. For the event categorisation they are required to have $p_{\mathrm{T}} > 30$ GeV and $|\eta| < 4.5$. To reduce the number of small-R jets originating from the pile-up interactions, the small-R jets are required to pass the JVT requirement if they are in the range $p_{\mathrm{T}} < 120$ GeV and $|\eta| < 2.5$.

In the analysis the VR track-jets are used as input of the b-tagging algorithm to correctly identify jets originating from the $H \rightarrow b\bar{b}$ decay. They are preferred over the standard calorimeter jets due to their higher efficiency to resolve objects in the high energy regions. Only central ($|\eta| < 2.5$) VR track-jets with $p_{\mathrm{T}} > 10$ GeV and with at least two tracks are considered.

Table 5.3 summarizes the jet collections and the respective kinematics cuts used in the analysis.

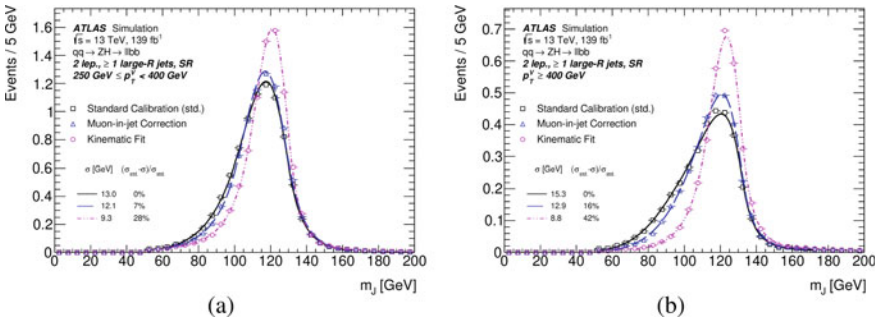

Fig. 5.1 Comparison of the large-R mass distributions when additional corrections are applied to the jet energy scale for the signal in the 2-lepton channel, 250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV (a) and $p_{\mathrm{T}}^{V} \geq$ 400 GeV regions for the dominant $qq \rightarrow ZH$ contribution. The distributions are fitted with a Bukin function [1] and the resolution values, σ, correspond to the width of the fitted function [2]

Table 5.3 Jet collections and requirements

Jet collection	p_{T}	η
Large-R jets	>250 GeV	$\lvert\eta\rvert < 2.5$
Small-R jets	>30 GeV	$\lvert\eta\rvert < 4.5$
VR track-jets	>10 GeV	$\lvert\eta\rvert < 2.5$

5.2 Event Selections

Events are categorised into 0-, 1- and 2-lepton channels depending on the number of charged leptons (electrons or muons) to target the $ZH \rightarrow \nu\bar{\nu}b\bar{b}$, $WH \rightarrow l\nu b\bar{b}$ and $ZH \rightarrow llb\bar{b}$ signature, respectively.

The online selection of the 0-lepton channel relies on the $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger with increasing threshold, from 70 GeV to 110 GeV, for the increasing of luminosity during the Run 2. In the 1-lepton electron sub-channel events are selected by the a low-p_{T} threshold unprescaled single electron trigger. The same trigger as in the 0-lepton channel is used in the 1-lepton muon sub-channel since muons are not included in the online $E_{\mathrm{T}}^{\mathrm{miss}}$ calculation. In the 2-lepton channel, the same trigger strategy as in the 1-lepton channel is adopted.

In the 0- and 2-lepton channels, the transverse momentum of the Z boson is reconstructed as $E_{\mathrm{T}}^{\mathrm{miss}}$ and as the transverse momentum of the two leptons system, respectively. In the 1-lepton channel, the transverse momentum of the W boson is reconstructed as the vectorial sum of the $E_{\mathrm{T}}^{\mathrm{miss}}$ and the lepton transverse momentum. In all the three lepton channels, the events are required to have the transverse momentum of the vector boson p_{T}^{V} grater than 250 GeV ($p_{\mathrm{T}}^{V} \geq 250$ GeV). Events are split in two bins of p_{T}^{V} (250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV, $p_{\mathrm{T}}^{V} \geq$ 400 GeV) to improve the analysis sensitivity. Events are also categorized in *signal regions* (SRs) and *control regions* (CRs). While the SR is the one expected to contain the larger fraction of signal, the

Table 5.4 Event selection for the three channels of the $VH(b\bar{b})$ *boosted* 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 *loose* lepton	1 *signal* lepton No second *loose* lepton		≥ 1 *signal* lepton 2 *loose* leptons	
$E_{\mathrm{T}}^{\mathrm{miss}}$	≥ 250 GeV	> 50GeV	–	–	
p_{T}^{V}	$p_{\mathrm{T}}^{V} \geq 250$ 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				
$\min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{jets})]$	$> 30°$	–			
$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},H_{\mathrm{cand}})$	$> 120°$	–			
$\Delta\phi\,(E_{\mathrm{T}}^{\mathrm{miss}},\, E_{\mathrm{T,trk}}^{\mathrm{miss}})$	$< 90°$	–			
$\lvert\Delta y(V,H)\rvert$	–	$\lvert\Delta y(V,H)\rvert < 1.4$			
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	

CR is a background enriched region designed to evaluate the contribution and shape of one of the main backgrounds. In the 0- and 1-lepton channels, the SRs are defined by vetoing b-tagged track-jets outside the Higgs candidate jet. More details on the event categorization are reported in Sect. 5.3.

In the following paragraphs, a description of the Higgs candidate and channel-specific selections are provided. Table 5.4 summarizes the selection applied in each of the three channels.

5.2.1 Higgs Candidate Reconstruction and Selection

All the events in the three lepton channels are required to have at least one large-R jet. When more than one large-R jet is found in the event, the one with the highest transverse momentum is used as Higgs candidate jet. Events are also required to have at least two track-jets ghost matched to the Higgs candidate. All the track-jets in the events are required to pass the VR jet *overlap removal* procedure to avoid pathological cases to the b-tagging algorithm in which the two axes of the jets are reconstructed inside both cones (see Sect. 4.5). The b-tagging algorithm used in the analysis is the MV2 algorithm with the b-tagging efficiency of 70%.[1] The b-tagging algorithm is applied to the two VR track-jets with the highest transverse momentum

[1] The 70% efficiency of the b-tagging algorithm corresponds to a c-jet and light-flavour jet rejections equal to 9 and 304, respectively.

matched to the Higgs candidate and the events are categorized according to the number of b-tagged matched track-jets. Events with no b-tagged track-jets, or with exactly one b-tagged track-jet, or with exactly two b-tagged track-jets, compose the 0-tag, 1-tag and 2-tag categories, respectively. The analysis is performed in the 2-tag region because it is the region with the largest sensitivity in which the two b-tagged track-jets represents the two Higgs decay products. The 0-tag and 1-tag regions are used only in the 1-lepton channel to extract the QCD multi-jet contribution (see Sect. 6.3.7). The mass of the Higgs candidate is reconstructed using the invariant mass of the large-R jets m_J and the $m_J > 50$ GeV requirement is adopted. This request is applied before any large-R jet mass corrections (*muon-in-jet* correction and Kinematic Fit).

5.2.2 0-Lepton Channel Selection

In the 0-lepton channel a specific selection is defined to isolate events containing a Z boson decaying into a pair of neutrinos, in addition to the Higgs boson selection. Events passing the online selection are required to have no *loose* leptons and $E_{\mathrm{T}}^{\mathrm{miss}} \geq 250$ GeV.

In this channel the QCD multi-jet (MJ) events are a relevant background. The jet energy mis-measurements could generate fake $E_{\mathrm{T}}^{\mathrm{miss}}$ which tends to be aligned with the mis-measured jet. It is not possible to use the MC sample to estimate the MJ contributions because the simulated events do not populate the analysis phase space with enough statistics. The MC multi-jet samples are only used as cross-check once the MJ contribution is estimated with a data driven method. Multi-jet events are suppressed after applying angular cuts in the separation between small-R jets, $E_{\mathrm{T}}^{\mathrm{miss}}$, $E_{\mathrm{T,trk}}^{\mathrm{miss}}$ and Higgs candidate jet ($H_{\mathrm{cand.}}$):

- $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, H_{\mathrm{cand.}}) > 120°$;
- $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, E_{\mathrm{T,trk}}^{\mathrm{miss}}) < 90°$;
- $\min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \text{small-}R \text{ jets})] > 30°$;

where $\Delta\phi(a, b)$ indicates the distance in the azimuthal angle between the two objects a and b. In the $\min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \text{small-}R \text{ jets})]$ calculation, only small-R jets with $p_{\mathrm{T}} > 70$ GeV geometrically outside[2] the Higgs candidate jet are considered. A detailed explanation on the $\min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \text{small-}R \text{ jets})]$ cut and of the p_{T} threshold of the small-R jets is reported in Sect. 6.3.7. The values of the angular cuts are tuned in a way that the remaining fraction of MJ contamination is of the order of 1% of the signal.

The efficiency of each selection cut applied in 0-lepton channel has been studied using simulated signal samples. The efficiency, ϵ, is evaluated as the ratio of the number of events that pass a cut, N_{cut}, over the total number of generated events,

[2] Since the radius parameter of the Higgs candidate jet is $R = 1.0$, only small-R jets with an angular separation grater that 1.0 with the Higgs candidate are considered ($\Delta R(\text{small-}R \text{ jets}, H_{\mathrm{cand.}}) > 1.0$).

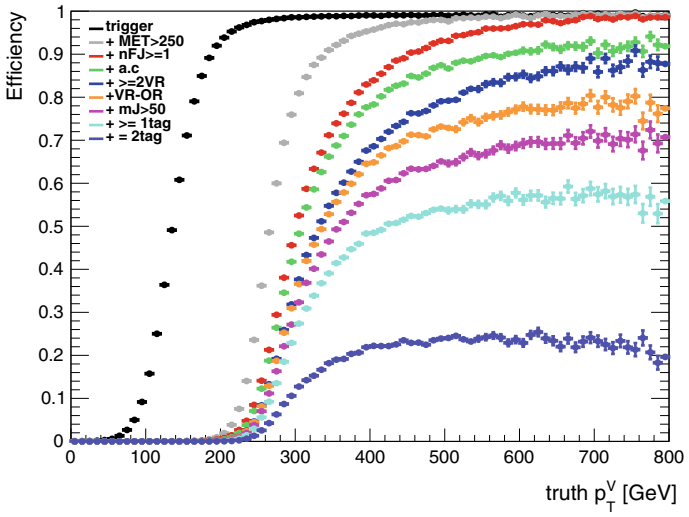

Fig. 5.2 Efficiency breakdown in 0-lepton channel. The cuts have been applied in cascade so in each step there is an additional request. Each curve with different colors shows the effect of cut cascade as described in the text

N_{all} ($\epsilon = N_{cut}/N_{all}$). Figure 5.2 shows the efficiency as a function of p_{T}^{V} evaluated at generator level (truth p_{T}^{V}). To study the efficiency the following cuts have been applied in cascade so in each step there is an additional request[3]:

- *trigger cut*: events that pass the trigger selection are required (black dots in Fig. 5.2);
- $E_{\mathrm{T}}^{\mathrm{miss}}$ *cut*: events with $E_{\mathrm{T}}^{\mathrm{miss}} \geq 250$ GeV are required (gray dots in Fig. 5.2);
- large-R *jet cut*: events with at least one large-R jet are required (red dots in Fig. 5.2);
- *anti-QCD cut*: events that pass the anti-QCD angular cuts are required (green dots in Fig. 5.2);
- *VR track-jets cut*: events with at least two VR track-jets ghost associated to the Higgs candidate jet are required (blue dots in Fig. 5.2);
- *VR overlap removal cut*: events that pass the VR overlap removal procedure are required (orange dots in Fig. 5.2);
- m_J *cut*: events with $m_J > 50$ GeV are required (magenta dots in Fig. 5.2);
- *1-tag cut*: events with at least one b-tagged track-jet are required (light blue dots in Fig. 5.2);
- *2-tag cut*: events with exactly two b-tagged track-jets are required (violet dots in Fig. 5.2).

[3] This means that the gray dots in Fig. 5.2 represent all the events that pass the trigger and $E_{\mathrm{T}}^{\mathrm{miss}}$ cuts divided by the total number of generated event.

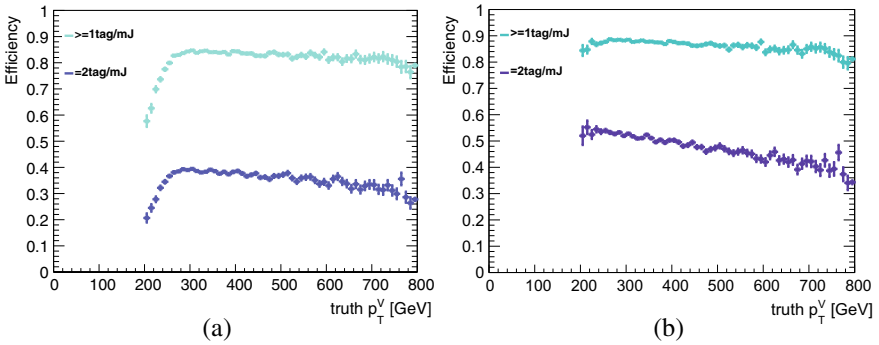

Fig. 5.3 **a** *b*-tagging efficiency in the 1-tag (light blue dots) and 2-tag (violet dots) region as a function of the truth transverse momentum of the vector boson. **b** *b*-tagging efficiency in the 1-tag (light blue dots) and 2-tag (violet dots) region as a function of the truth transverse momentum of the vector boson requiring also the VR track-jets reconstruct correctly the *b*-hadron

All the efficiencies have a turn on and then a flat trend. After applying all the cuts, only 20% of the signal events pass the 0-lepton event selection. As expected, the cut discarding most events is the *b*-tagging requirement. To understand better the impact of a single cut, it is needed to normalize the number of events that pass a specific cut to the number of events that pass the previous cut. In the following only the impact of the *b*-tagging algorithm is discussed. For completeness, all the other plots are reported in Appendix C. Figure 5.3a shows the number of events passing the event selection in the 1-tag (light blue dots) and 2-tag regions (violet dots) divided the number of events that pass all the event selection cuts except for the *b*-tagging request. The plot shows that both curves have a decreasing trend at high p_{T}^{V} values because the VR track-jets start to become closer and the *b*-tagging algorithm is less efficient. Knowing that the average *b*-tagging efficiency is 70%, in the 1-tag region the expected efficiency is 91%.[4] The observed efficiency in the 1-tag region is around 80% which is a lower than the expected value. Differently, the efficiency in the 2-tag region is 35%, while the expected efficiency is 49%. The discrepancy between the expected and the observed values has been further investigated. Figure 5.3b shows the efficiency in the 1-tag and 2-tag regions considering only events in which the VR track-jets reconstruct correctly the *b*-hadron. It is possible to know if the VR track-jet really contains the *b*-hadron looking at the generator level information of the simulated event. In this case the efficiency in the 1-tag region is approximately 91%, while in the 2-tag region is around 49%. Moreover, it has been also noticed that 25% of the signal events with at least three VR track-jets[5] have only one of the two leading VR track-jets which contains one *b*-hadron and the other VR track-jet

[4] In the 1-tag region either the leading or the sub-leading VR track-jets is *b*-tagged so the expected efficiency can be calculated subtracting to the unity the probability to have zero *b*-tagged track-jets which is 9% ($0.3 \times 0.3 = 0.09 = 9\%$).

[5] About 30% of the signal events have at least three VR track-jets.

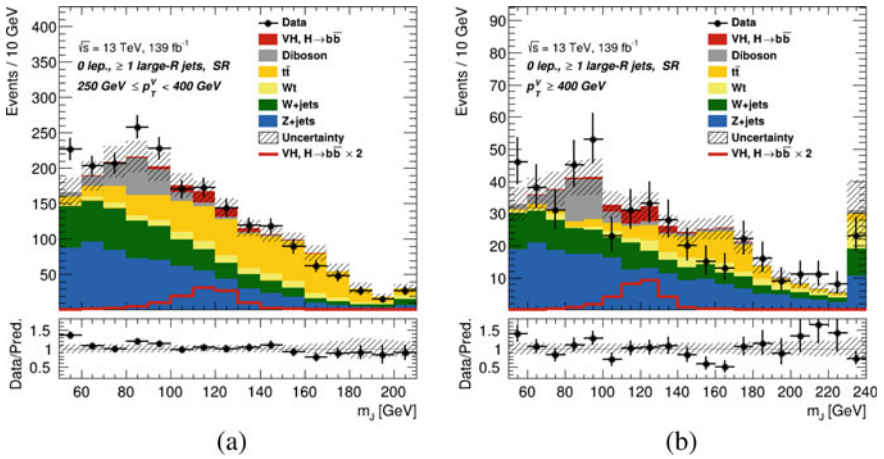

Fig. 5.4 The invariant large-R jet mass m_J pre-fit distribution in the 0-lepton channel in the SR in the 250 GeV $\leq p_{\mathrm{T}}^V < 400$ GeV (**a**) and $p_{\mathrm{T}}^V \geq 400$ GeV (**b**) momentum ranges. The data are shown as black dots. The background contributions are shown as filled histograms. The Higgs boson signal is shown as a filled histogram on the top of the background contribution, and unstacked as an unfilled histogram, multiplied by a factor 2. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background is shown in the lower panel

contains a light-hadron. The other b-hadron is contained in the third VR track-jet. This result shows that a possible way to recover the inefficiency of the b-tagging algorithm is to apply it considering the leading three VR track-jets instead of considering only the two leading VR track-jets.

The large-R jet mass m_J pre-fit distributions for data and simulated samples in the SR in the two p_{T}^V regions are shown in Fig. 5.4. The data are shown as black dots, while the signal and background contributions[6] are shown as filled histograms. The Higgs boson signal is also shown unstacked as an unfilled histogram multiplied by the factor indicated in the legend. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background events is shown in the lower panel to highlight the agreement between data and MC simulation. The same style and convention are used also for the following plots in this thesis.

The dominant backgrounds of this channel are the Z+jets, W+jets and top processes. For this reason, a dedicated CR is set to model the top background (see Sect. 5.3.3).

[6] The contribution of the multi-jet background is not shown in the plots because, as discussed in the text, it is negligible after applying the event selection cuts. More information of the multi-jet background suppression can be found in Sect. 6.3.7.

5.2.3 1-Lepton Channel Selection

In 1-lepton channel, a set of cut is applied to select events containing a $W \to l\nu$ decay. All the events in the 1-lepton channel are required to have a *signal* lepton, and a veto on any additional *loose* leptons is applied.

Lepton isolation requirements remove most of the non-prompt lepton background. To additionally suppress the multi-jet background, a cut on $E_{\mathrm{T}}^{\mathrm{miss}}$ ($E_{\mathrm{T}}^{\mathrm{miss}} > 50$ GeV) is applied in the electron sub-channel. In the muon sub-channel, such cut is not applied because there are few events from the multi-jet background. More details on the multi-jet estimate in 1-lepton channel can be found in Sect. 6.3.7.

In order to reduce the contribution from the top and W+jets production, a further selection on the rapidity difference between the Higgs-candidate jet and the W boson is applied, $|\Delta y(W, H_{\mathrm{cand.}})| < 1.4$. To calculate the rapidity of the W boson it is necessary to fully reconstruct its momentum. Neglecting off-shell effects and W boson width, and assuming[7] $m^l = m^\nu = 0$ and $E_{\mathrm{T}}^{\mathrm{miss}} = p_T^\nu$, the longitudinal momentum of the neutrino is estimated by the following equation extracted constraining the lepton + neutrino system to have the W boson mass:

$$p_z^\nu = \frac{1}{2(p_{\mathrm{T}}^l)^2}\left[Xp_z^l \pm E^l\sqrt{X^2 - \left[m_{TW}^2 + 2p_{\mathrm{T}}^l E_{\mathrm{T}}^{\mathrm{miss}}\cos\left(\Delta\phi(l, E_{\mathrm{T}}^{\mathrm{miss}})\right)\right]^2}\right] \quad (5.1)$$

with

$$X = m_W^2 + 2p_{\mathrm{T}}^l E_{\mathrm{T}}^{\mathrm{miss}}\cos\left(\Delta\phi(l, E_{\mathrm{T}}^{\mathrm{miss}})\right) \quad (5.2)$$

where the l and ν superscripts represent the charged lepton and the neutrino, respectively, $\Delta\phi(l, E_{\mathrm{T}}^{\mathrm{miss}})$ is the azimuthal angle between the lepton and the $E_{\mathrm{T}}^{\mathrm{miss}}$, and m_{TW} is the transverse mass of the W boson defined as $m_{TW} = \sqrt{2p_{\mathrm{T}}^l E_{\mathrm{T}}^{\mathrm{miss}}(1 - \cos\left(\Delta\phi(l, E_{\mathrm{T}}^{\mathrm{miss}})\right))}$. This method leads to two solutions, the retained solution is the one that minimises the difference between the longitudinal boost of the W and Higgs bosons. The equation has imaginary solutions[8] if the discriminant is less zero which means that the transverse mass of the W boson m_{TW} is larger than the W boson mass m_W. In this particular case, the discriminant is set to zero which means setting the W transverse mass to the W mass ($m_{TW} = m_W$). Preliminary studies show that the cut on the $|\Delta y(W, H_{\mathrm{cand.}})|$ discards 20% of the background events and 5% of signal events. The rapidity cut brings an improvement of 5% in the expected significance.

The efficiency of each selection cut used in the 1-lepton channel has been studied using MC signal samples. As in the 0-lepton channel, the efficiency is evaluated as the number of events that pass a cut over the total number of generated events. Figure 5.5 shows the efficiency of 1-lepton cuts as a function of truth p_{T}^V. To study

[7] The assumption of neglecting the lepton and neutrino masses is explained by the fact that the lepton and neutrino have masses much smaller that the W boson mass.

[8] The imaginary solutions are a consequence of the finite resolution of the detector.

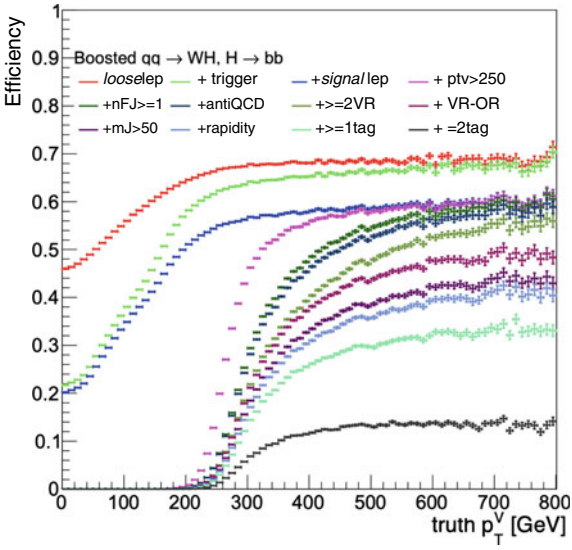

Fig. 5.5 Efficiency breakdown in 1-lepton channel. The cuts have been applied in cascade so in each step there is an additional request. Each curve with different colors shows the effect of cut cascade as described in the text

the efficiency the following cuts have been applied in cascade and in each step there is an additional request:

- *loose lepton cut*: events with a *loose* electron or *loose* muon are required (red dots);
- *trigger cut*: events that pass the trigger selection are required (lime dots);
- *signal lepton cut*: events with a *signal* electron or *signal* muon are required (blue dots);
- p_{T}^{V} *cut*: events with $p_{\mathrm{T}}^{V} > 250$ GeV are required (magenta dots);
- large-R *jet cut*: events with at least one large-R jet are required (dark green dots);
- *anti-QCD cut*: events that pass the anti-QCD cut ($E_{\mathrm{T}}^{\mathrm{miss}} > 50$ GeV in the electron sub-channel) are required (dark blue dots);
- *VR track-jets cut*: events with at least two VR track-jets ghost associated to the Higgs candidate jet are required (crocodile dots);
- *VR overlap removal cut*: events that pass the VR overlap removal procedure are required (purple dots);
- m_J *cut*: events with $m_J > 50$ GeV are required (violet dots);
- *rapidity cut*: events with $|\Delta y(W, H_{\mathrm{cand.}})| < 1.4$ are required (azure dots);
- *1-tag cut*: events with at least one b-tagged track-jet are required (turquoise dots);
- *2-tag cut*: events with exactly two b-tagged track-jets are required (black dots).

In the 1-lepton channel only 10% of the signal events has been selected and most of the events do not pass the *loose* lepton cut and the b-tagging requirements. The request of one electron or one muon in the final states reduces the number of the leptonic W events to 78% because the events with the W boson decays into $\tau + \nu$ would pass the event selection requirements only if the τ lepton decays leptonically. The observed efficiency is lower than this fraction because of the geometrical acceptance and lepton reconstruction efficiency.

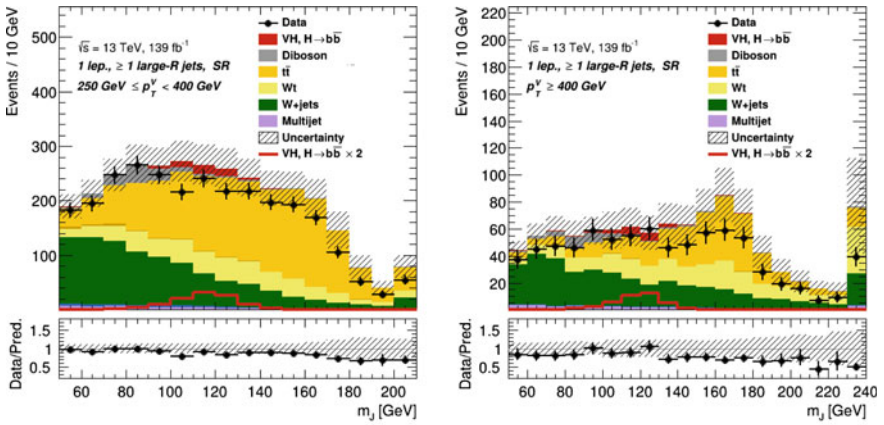

Fig. 5.6 The invariant large-R jet mass m_J pre-fit distribution in the 1-lepton channel in the SR in the 250 GeV $\leq p_{\mathrm{T}}^V <$ 400 GeV (**a**) and $p_{\mathrm{T}}^V \geq$ 400 GeV (**b**) momentum ranges. The data are shown as black dots. The background contributions are shown as filled histograms. The Higgs boson signal is shown as a filled histogram on the top of the background contribution, and unstacked as an unfilled histogram, multiplied by a factor 2. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background is shown in the lower panel

As in the 0-lepton channel the impact of each single cut has been studied. Almost all the events with truth $p_{\mathrm{T}}^V \geq 250$ GeV pass the *trigger cut* and less than 5% of the events are discarded. The *signal lepton cut* has an efficiency of 90% with a flat trend. With the *anti-QCD cut* less than 5% of signal events are discarded. The *rapidity cut* as an efficiency of 95%. The efficiencies of the large-R jet and VR track-jets related cuts have a similar trend in 0-lepton and 1-lepton channel. Also the efficiencies of the b-tagging cuts agree between 0-lepton and 1-lepton channel.

The large-R jet mass m_J pre-fit distributions for data and simulated samples in the SR in the two p_{T}^V regions are shown in Fig. 5.6. The dominant background contributions are the W+jets and top processes. Dedicated CRs are set to model the top backgrounds (see Sect. 5.3.3).

5.2.4 2-Lepton Channel Selection

In the 2-lepton channel, a Z boson decaying into two same flavour leptons (ee, $\mu\mu$) is reconstructed together with the Higgs candidate jet. Therefore selected events have exactly two same flavour *loose* leptons. Beside this request, at least one *signal* lepton is required. Moreover in the muon sub-channels, the two leptons must have opposite

charge. The charge requirement is not applied to the di-electron events due to the higher rate of charge misidentification.

To reduce the Z+jets background, a cut on the rapidity difference between the Z and H bosons is applied. In this case, the four momentum of the Z boson is fully reconstructed using the information of the two leptons system. The same threshold as in 1-lepton channel is applied, $|\Delta y(Z, H_{\text{cand.}})| < 1.4$. To further suppress the non-resonant background, the invariant mass of the di-lepton system must be consistent with the Z boson mass, 66 GeV $< m_{ll} <$ 116 GeV.

The lepton-p_{T} imbalance $(p_{\text{T}}^{l_1} - p_{\text{T}}^{l_2})/p_{\text{T}}^{Z}$ is sensitive to the Z boson polarization which is found to be different between signal and Z+jets events [3]. The lepton-p_{T} imbalance can be used as a discriminant between signal and background events. The two leptons coming from the signal events usually have the same transverse momentum ($p_{\text{T}}^{l_1} \sim p_{\text{T}}^{l_2}$), while in case of Z+ jets events the distribution of the lepton-p_{T} imbalance has a flat trend. To discard background events, the lepton-p_{T} imbalance is required to be less than 0.8.

The efficiency of each selection cut used in the 2-lepton channel has been studied using simulated signal samples. As before, the efficiency is evaluated as the number of events that pass a cut over the total number of generated events. Figure 5.7 shows the efficiency of the 2-lepton cuts as a function of the truth p_{T}^{V}. To study the efficiency, the following cuts have been applied in cascade and in each step there is an additional request:

- *loose lepton cut*: events with two *loose* electrons or two *loose* muons are required (red dots);
- *trigger cut*: events that pass the trigger selection are required (lime dots);
- m_{ll} *cut*: events with the invariant mass of the di-lepton system consistent with the Z boson mass, 66 GeV $< m_{ll} <$ 116 GeV. (blue dots);
- p_{T}^{V} *cut*: events with $p_{\text{T}}^{V} \geq 250$ GeV are required (magenta dots);
- large-R *jet cut*: events with at least one large-R jet are required (light blue dots);
- *VR track-jets cut*: events with at least two VR track-jets ghost associated to the Higgs candidate jet are required (green dots);
- *VR overlap removal cut*: events that pass the VR overlap removal procedure are required (gray dots);
- m_J *cut*: events with $m_J > 50$ GeV are required (violet dots);
- *rapidity cut*: events with $|\Delta y(Z, H_{\text{cand.}})| < 1.4$ are required (golden dots);
- *lepton-p_{T} imbalance cut*: events with $(p_{\text{T}}^{l_1} - p_{\text{T}}^{l_2})/p_{\text{T}}^{Z} < 0.8$ are required (light brown dots);
- *1-tag cut*: events with at least one b-tagged track-jet are required (brown dots);
- *2-tag cut*: events with exactly two b-tagged track-jets are required (light red dots).

As in the 1-lepton channel only 10% of the simulated signal events pass the full cut cascade. The request of two *loose* leptons in the final state removes 40–50% of the events. The shape of the efficiency of the *trigger cut* has a discontinuity because in the muon sub-channel the $E_{\text{T}}^{\text{miss}}$ trigger is used from $p_{\text{T}}^{V} > 150$ GeV. Events with a p_{T}^{V} below 150 GeV have been selected using a single muon trigger. All the di-muons events used in the analysis have a $p_{\text{T}}^{V} \geq 250$ GeV so the single muon trigger is never

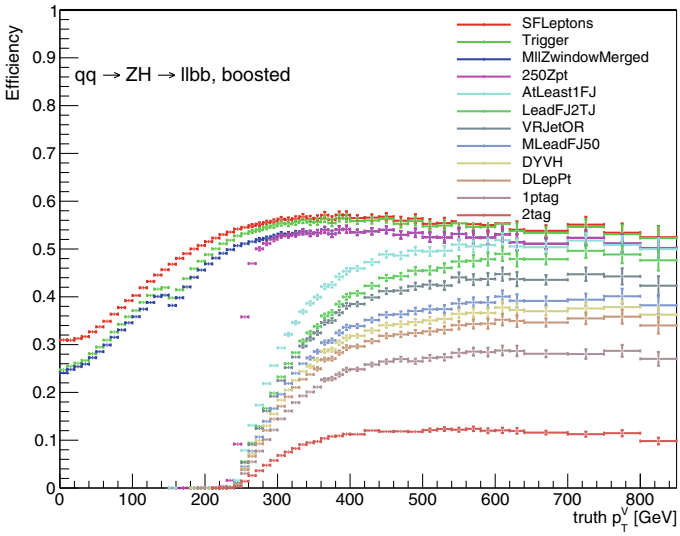

Fig. 5.7 Efficiency breakdown in 2-lepton channel. The cuts have been applied in cascade so in each step there is an additional request. Each curve with different colors shows the effect of cut cascade as described in the text

applied. After applying the *loose lepton* and *trigger cuts*, it has been tested the request of having two opposite muons in the muon-sub channel. This cut is fully efficient and for this reason is not reported in Fig. 5.7. The m_{ll} *cut* is almost fully efficient, only 2% of the events are discarded. After a sharp turn-on, the p_{T}^{V} *cut* is fully efficient. With the *lepton-*p_{T} *imbalance cut* less than 10% of the events are removed. The efficiency of the jets related cuts and *b*-tagging requirements have been compared among the three lepton channels and the results of the 2-lepton channel are in agreement with the one of the 0-lepton and 1-lepton channel. Also the rapidity cut has the same trend in 1- and 2-lepton channels.

The large-R jet mass pre-fit distributions for data and simulated sample in the SR in the two p_{T}^{V} regions are shown in Fig. 5.8. In this channel, the dominant background is the Z+jets process.

5.3 Event Categorization

After applying the requirements described above, the events in the three lepton channels are categorized depending on the p_{T}^{V}. In 0-lepton and 1-lepton channels, the events are further categorized depending on the number of small-R jets and on the number of *b*-tagged track-jets outside the Higgs candidate jets. At the end of this section there is a summary of the analysis region definition.

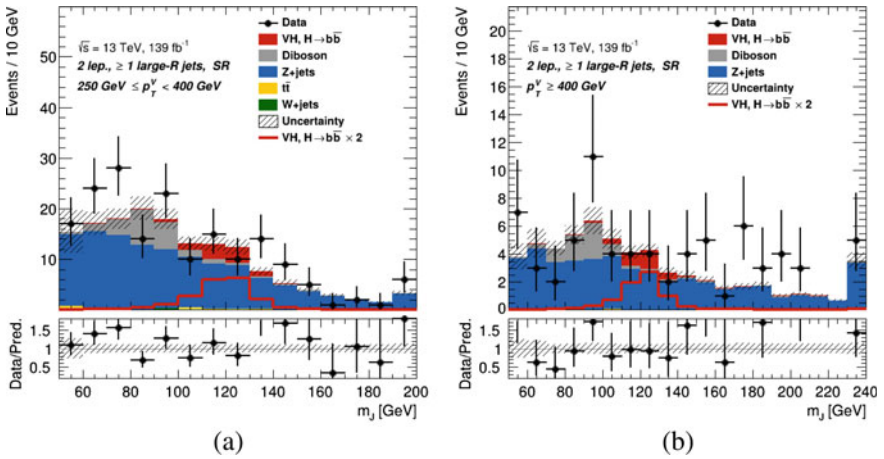

Fig. 5.8 The invariant large-R jet mass m_J pre-fit distribution in the 2-lepton channel in the SR in the 250 GeV $\leq p_{\mathrm{T}}^V <$ 400 GeV (**a**) and $p_{\mathrm{T}}^V \geq$ 400 GeV (**b**) momentum ranges. The data are shown as black dots. The background contributions are shown as filled histograms. The Higgs boson signal is shown as a filled histogram on the top of the background contribution, and unstacked as an unfilled histogram, multiplied by a factor 2. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background is shown in the lower panel

5.3.1 p_T^V Splitting

The events are categorized depending on the p_{T}^V because the phase-space with high signal-to-background ratio is at high values of the vector boson transverse momentum. Moreover, BSM effects may be more pronounced in the high-p_{T} region. Two regions are considered: a medium energy region and a high energy region, 250 GeV $\leq p_{\mathrm{T}}^V <$ 400 GeV and $p_{\mathrm{T}}^V \geq$ 400 GeV, respectively. These p_{T}^V intervals are coherent with the cuts used in the STXS categorization (see Sect. 1.5). Figure 5.9 shows the p_{T}^V pre-fit distributions in the SR in the three lepton channels.

5.3.2 Signal Region Splitting

In 0- and 1-lepton channels the SR is defined by requiring to have zero b-tagged track-jets outside the Higgs candidate jet in order to enhance the top background rejection. It is possible to further discriminate between the top process and the signal using the jet multiplicity of the two processes. The difference in jet multiplicity can be easily deduced from the leading order Feynman diagrams for the $t\bar{t}$ and WH processes shown in Fig. 5.10. The $t\bar{t}$ events passing the event selection are mainly

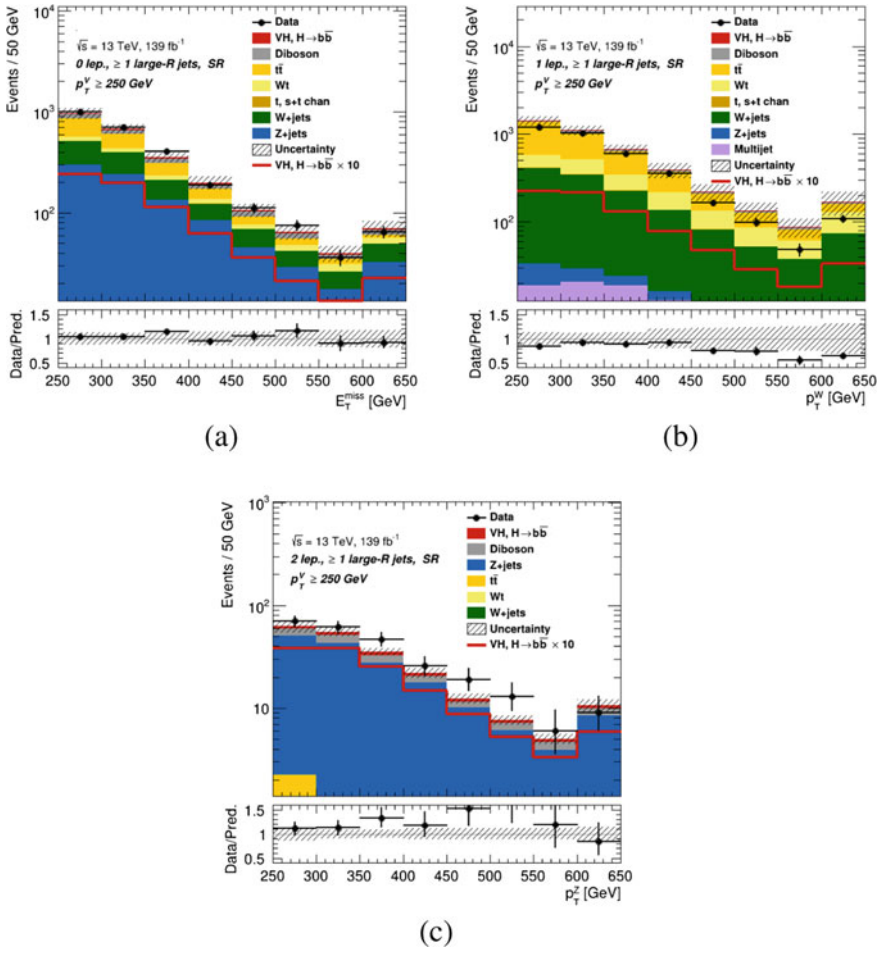

Fig. 5.9 The p_{T}^{V} pre-fit distributions in 0-lepton (**a**), 1-lepton (**b**) and 2-lepton (**c**) SRs for $p_{\mathrm{T}}^{V} \geq 250$ GeV. The data are shown as black dots. The background contributions are shown as filled histograms. The Higgs boson signal is shown as a filled histograms on top of the background contribution, and unstacked as an unfilled histogram, multiplied by a factor 10. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background is shown in the lower panel

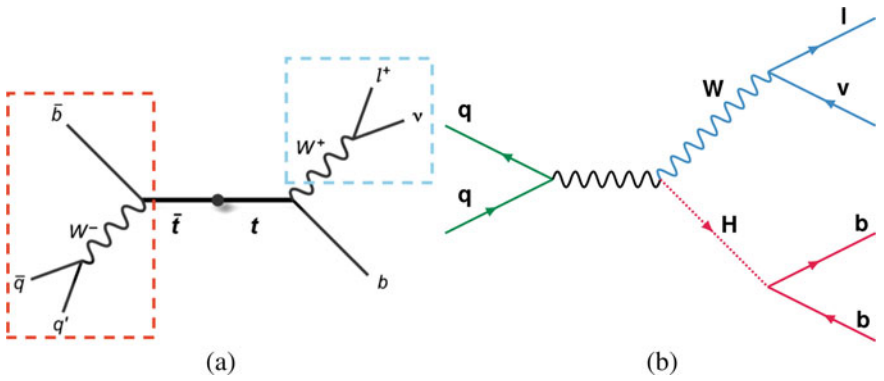

Fig. 5.10 Leading order Feynmam diagram for the $t\bar{t}$ (**a**) and WH (**b**) processes. The red box in the Feynman diagram of the $t\bar{t}$ process indicates the decay products of a hadronically decaying top quark which can be wrongly selected as Higgs candidate jet. The light blue box includes the decay products of a W boson which is correctly reconstructed as V boson candidate

constituted by events in which one W boson decays leptonically and the other one decays hadronically. The W boson that decays leptonically is expected to be identified by the vector boson selection (light blue box in Fig. 5.10a). One of the two jets from the hadron decay of the W boson could be mis-identified as a b-jet. In the red box, this jet together with the b-quark of the top decay can be wrongly selected as Higgs candidate (red box in Fig. 5.10a). As a result, the event has an additional b-jet outside the Higgs candidate jet.

An additional way to discriminate top events from signal events is study the hadron activity of the event because $t\bar{t}$ events are characterised by more activity. The angular distance between the Higgs candidate jet and the small-R jets is studied to avoid using jets associated to the Higgs candidate jet.[9] Figure 5.11 shows the angular distance between the Higgs candidate and the leading small-R jet for signal VH events in 1-lepton signal region. The request $\Delta R > 1$ ensures that the small-R jet is not matched to the Higgs candidate jet.

To assess the jet activity, the p_{T} distribution of the leading small-R jet not matched to the Higgs candidate jet is studied. It is expected that the jets produced by $t\bar{t}$ events have higher p_{T} with respect to the jets produced by signal events. Figure 5.12 shows the p_{T} distribution of the leading small-R jet not matched to the Higgs candidate jet in 1-lepton SR, 250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV, for the major processes. The distribution of each process is normalized to the total number of events of the process that pass the event selection. Looking at the plot, the distribution of the $t\bar{t}$ events has a maximum at $p_{\mathrm{T}} \sim 30$ GeV, while the maximum of the distribution of signal events is around $p_{\mathrm{T}} \sim 20$ GeV. To discriminate signal from $t\bar{t}$ events, it is required that the small-R jets

[9] Small-R jets are used in these studies instead of VR track-jets in order to have an easier estimate of the systematics uncertainties.

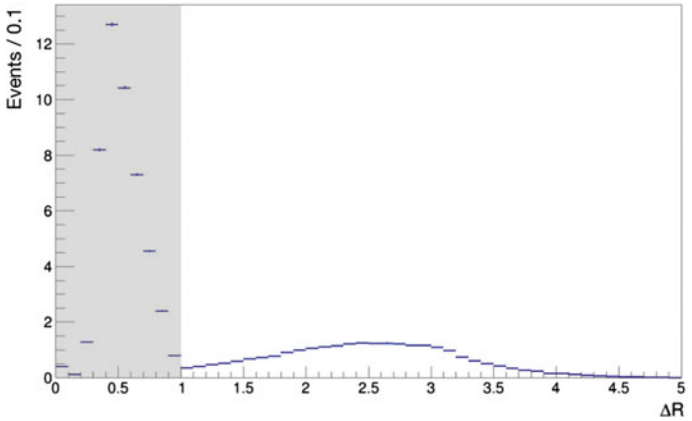

Fig. 5.11 Distribution of the angular distance between the Higgs candidate and the leading small-R jet for VH signal events in 1-lepton SR, 250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV. The gray area represents all the signal events in which the selected small-R jet is matched to the Higgs candidate

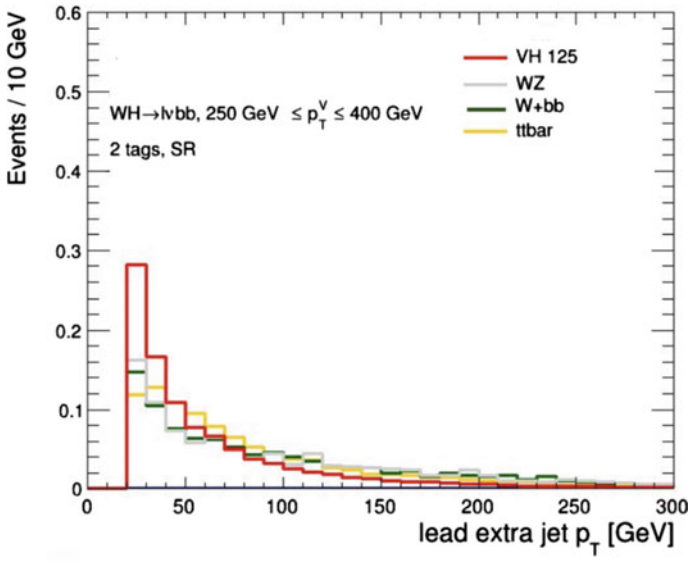

Fig. 5.12 p_{T}^{V} distribution of the leading small-R jet outside the Higgs candidate jet in 1-lepton SR, 250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV, for the main processes. The distribution of each process is normalized to the total number of events of the process that pass the event selection

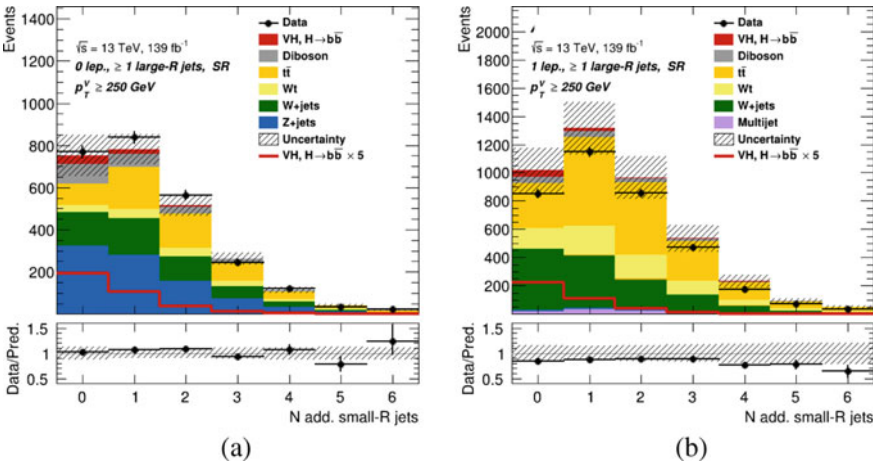

Fig. 5.13 Number of small-R jets non-matched to the Higgs candidate jet distribution in 0-lepton (**a**) and 1-lepton (**b**) SRs for $p_{\mathrm{T}}^{V} \geq 250$ GeV. The data are shown as black dots. The background contributions are shown as filled histograms. The Higgs boson signal is shown as a filled histograms on top of the background contribution, and unstacked as an unfilled histogram, multiplied by a factor 5. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background is shown in the lower panel

outside the Higgs candidate jet must have $p_{\mathrm{T}} > 30$ GeV. This p_{T} cut is also aligned with one applied on the small-R jets in the STXS framework (see Sect. 1.5).

The final step is to apply a cut on the jet multiplicity. Figure 5.13 shows the distribution of the number of small-R jets outside the Higgs candidate in the 0-lepton and 1-lepton channel. In both channels the distribution of $t\bar{t}$ events has different shape with respect to the shape of the signal events which has a peak at zero.

Requiring zero small-R jets outside the Higgs candidate jet it is possible to discard about 60(70%) of backgrounds events and 35% of signal events in 0-lepton (1-lepton) channel. In the analysis the region with zero small-R jets outside the Higgs candidate jet is called *high purity signal region* (HP SR) because it is the region with highest signal-to-background ratio and less top events. To avoid signal loss, events in the SR with one or more small-R jets are used to define the *low purity signal region* (LP SR). This categorization of the SR is only applied in the 0- and 1-lepton channels. It is not applied in the 2-lepton channel because the $t\bar{t}$ process is an almost negligible background. Preliminary studies show that the splitting of the SR brings a 30% (17%) gain in the expected significance in 1-lepton (0-lepton) channel. The improvement in the 1-lepton channel is bigger with respect to the one in 0-lepton channel because in 1-lepton channel the fraction of $t\bar{t}$ events is higher than in 0-lepton channel.

The large-R jet mass m_J pre-fit distributions for data and simulated sample in the HP SR and LP SR in the two p_{T}^{V} regions in 0-and 1-lepton channels are shown in Figs. 5.14 and 5.15.

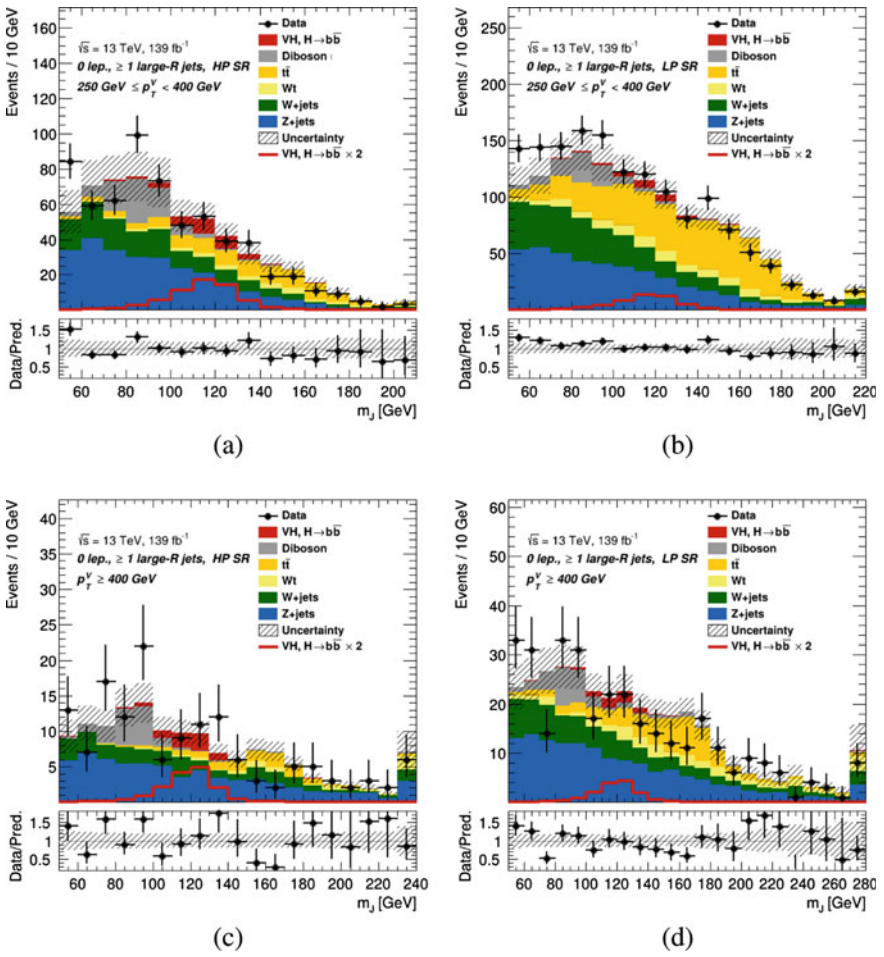

Fig. 5.14 The invariant large-R jet mass m_J pre-fit distributions in the 0-lepton channel in the HP SR (left) and LP SR (right) in the 250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV (top) and $p_{\mathrm{T}}^{V} \geq$ 400 GeV (bottom) region. The data are shown as black dots. The background contributions are shown as filled histograms. The Higgs boson signal is shown as a filled histogram on the top of the background contribution, and unstacked as an unfilled histogram, multiplied by a factor 2. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background is shown in the lower panel

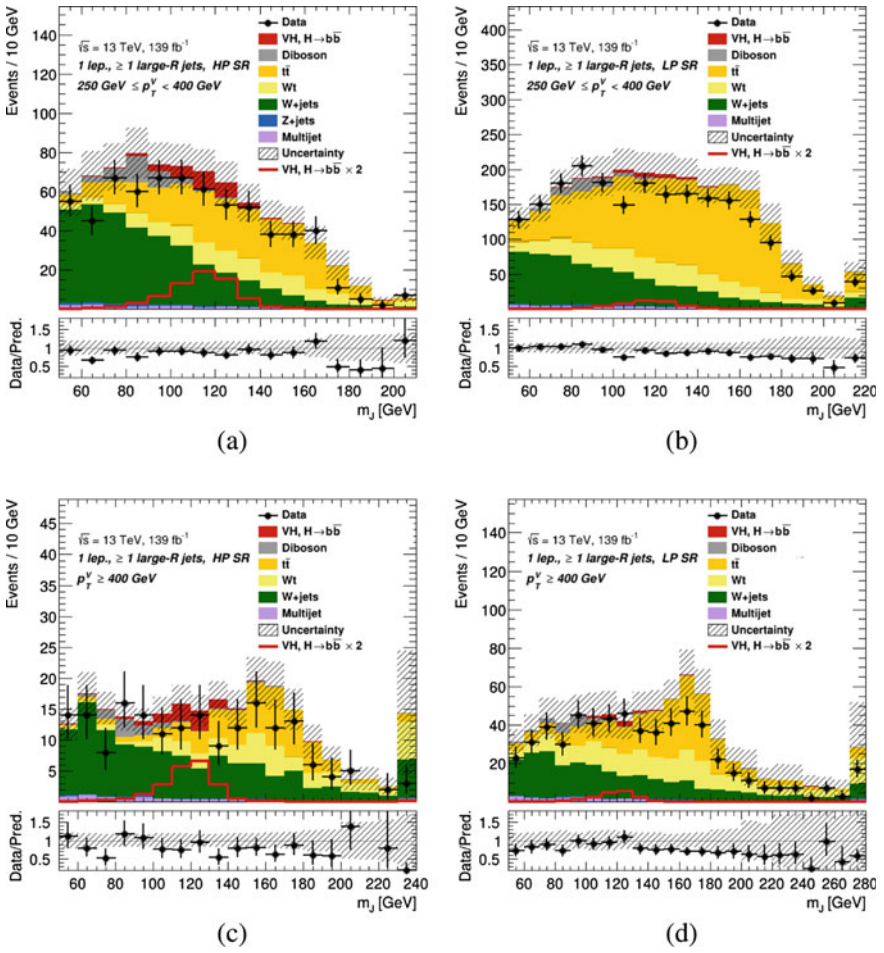

Fig. 5.15 The invariant large-R jet mass m_J pre-fit distributions in the 1-lepton channel in the HP SR (left) and LP SR (right) in the 250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV (top) and $p_{\mathrm{T}}^{V} \geq$ 400 GeV (bottom) region. The data are shown as black dots. The background contributions are shown as filled histograms. The Higgs boson signal is shown as a filled histogram on the top of the background contribution, and unstacked as an unfilled histogram, multiplied by a factor 2. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background is shown in the lower panel

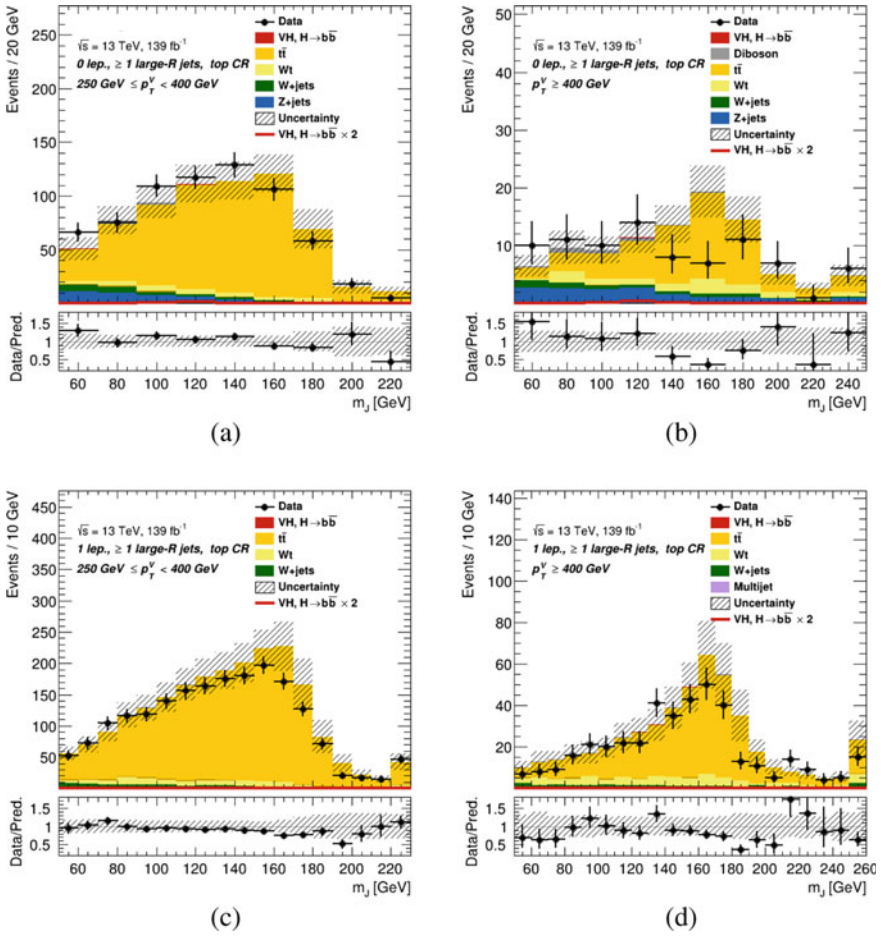

Fig. 5.16 The invariant large-R jet mass m_J pre-fit distributions in the 0-lepton (top) and 1-lepton channel (bottom) channel in the CR in the 250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV (left) and $p_{\mathrm{T}}^{V} \geq$ 400 GeV (bottom) region. The data are shown as black dots. The background contributions are shown as filled histograms. The Higgs boson signal is shown as a filled histogram on the top of the background contribution, and unstacked as an unfilled histogram, multiplied by a factor 2. The size of the combined statistical and systematic uncertainty for the sum of the signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the signal and background is shown in the lower panel

5.3.3 Control Region Definition

The $t\bar{t}$ process is one of the main background in the 0- and 1-lepton channel and a way to further study it is building a control region. In the 0-lepton and 1-lepton channel, the CRs are defined requiring at least one b-tagged track-jet outside the Higgs candidate jet.[10] As illustrated in Fig. 5.10a, the $t\bar{t}$ decay can give at least one stand-alone b-quark outside the Higgs candidate jet.

Figure 5.16 shows the m_J pre-fit distributions in the CRs in 0- and 1-lepton channels. The plots confirm that most of the events in the CRs are from the $t\bar{t}$ background process. Due to the low statistic in the 0-lepton channel a coarser binning is chosen with respect to the 1-lepton channel. Moreover, comparing the data to the MC prediction, the distributions show a mis-modelling at around the top mass.

In the 0-lepton channels 80% (56%) of the events in the CRs are from the top process in the 250 GeV $\leq p_{\mathrm{T}}^{V} <$ 400 GeV ($p_{\mathrm{T}}^{V} \geq$ 400 GeV) region, while in 1-lepton channel the fraction of $t\bar{t}$ events is ∼90%. In 1-lepton channel the fraction of $t\bar{t}$ events in the CRs is higher than in the 0-lepton channel because the signal final state is more similar to the $t\bar{t}$ final state.

5.3.4 Summary of the Signal and Control Regions

According to the event categorization described in the previous sub-sections, in the analysis ten SRs and four CRs are considered. The SRs and CRs are summarized in Table 5.5.

Table 5.5 Summary of the definition of the analysis regions. Regions with relatively large signal purity are marked with the label SR. Background enriched regions are marked with the label CR

Channel	Categories					
	250 < p_T^V < 400 GeV			p_T^V > 400 GeV		
	0 add. b track-jets		≥ 1 add. b track-jets	0 add. b track-jets		≥ 1 add. b track-jets
	0 add. small-R jets	≥ 1add. small-R jets		0 add. small-R jets	≥ 1add. small-R jets	
0-lepton	SR	SR	CR	SR	SR	CR
1-lepton	SR	SR	CR	SR	SR	CR
2-lepton	SR			SR		

[10] In this case, the b-tagged track-jet is considered outside the Higgs candidate jet if it is not ghost-matched to the Higgs candidate jet.

References

1. Bukin AD (2007) Fitting function for asymmetric peaks. arXiv:0711.4449 [physics.data-an]
2. Aad G et al. (2020) Measurement of the associated production of a Higgs boson decaying into b-quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13TeV$ with the ATLAS detector. arXiv:2008.02508 [hep-ex]
3. Goncalves D, Nakamura J (2018) Role of the Z polarization in the $H \to b\bar{b}$ measurement. Phys Rev D 98:093005. https://doi.org/10.1103/PhysRevD.98.093005. arXiv:1805.06385 [hep-ph]