CMS Physics Analysis Summary

Contact: cms-pag-conveners-higgs@cern.ch 2015/03/12

Search for a Standard Model-like Higgs boson decaying into $WW \rightarrow \ell\nu q\bar{q}'$ in exclusive jet bins in pp collisions at $\sqrt{s} = 8$ TeV

The CMS Collaboration

Abstract

A search for a Standard Model Higgs boson decaying into the WW final state is performed with an integrated luminosity of up to 19.3 fb$^{-1}$ of pp collisions at $\sqrt{s} = 8$ TeV in the high mass region $600 < m_H < 1000$ GeV. This search is performed in the semi-leptonic channel, where the hadronically decaying W boson is highly boosted and its decay products are contained in one jet. Jet substructure techniques are used in identifying the hadronically decaying W. The analysis is divided into categories of additional jet activity in order to highlight the gluon fusion and vector boson fusion production mechanisms improving the sensitivity. No evidence for a standard model-like Higgs boson has been found in the investigated mass region. The results are also interpreted in a beyond the Standard Model Higgs scenario, based on an effective theory which predicts the existence of heavy Higgs singlet, with the mass of the lighter one being around 125 GeV.
1 Introduction

The Standard Model (SM) of electroweak interactions [1–3] relies on the existence of the Higgs boson, $H$, a scalar particle associated with the field responsible for spontaneous electroweak symmetry breaking [4–9]. The observation of a Higgs boson with a mass of 125 GeV [10–12] is consistent with the theoretical constraint coming from the unitarization of WW scattering at high energies [13–22]. However, it is still important to continue searching for the SM-like Higgs boson in the high mass region. Several popular scenarios, such as general two-Higgs-doublet models [23] or models in which the SM Higgs boson mixes with a heavy electroweak singlet [24–37], predict the existence of additional resonances at high mass, with couplings similar to the SM Higgs boson. This study reports the search for a SM-like Higgs boson decay to two $W$ bosons in the mass range from $600 < m_H < 1000$ GeV. In addition, this analysis shows the first selection of the vector boson scattering topology at high masses, paving the way for future investigations of the WW scattering process in the semi-leptonic final state.

The semi-leptonic final state investigated in this analysis can be triggered and separated from the QCD multijet production due to the presence of one lepton from the leptonic decay of one $W$ boson, and shows a large branching fraction due to the hadronic decay of the other $W$ boson. Because of the large invariant mass searched for, the $W$ bosons have a large transverse momentum, which causes the hadronic decay products to be collimated to the point of being reconstructed as a single jet in the detector. Jet substructure techniques are therefore employed for identifying single jets that have originated from a highly boosted hadronically decaying $W$ boson.

The analysis has been performed on the pp-collision data sample recorded by the CMS detector at the center-of-mass energy of 8 TeV, corresponding to an integrated luminosity of 19.3 fb$^{-1}$. The obtained results are interpreted as a search for a SM Higgs boson, as well as in the context of a beyond the SM (BSM) scenario, based on an effective theory where a second scalar particle besides the resonance at 125 GeV completes the unitarization of the WW scattering [38].

2 The CMS detector

The CMS detector, described in detail in [39], is a multipurpose apparatus designed to study high transverse momentum ($p_T$) physics processes in proton-proton collisions. It consists of pixel and silicon-strip tracker up to a pseudorapidity of $|\eta| < 2.5$ which, together with a 3.8 T solenoid, provides a track momentum resolution of 1% at 100 GeV; a granular electromagnetic crystal-calorimeter, extending up to $|\eta| < 3$, with an energy resolution of about 3%/\sqrt{E} [40]; a hadronic calorimeter, extending up to $|\eta| < 5$, with an energy resolution of 100%/\sqrt{E} and a muon system, capable of reconstructing and identifying muons up to $|\eta| < 2.4$. The detector is nearly hermetic, allowing for measurements of the missing transverse energy ($E_T^{miss}$) in the event. The right-handed coordinate system is used by CMS [39] with the origin at the nominal interaction point, the x-axis pointing to the center of the LHC, the y-axis pointing up perpendicular to the LHC plane and the z-axis along the counterclockwise-beam direction. The polar angle $\theta$ is measured from the positive z-axis and the azimuthal angle $\phi$ is measured in the x-y plane. The pseudorapidity is defined as $\eta = -\ln(\tan(\theta/2))$.

3 Analysis Strategy

The analysis is performed studying the distribution of the fully reconstructed mass of the WW system, $m_{\ell\nu jj}$, which is expected to give a handle in distinguishing between a peaking shape for
the expected signal, increasing in width as a function of the mass hypothesis, and a smooth non-resonant falling background. Data from signal-free control regions are used to estimate normalization and shapes of the dominant backgrounds, reducing the dependence on the simulation. A heavy SM-like Higgs boson may be produced through gluon-gluon fusion (ggH) or vector boson fusion (VBF) processes, where the VBF production is expected to play a big role in the high mass region, reaching almost 50% of the total production cross section at a mass of 1 TeV. For this reason, the analysis is divided into bins of additional jet activity, to classify if the event is consistent with the VBF or ggH production mechanism. This improves the sensitivity over the previous CMS analysis [41] and the VBF category can be used as a first benchmark for di-boson scattering measurements in this final state.

4 Data and Simulation samples

Data have been collected by the CMS experiment with single lepton triggers, with a typical on-line $p_T$ threshold of 24 GeV for muons and 27 GeV for electrons. The pseudorapidity range for muons (electrons) extends up to $|\eta| < 2.4$ ($2.5$). The trigger efficiency is about 94% (90%) for muon (electron), with a small dependence, few percent, on $p_T$ and $\eta$. Simulated events are corrected for the trigger efficiency as a function of lepton $p_T$ and $\eta$.

The main source of background for this analysis is represented by single W boson production in association with jets and it is primarily estimated from data. Monte Carlo (MC) simulations are used to estimate residual contributions of the background, and to measure the selection efficiency of the Higgs signal. We use the MADGRAPH 5 1.3.30 [42] event generator to simulate the W boson and Drell-Yan production in association with up to four jets, referred to as W+jets and Z+jets respectively, which are showered with the PYTHIA v6.424 generator [43], with Z2* Tune [44]. In addition, we use the HERWIG++ 2.4 [45] generator to produce an alternate sample of W+jets events, which has a different parton shower model than PYTHIA, in order to estimate systematic uncertainties. Top-quark pair events are simulated using POWHEG 1.0 [46–49]. Single top production is also modeled using POWHEG [46–49]. Electroweak diboson WW production is simulated with aMC@NLO [50], WZ and ZZ are generated using PYTHIA, while WW electroweak scattering is simulated via PHANTOM [51]. The POWHEG generator has been used to produce Higgs signal events, and the showering has been performed with PYTHIA. For this analysis, samples with Higgs mass hypotheses ranging from 600 to 1000 GeV have been used, where the Higgs lineshape is reweighted according to the complex pole scheme calculation [52–55]. We correct the signal produced via ggH for the effect of interference with the SM $gg \rightarrow WW$ production process, using the method proposed in [53], computing the interference effect with the MCFM v6.6 generator [56]. A similar procedure is used for the signal produced via VBF, interfering with the SM $qq \rightarrow WW$ process, using PHANTOM generator. All simulated samples are corrected for any data-MC difference in the trigger and physics objects identification efficiencies.

In the BSM interpretation, we search for a electroweak singlet scalar, where a heavy SM-like Higgs boson mixes with the recently discovered candidate with a mass close to 125 GeV. Phenomenologically, the couplings of the two gauge eigenstates become related by unitarity and the effective coupling strength of the light Higgs boson is therefore reduced with respect to the SM case. If we define $C$ ($C'$) as the scale factor of the couplings of the low (high) mass with respect to the SM, one can fix $C^2 + C'^2 = 1$ as the unitarity condition to be preserved. Indirectly, we can set an upper limit at 95% CL on $C^2 < 0.446$ using the signal strength fits to the H(125) candidate as obtained in [12]. The heavy Higgs signal strength, $\mu'$, and the modified width, $\Gamma'$,
are defined as:

\[ \mu' = C'^2(1 - BR_{\text{new}}) \times \mu_{\text{SM}} \]  

(1)

\[ \Gamma' = \frac{C'^2}{(1 - BR_{\text{new}})} \times \Gamma_{\text{SM}} \]  

(2)

where \( BR_{\text{new}} \) is the branching fraction of the heavy Higgs to non-SM-like decay modes. The BSM heavy Higgs line shape is evaluated by reweighting the SM POWHEG samples following the procedure outlined in [57]. The model is implemented as a rescaling of the SM at NLO in QCD and LO in EWK. In this reweighting procedure, we use the ratio of two relativistic Breit-Wigner shapes: one for the narrower BSM signal, the other one for the generated SM Higgs boson. Notice that due to the poor resolution of this channel, departures from the Breit-Wigner approximation are not expected to modify significantly the final results.

The contribution due to the interference between the BSM Higgs and the background is estimated in a different way between ggH and VBF production modes. For ggH, the interference is assumed to scale according to the modified coupling of the Higgs boson as: \( \mu(I)_{\text{BSM}} = \mu_{\text{SM}}C'^2 + ls_{\text{SM}}C' \), where \( \mu (I) \) is the signal+interference strength in the BSM case. This assumption is based on the hypothesis that the couplings are similar to the SM case and simply rescaled due to unitarity constraints; this procedure has been validated using MCFM. On the other hand, to estimate the interference effect on VBF signal in the BSM scenario, simulations with PHANTOM are used to extract a dedicated set of weights for each \( C' \) value considered in the analysis. We assign an extra uncertainty of 10% to the interference for ggH, according to the multiplicative/additive scheme, while for VBF it is at the level of 9-11%, which has been estimated varying up and down renormalization and factorization scales in the event generation. In addition, the signal shape is very affected by the interference correction, in particular for high mass hypotheses in the VBF channel (\( m_H > 800 \) GeV), where a low-mass exponential tail is added on top of the signal peak.

5 Event Reconstruction

We search for a fully reconstructed SM-like Higgs boson \( H \rightarrow WW \rightarrow \ell\nu q\bar{q} \). The final state signature of this analysis is represented by an isolated charged lepton, either electron or muon, missing transverse energy \( E_T^{\text{miss}} \) and a jet containing the entire hadronic decay of a W boson. We then further categorize the events depending on lepton flavour and additional jet activity outside of the jet considered as the hadronic W candidate, in order to enhance the analysis sensitivity over the explored mass range. In the following sections, the selection criteria are explained before describing the categorization.

5.1 Leptonically decaying W boson, \( W_\ell \)

Muons are measured with the tracker and the muon system, within \( |\eta| < 2.4 \). Electrons are detected as tracks in the tracker pointing to energy clusters in the ECAL, within \( |\eta| < 2.5 \), excluding the transition region between the barrel and endcap, \( |\eta| \in [1.44, 1.57] \). Muons (electrons) are required to have a momentum transverse to the beam direction, \( p_T \), greater than 30 GeV (35 GeV). The lepton candidates are required to be isolated from other detector activity and compatible with the primary vertex of the event, which is chosen as the vertex with the highest \( \sum p_T^2 \) of its associated tracks. According to simulation, this requirement provides the correct assignment for the primary vertex in more than 99% of cases in both signal and background events. The isolation of e or \( \mu \) leptons is ensured by applying requirements on the sum of the transverse energies of all reconstructed particle-flow candidates [58], charged or
neutral, within a cone of $\Delta R < 0.4$ around the lepton direction, after subtracting the average pileup energy estimated using a jet area technique [59] on an event-by-event basis. We define $\Delta R \equiv \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}$, where $\Delta \eta$ and $\Delta \phi$ are differences in pseudorapidity and in azimuthal angle between two measured particles in radians. To reduce the backgrounds from the Drell-Yan and SM diboson processes, we veto the presence of any other loosely identified lepton in the event, with $p_T > 20$ GeV for electrons and $p_T > 10$ GeV for muons.

The leptonic W candidate is obtained from the lepton and the missing transverse energy, $E^\text{miss}_T$, in the event. Therefore, an accurate $E^\text{miss}_T$ measurement is essential for reconstructing the full event kinematics of the WW system. We use $E^\text{miss}_T$ measured in the event from particle-flow reconstruction [60], which combines the information from all CMS sub-detectors to reconstruct each particle. The $E^\text{miss}_T$ resolution, measured as a function of the sum $\sum E_T$ of the particle-flow objects in the event, varies from 4% at $\sum E_T = 60$ GeV to 10% at $\sum E_T = 350$ GeV [61]. In addition, the unmeasurable longitudinal component of the neutrino momentum is reconstructed by requiring the lepton-neutrino pair to have the invariant mass of a W boson, using a kinematic fit. The ambiguity in the involved second-order equation is resolved by taking the solution that yields the smallest $|p_z|$ value which, when evaluated in simulation, is closest to the true $|p_z|$ approximately 75% of the time.

5.2 Hadronically decaying W boson, $W^\text{had}$

Jets are reconstructed from calorimeter and tracker information using the particle flow algorithm. In addition, charged-particle tracks not originating at the primary vertex are not considered for jet clustering [62]. A jet quality requirement, primarily based on the energy balance between charged, neutral and electromagnetic components, is also applied.

Two different clustering algorithms, the anti-$k_T$ and Cambridge-Aachen algorithms (CA) [63–65], are used to reconstruct jets in the event, considering the whole detector pseudorapidity coverage up to $|\eta| < 5$. Cambridge-Aachen jets are clustered with a distance parameter of $\Delta R = 0.8$ (CA8) and are used for reconstructing the hadronically decaying W boson, $W^\text{had}$, where the entire W boson decays into a single jet. Anti-$k_T$ jets, clustered with a distance parameter of $\Delta R = 0.5$ (AK5), are used for categorizing the additional jet activity in the event and b-quark jet tagging. Any reconstructed jet overlapping with the isolated lepton, within $\Delta R = 0.5$ (0.8) for AK5 (CA8) jets, is removed in order to avoid double counting of the lepton as a jet.

The CA jets with larger $\Delta R$ parameter are used to increase signal acceptance of $W^\text{had}$. For signal events falling in the detector acceptance, approximately 65% (85%) of the $W^\text{had}$ decay products are contained in a cone of $\Delta R < 0.8$ for a signal mass of 600 GeV (1 TeV). Alternatively, approximately 15% (42%) of the $W^\text{had}$ decay products are separated by a distance of $\Delta R < 0.5$ for a signal mass of 600 GeV (1 TeV) and would not be reconstructed by the standard CMS AK5 jet finding algorithm; thus, the larger cone CA8 jet affords more signal acceptance in the single jet signature, while not losing events when the decay products are separated by $\Delta R < 0.5$. In addition, in the analysis where the jets are resolved, the combinatorial background is much larger.

6 Event Selection

The large invariant mass of the signal translates into a large transverse momentum of the W bosons. Therefore, the $p_T$ of both leptonically decaying W and of the hardest CA8 jet in the event are required to be above 200 GeV. We choose the highest $p_T$ CA8 jet in the event to uniquely be the $W^\text{had}$ candidate. At the same time, the missing transverse energy is required to
be above 50 GeV (70 GeV) for the muon (electron) channel, in order to suppress any possible contribution to the analysis from QCD multijet events. In addition, there are specific topological cuts which require that the $W$ bosons are back-to-back: the distance between the lepton and the $W$-jet should be large, $\Delta R_{l,j} > \pi/2$; the azimuthal distance between the missing energy and the $W$-jet should be large, $\Delta \Phi_{E_{\text{miss}},j} > 2.0$; and finally the azimuthal distance between the leptonically decaying $W$ boson and the $W$-jet should be $\Delta \Phi_{V,j} > 2.0$.

In addition to these kinematic cuts, we make further requirements on the additional AK5 jets in the event with a $p_T > 30$ GeV. These AK5 jets are required to be a distance of $\Delta R > 0.8$ from the $W_{\text{had}}$ candidate. To reduce the amount of the $t\bar{t}$ background, we veto the presence of any b-tagged AK5 jets in the event by selecting on the CSV discriminant [66], choosing a 70% on signal efficiency as optimal working point. Such a veto is only possible in the tracking region of the detector, for jets up to $|\eta| < 2.5$. Then, we categorize the events based on the number of AK5 jets with a $p_T > 30$ GeV: if there are 0 or 1 AK5 jets, they are classified in the 0+1 jet category, otherwise if there are 2 or more AK5 jets, they are candidates for the 2 jet category. For the 2 jet bin, we then select the 2 highest $p_T$ AK5 jets as VBF tag-jets, making the following cuts to isolate the VBF production mechanism:

- $\Delta \eta_{jj} > 3$
- $m_{jj} > 250$ GeV
- $m_{W_{\text{had}}+j} > 200$ GeV
- $m_{W_{\text{ lept}}+j} > 200$ GeV

where $W_{\text{had}}+j$ and $W_{\text{ lept}}+j$ are the 4-vector combinations of the $W$ candidates and their nearest AK5 jet. B-veto and top mass cut requirements are introduced to suppress top backgrounds from single top and $t\bar{t}$ production. The efficiencies of those cuts, on both signal ($m_H = 800$ GeV) and background, for the 2-jet bin category are reported in Table 1 together with their contribution to the improvement of the signal purity ($S/B$).

<table>
<thead>
<tr>
<th>cut</th>
<th>$\varepsilon$(ggH)</th>
<th>$\varepsilon$(VBF)</th>
<th>$\varepsilon$(single-$t$)</th>
<th>$\varepsilon$(VV)</th>
<th>$\varepsilon$(W+jets)</th>
<th>$\varepsilon$(t$t$)</th>
<th>S/B</th>
</tr>
</thead>
<tbody>
<tr>
<td>b-veto tag jet (2-jet bin)</td>
<td>0.91</td>
<td>0.94</td>
<td>0.44</td>
<td>0.89</td>
<td>0.82</td>
<td>0.40</td>
<td>$1.8 \times 10^{-2}$</td>
</tr>
<tr>
<td>hadronic top mass cut</td>
<td>0.82</td>
<td>0.96</td>
<td>0.60</td>
<td>0.78</td>
<td>0.69</td>
<td>0.44</td>
<td>$2.8 \times 10^{-2}$</td>
</tr>
<tr>
<td>leptonic top mass cut</td>
<td>0.72</td>
<td>0.96</td>
<td>0.45</td>
<td>0.68</td>
<td>0.64</td>
<td>0.37</td>
<td>$4.0 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

Table 1: Top killing selection efficiencies for signals (ggH and VBF with $m_H = 800$ GeV) and each background source. Purity defined as S/B is reported after each cut. B-veto efficiency refers to all the selected events in the 2-jet bin category after applying acceptance cuts and $W$-jet identification selections which are described below. In the purity evaluation the $m_{lvj}$ of the final decay system is required to be between 550-1500 GeV.

Cuts are also applied on the jet substructure to further discriminate jets coming from a hadronically decaying $W$ boson from QCD jets, originating from quarks and gluons. The primary observable used to separate jets originating from $W$ boson decays from QCD ones is the jet mass itself, with improved separation between signal and background by means of a jet grooming algorithm [67]. In particular, for this analysis the pruning implementation has been chosen [68], as it is specifically designed to remove soft QCD and pileup contributions from the jet. In the pruning technique a jet is reclustered using all the particles used to build a CA8 jet, ignoring in each recombination step the softer "protojet" if the recombination is softer than a given threshold $z_{\text{cut}} = 0.1$ and forms an angle $\Delta R$ wider than $D_{\text{cut}} = 0.5 \times m_{\text{orig}}/p_T^{\text{orig}}$ with respect to the previous step. $m_{\text{orig}}$ and $p_T^{\text{orig}}$ are the mass and transverse momentum of the original CA8 jet.
The hardness $z$ of a recombination is defined as $z = \min(p_T^i, p_T^j) / p_T^0$, where $p_T^i$ and $p_T^j$ are the $p_T$ of the two protojets to be combined and $p_T^0$ is the $p_T$ of the combination of the two protojets. Fig. 1, left, shows the invariant mass of the W boson candidates, after the analysis selections described in this Section, for a signal SM Higgs boson of 600 GeV and for the simulated background of W+jets. Dashed lines correspond to the distributions before the application of the pruning algorithm, while continuous lines after pruning. Jets reconstructed from W$_{\text{had}}$ show a distinctive peak near the W mass, which is narrowed and centred in scale by the pruning, while the background jets get correctly assigned a smaller mass on average, enhancing the separation between the two samples.

In addition to the pruned jet mass, we use information about the jet shape to further reduce QCD jets coming from the background. N-subjettiness was introduced in [69] and is a generalized jet shape observable which defines a measure, $\tau_N$, for a jet to have N subjets. For N candidate subjets of a given jet, we can define the N-subjettiness observables as:

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min\{\Delta R_{1,k}, \Delta R_{2,k}, \cdots, \Delta R_{N,k}\}$$

(3)

where $k$ runs over all constituent particles. The normalization factor is $d_0 = \sum_k p_{T,k} R_0$ and $R_0$ is the original jet distance parameter. The $\tau_N$ observable has a small value if the jet is consistent with having N or less subjets. Thus, for boosted W identification, i.e. for discrimination between W-jets with 2 subjets and QCD-jets consistent with 1 subjet, the ratio $\tau_2/\tau_1$ is of particular interest and tends to smaller values for signal W-jets. The subjet axes are obtained by running the exclusive $k_T$ algorithm [70], reversing the last N clustering steps. The axes can be optimized to minimize the N-subjettiness value. As a default definition for the axes, we use a “one-pass” optimization of the exclusive $k_T$ axes, where one step of the iterative optimization is performed. In Fig. 1, the $\tau_2/\tau_1$ distributions can be seen after selecting a jet to have a pruned mass from 65-105 GeV. In this figure, one can see that $\tau_2/\tau_1$ provides additional discrimination between signal and background. Two signal masses are shown, 600 GeV and 1000 GeV, where the difference between the two is small. We cut on the N-subjettiness variable $\tau_2/\tau_1$, choosing the jet axis via the one-pass $k_T$ optimization, in order to further reduce background contributions.

The combination between pruned jet mass and $\tau_2/\tau_1$ selections are exclusively used to identify the hadronically decaying W boson, W$_{\text{had}}$. A jet passing this criteria is referred to as a W-tagged jet. The final WW system is defined using the four-vector of $W_\ell$ and the ungroomed four-vector of W$_{\text{had}}$, since no dedicated correction for groomed jets are currently available in CMS. Jet substructure approach is fundamental in order to boost the sensitivity in the considered final state. These techniques have been already used in CMS for both new physics searches in the VV semi-leptonic final state [71] and W+jets differential cross section measurement [67]. Eventually, detailed studies on hadronically decaying W bosons identification and jet substructure performances have been performed on data in an enriched semi-leptonic $t\bar{t}$ sample, as reported in [72].

After applying the kinematic selections, with no jet substructure cuts except a loose cut on the pruned jet mass in the range 40-130 GeV, we make comparisons of the data and MC contributions. In Fig. 2, we show the $W_\ell$ $p_T$ (left) and the W$_{\text{had}}$ $p_T$ (right) distribution for events in the 0+1 jet category for the muon channel. In Fig. 3, before cuts to isolate the VBF production, we show the absolute value $|\Delta p_T|$ (left) and the $M_{jj}$ (right) distributions for events in the 2 jet category for the muon + jets final state. We see that in the 0+1 jet bin case, the background is predominantly coming from W+jets, while, for the 2 jet bin case, there is a more even fraction of W+jets and $t\bar{t}$ backgrounds although the $t\bar{t}$ is proportionally reduced further when the cuts
Figure 1: Signal and background distributions from simulation of pruned jet mass (left) and \( \tau_2/\tau_1 \) (right) after analysis level cuts described in Section 6. On the left plot, we also show the ungroomed jet mass as dotted lines to underline the effect of pruning.

on \( m_{W_\text{had}} \) and \( m_{W_\ell} \) are added.

Figure 2: Leptonic W \( p_T \) (left) and CA8 jet \( p_T \) (right) for the muon channel in the 0+1 jet bin category.

7 Background and Signal Estimation

The final discriminating variable in the analysis is the shape of the three-body \( m_{\ell\nu} j \) distribution. The signal region is defined around the W boson mass: a jet is considered a W-jet candidate if its pruned mass \( m_J \), computed from the sum of the four-momenta of the constituents surviving the pruning, falls in the range \( 65 < m_J < 105 \) GeV.

The signal normalization and shape are estimated from simulation, with data-to-MC correc-
7.1 Non-dominant backgrounds and signal estimation

Since the jet substructure reproduced in Monte Carlo (MC) events depends on the details used in the simulation, a sample of $W_{\text{had}}$ have been isolated to study the effect of jet selections in data. By applying all analysis requirements, but requiring at least one b-tagged jet in the event, a sample of $W$ boson, decaying hadronically into a single jet, can be isolated in a region of nearly pure $t\bar{t}$ and single-top events. We use this top-enriched control sample to validate our $W$-jet tagging method.

From the comparison between data and MC, a normalization correction factor for $t\bar{t}$ and single top is evaluated in the signal region. For the 0+1 jet category, the scale factor is measured to be $0.91 \pm 0.08$ ($0.89 \pm 0.08$) in the muon (electron) channel. For the 2 jet category, where we merge electron and muon events to increase the statistics, the scale factor is measured to be $1.09 \pm 0.25$, where the larger uncertainty comes from a reduced amount of statistics after applying all VBF category cuts. We check that as we apply successive cuts in the 2 jet category, we find that the various scale factors are always consistent within statistical errors and with the scale factor derived in the 0+1 jet bin.

A simultaneous fit to the jet mass distributions for the events passing and failing the jet mass and $t_2/t_1$ selections is used to extract a data-to-MC efficiency scale factor for identifying merged $W$ bosons. The scale factor for $W$-tagging is $0.93 \pm 0.09$ [72]. In addition, the $W_{\text{had}}$ mass peak and resolution are extracted from the same fit and are measured to be $82.7 \pm 0.3$ GeV and $7.6 \pm 0.4$ GeV in simulation and $84.1 \pm 0.4$ GeV and $8.4 \pm 0.6$ GeV in data. The larger resolution
in data with respect to that of the simulation is in agreement with past CMS measurements of jet energy resolution [73].

The SM-like and BSM Higgs hypotheses are modelled using a Crystal Ball function in the ggH channel, while a composite model, given by a Crystal Ball plus an exponential shape, is considered in the VBF production mode. The latter model is able to account for large interference effects between the signal and the SM background, which increases with the mass of the considered Higgs-like resonance. The shape, normalization and efficiency scale factors all are applied to the signal in addition to the WW/WZ/ZZ background.

7.2 W+jets background estimation

We estimate the main W+jets background contribution from data via a sideband method, in order to extract both its shape and normalization. As previously mentioned, the signal region (SR) is defined around the W boson mass, requiring the pruned jet mass to fall within \( m_J = [65-105] \) GeV, while an event falls in the lower sideband (LSB) if \( m_J \) has a value in [40-65] GeV, or in the upper sideband region when \( m_J = [105,130] \) GeV.

The background estimation procedure, described below, is the same for both the 0+1 and 2 jet categories; the only difference is that in the 2 jet bin we combine the statistics from the electron and muon channels in order to reduce the statistical uncertainty related to sideband fit extrapolation. We verify that the shapes are the same between the two lepton flavour channels in the 2 jet bin, so that we can parametrize them with a single analytical function.

The W+jets normalization is obtained from a sideband fit of data in the \( m_J \) spectrum, as shown in Fig. 4 for the 0+1 jet bin (left) and 2 jet bin (right), where data events inside the signal region are not used in the fit. All the background contributions, except for W+jets, are taken from the simulation and their distributions are parametrized with models determined with dedicated fits on each simulated sample and checked, when possible, in the top-enriched control region. The uncertainty band includes contributions from the fit parameter errors and normalization uncertainties for the other background contributions, detailed in Sec. 8.

![Figure 4: Data and MC distribution for the \( m_J \) sideband and signal region is shown for the 0+1 (2) jet category on the left (right). The W+jets normalization is extracted from a fit to the sideband region excluding the signal one.](image)

The W+jets shape in the signal region is determined from data in the lower sideband, through
an extrapolation function $\alpha_{\text{MC}}(m_{\ell\nu j})$ derived from the W+jets simulation, defined as:

$$\alpha_{\text{MC}}(m_{\ell\nu j}) = \frac{F_{\text{MC,SR}}(m_{\ell\nu j})}{F_{\text{MC,LSB}}(m_{\ell\nu j})}$$  \hspace{1cm} (4)$$

where the function $\alpha_{\text{MC}}(m_{\ell\nu j})$ is determined from MC in order to account for the correlations between the jet mass and the mass of the three-body system, while $F_{\text{MC,SR}}(m_{\ell\nu j})$ and $F_{\text{MC,LSB}}(m_{\ell\nu j})$ are the probability density functions used to describe the $m_{\ell\nu j}$ spectrum in simulation for the signal region and low sideband region, respectively.

The HERWIG sample is used as benchmark, as this generator is known to better describe the jet substructure [67]. The shape of $\alpha_{\text{MC}}(m_{\ell\nu j})$ can be seen in Fig. 5: the 0+1 jet category (left) and 2 jet category (right). The red and blue dashed-dot lines are the fits of the $m_{\ell\nu j}$ shape in the signal and in the low sideband region, respectively. Their ratio is given by the black line and the black (green) shaded regions correspond to the 1σ (2σ) bands of the fit parameters; the structure of the uncertainty band is due to the functional form that is chosen. For the 0+1 jet bin, the difference between the HERWIG and PYTHIA determinations is taken as an additional systematic uncertainty due to the choice of the alternate parton shower model, labelled as “Alternate PS” in Fig. 5. The effect from parton shower modelling is minimal in the 2 jet bin due to larger statistical errors. The systematic uncertainty associated with the chosen fit function for the sideband and signal regions is also considered and is shown as “Alternate Function” in Fig. 5. In addition, systematic effects due to uncertainties in the jet energy scale and resolution measurements are also considered and shown in Fig. 5.

The W+jets shape in the signal region is then extrapolated from the lower sideband through the $\alpha_{\text{MC}}(m_{\ell\nu j})$ function:

$$F_{\text{data,SR}}(m_{\ell\nu j}) = \alpha_{\text{MC}}(m_{\ell\nu j}) \times F_{\text{data,LSB}}(m_{\ell\nu j})$$  \hspace{1cm} (5)$$

Fig. 6 shows the $m_{\ell\nu j}$ distribution in the LSB for data and MC, for the 0+1 jet category on the left and 2 jet category on the right. The simulated distributions are parametrized with
functions determined with dedicated fits on each single simulated sample. The non-dominant backgrounds are fixed to MC expectations corrected by data-to-MC scale factors, while the W+jets parameters are left free to float and the uncertainty bands reflect the fit parameter errors. This latter component is then scaled according to Eq. 5 to get the W+jets shape in the signal region.

![Figure 6: Distributions of the LSB region for data and MC for 0+1 jet category (left) and 2 jet category (right).](image)

In Fig. 7, the final \(m_{\ell\nu_j}\) distributions in the signal region are shown for the 0+1 jet category: muon channel on the left, electron channel on the right. The final \(m_{\ell\nu_j}\) distribution in the signal region for the 2 jet bin is shown in Fig. 8. The uncertainty band takes into account both the fit uncertainty on the \(m_{\ell\nu_j}\) spectrum in the LSB region, the uncertainty on the alpha function, as shown in Fig. 5, and the normalization uncertainties for all the background contributions, which are detailed in Sec. 8.

![Figure 7: Distributions in the signal region for data and background expectation for 0+1 jet category: muon channel (left) and electron one (right).](image)
8 Systematic uncertainties

In this section, we discuss in details the systematic uncertainties on background, Section 8.1, and signal sources, Section 8.2, considered in the analysis.

8.1 Systematic effects on the background

The systematic uncertainty on the W+jets background normalization is dominated by the fit uncertainty associated with the number of data events in the $m_J$ sideband regions; residual bias due to the functional parametrization is also taken into account and added in quadrature to the statistical uncertainty. It is 5(8)% in 0+1 jet category for muon (electron), while at the level of 25% in the VBF channel.

The uncertainty on the $t\bar{t}$ normalization comes from the uncertainty on the data-to-simulation scale factors derived in the top-enriched control sample, which is estimated to be 6.5% in the 0+1 jet bin and 26.5% in the 2 jet one.

Systematic uncertainties coming from single lepton trigger and lepton identification efficiencies are derived thorough dedicated tag and probe studies on $Z \rightarrow \mu^+\mu^-$ and $Z \rightarrow e^+e^-$ events; they are, respectively, equal to 1%(1%) and 2%(2%) for $\mu(e)$ final states. The uncertainties coming from lepton energy scale and resolution are found to be small, less than 2%, for all the background samples, as well as the uncertainty due to the dedicated data-to-MC scale factor used for the b-tagging (< 3.5%).

Systematic uncertainties coming from jet energy scale and resolution are summarized for each background component in Table 2.

The theoretical uncertainties on the WW inclusive cross section is assigned to be 20%, taken from the relative difference in the mean value between the recent CMS cross section measurement at $\sqrt{s} = 8$ TeV and the SM expectation [74]. An additional systematic uncertainty on the WW normalization comes from the uncertainty on the W-tag scale factors, which is reported in Section 7.1. The same uncertainties derived for WW are also used for WZ, ZZ and WW electroweak production.
8.2 Systematic effects on the signal

The systematic uncertainties on the W+jets \( m_{\ell \nu j} \) shape in the signal region come from two separate contributions, as indicated in Eq. 5: the uncertainty on the extrapolation function \( \alpha_{MC}(m_{\ell \nu j}) \) and the uncertainty on the parameters of the fit to the \( m_{\ell \nu j} \) spectrum in the low-mass sideband region. The uncertainty on \( \alpha_{MC}(m_{\ell \nu j}) \), computed using the default HERWIG W+jets simulated sample, is related to the uncertainties on the \( m_{\ell \nu j} \) fit parameters for the numerator and the denominator of the ratio and it is shown in Fig. 5 as a function of the reconstructed invariant mass. The pink and the yellow dashed lines denote, respectively, the \( \alpha_{MC}(m_{\ell \nu j}) \) functions derived from the alternative parton shower model (PYTHIA) and an alternative fit parametrization. To account for these additional shape variations, in the statistical analysis we increase the HERWIG-based uncertainties (shaded bands) by a factor of 2 for all the mass points. In the same way, in order to reduce potential bias in the final limit extraction due to a possible wrong description of the W+jet shape, we increase also by a factor 2 the uncertainty band on the W+jet shape fitted in sideband region.

### Table 2: List of systematic uncertainties on background normalisation: left part of the table refers to 0+1 jet bin, right to 2-jet bin category.

<table>
<thead>
<tr>
<th>Syst. uncertainty</th>
<th>( W+\text{jets} )</th>
<th>( \text{tt} )</th>
<th>single ( t )</th>
<th>( \text{VV} )</th>
<th>( W+\text{jets} )</th>
<th>( \text{tt} )</th>
<th>single ( t )</th>
<th>( \text{VV} )</th>
<th>( \text{WW}_{ewk} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminosity</td>
<td>-</td>
<td>2.6%</td>
<td>2.6%</td>
<td>2.6%</td>
<td>-</td>
<td>-</td>
<td>2.6%</td>
<td>2.6%</td>
<td>2.6%</td>
</tr>
<tr>
<td>Bkg. Cross Sec.</td>
<td>-</td>
<td>-</td>
<td>30%</td>
<td>20%</td>
<td>-</td>
<td>-</td>
<td>30%</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>Trigger Eff.</td>
<td>-</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
<td>-</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Lepton Eff.</td>
<td>-</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>-</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>B-Tagging</td>
<td>-</td>
<td>1.7%</td>
<td>3.3%</td>
<td>0.6%</td>
<td>-</td>
<td>1.5%</td>
<td>3%</td>
<td>0.5%</td>
<td>0.7%</td>
</tr>
<tr>
<td>W-Tagging</td>
<td>-</td>
<td>-</td>
<td>9.3%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.3%</td>
<td>9.3%</td>
<td>-</td>
</tr>
<tr>
<td>Top Normalization</td>
<td>-</td>
<td>6.5%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26.5%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( W+\text{jet} ) Normalization</td>
<td>5-8%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lepton Scale</td>
<td>-</td>
<td>-</td>
<td>0.4%</td>
<td>1%</td>
<td>-</td>
<td>-</td>
<td>0.5%</td>
<td>1.5%</td>
<td>1%</td>
</tr>
<tr>
<td>Lepton Res.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jet Scale (JES)</td>
<td>2.7%</td>
<td>4%</td>
<td>4.1%</td>
<td>3%</td>
<td>2.1%</td>
<td>4.1%</td>
<td>7.1%</td>
<td>7.5%</td>
<td>4.6%</td>
</tr>
<tr>
<td>Jet Res. (JER)</td>
<td>1%</td>
<td>0.4%</td>
<td>0.9%</td>
<td>0.7%</td>
<td>1.9%</td>
<td>3.1%</td>
<td>8.3%</td>
<td>4.3%</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

8.2 Systematic effects on the signal

The systematic uncertainties on the signal cross section upper limits includes the uncertainty on the integrated luminosity of the sample (2.6%) and the uncertainties on the signal efficiency. Some of the systematics sources on the signal efficiency are: the electron and the muon energy scale and resolution, the jet energy scale and resolution. The event selection is applied to signal samples after varying the lepton and jet four-momenta within one respective standard deviation (or applying an appropriate energy/momentum smearing in case of resolution uncertainties) of the corresponding uncertainty on energy/momentum scale and resolution. In this process, variations on the lepton and jet four-momenta are propagated consistently to the \( E_T^{\text{miss}} \) vector. The signal efficiency is then re-calculated separately for each source of systematics. The largest relative change in signal efficiency, compared to the default value, is taken as the systematic uncertainty for that specific source.

We also include systematics in signal efficiency due to uncertainties on data-to-simulation scale factors for the W-tag identification (derived from the top-enriched control sample, Section 7.1), lepton trigger/identification/isolation (derived from \( Z \rightarrow \ell^+ \ell^- \) events), and b-tag identification efficiencies.

The source of systematic uncertainties on the \( m_{\ell \nu j} \) signal shape are the uncertainties on the scale and the resolution of the jet energy, electron energy, muon momentum and interference correc-
### Table 3: List of systematic uncertainties on signal (ggH and VBF) normalisation: left part of the table refers to 0+1 jet bin, right to 2-jet bin category. (* stands for mass dependent systematics)

<table>
<thead>
<tr>
<th>Syst. uncertainty</th>
<th>ggH</th>
<th>VBF</th>
<th>ggH</th>
<th>VBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminosity</td>
<td>2.6%</td>
<td>2.6%</td>
<td>2.6%</td>
<td>2.6%</td>
</tr>
<tr>
<td>PDF gg</td>
<td>-</td>
<td>9.1%</td>
<td>-</td>
<td>9.1%</td>
</tr>
<tr>
<td>PDF qq</td>
<td>-</td>
<td>5%</td>
<td>-</td>
<td>5%</td>
</tr>
<tr>
<td>ggH0In</td>
<td>26%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ggH2In</td>
<td>6%</td>
<td>-</td>
<td>19%</td>
<td>-</td>
</tr>
<tr>
<td>Int ggH</td>
<td>10%</td>
<td>-</td>
<td>10%</td>
<td>-</td>
</tr>
<tr>
<td>Int vbfH</td>
<td>-</td>
<td>10%</td>
<td>-</td>
<td>10%</td>
</tr>
<tr>
<td>Trigger eff.</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Lepton eff.</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>B-Tagging</td>
<td>0.5%*</td>
<td>0.2%*</td>
<td>0.5%*</td>
<td>0.2%*</td>
</tr>
<tr>
<td>W-Tagging</td>
<td>9.3%</td>
<td>9.3%</td>
<td>9.3%</td>
<td>9.3%</td>
</tr>
<tr>
<td>Lepton Scale</td>
<td>2.1%*</td>
<td>1.5%*</td>
<td>3.5%*</td>
<td>1.8%*</td>
</tr>
<tr>
<td>Jet Scale (JES)</td>
<td>3.9%*</td>
<td>4.4%*</td>
<td>5.0%*</td>
<td>4.5%*</td>
</tr>
<tr>
<td>Jet Res (JER)</td>
<td>2.5%*</td>
<td>3.5%*</td>
<td>8.0%*</td>
<td>10.6%*</td>
</tr>
</tbody>
</table>

A summary of the impact of all sources of systematic uncertainty on the signal is presented in Table 3. The dominant systematic uncertainty on the signal normalisation is related to the theoretical uncertainty on interference with SM processes (ggWW [57] and qqWW) and the theoretical uncertainty on ggH cross section in exclusive jet bins. The dominant systematic uncertainty on the $m_{\ell \nu j}$ signal shape comes from the jet energy scale and resolution: approximately 3%(2%) on the position for the 0+1 jet bin and 4% (3%) for the VBF case, 2%(5%) on the width for 0+1 jet bin, 5%(10%) on 2 jet bin.

### 8.3 Bias from fit function choice

Several cross-checks of the choices on the functional forms of the fit shapes have been performed. For the $m_{lj}$ fit, the fit of the LSB region and for the signal shape choice, we check the effect of alternate shapes on the analysis results. In the $m_{lj}$ fit, we generate pseudo-experiments with one function and perform the fits with the alternate (and vice versa) and the residual bias on the W+jets normalization from this procedure is added to the systematic uncertainty. For the LSB fit and signal fit, we again generate pseudo-experiments with one function and perform the fits with the alternate shape and check the final effect on the yield of the background and signal, as well as the effect on the expected exclusion sensitivity. We find that the bias is typical less than 10% for all the categories and for the whole investigated mass range.

### 9 Results

Based on the normalization and shape of the $m_{\ell \nu j}$ for signal and background reported in Fig. 7 and in Fig. 8, we set upper limits on the SM-like Higgs boson production cross section. The three categories are used simultaneously in an unbinned maximum likelihood fit, using the modified frequentist asymptotic CL$_S$ [75, 76] with profile likelihood as the test statistic in com-
puting the exclusion limits. Systematics uncertainties are included in the frequentist statistical interpretation as log-normal constrained nuisance parameters. Exclusion limits at a 95% confidence level are presented on the production cross section times branching fraction for the Higgs boson compared to the SM expectation in Fig. 9: (left) the exclusion limit obtained from the combination among ggH and VBF channels, (right) the comparison between the combined sensitivity and the one coming from the individual 0+1 and 2 jet bin categories. An expected sensitivity to exclude the SM-like Higgs boson varies from 1.1 times the SM cross-section at 600 GeV to 3.3 times the SM cross-section at 1000 GeV. Local excesses of 2.64 $\sigma$ and 2.56 $\sigma$ are observed in the mass range between 700 and 800 GeV.

Figure 9: (Left) Observed (solid) and expected (dashed) 95% CL upper limit on $\sigma/\sigma_{SM}$, obtained via asymptotic CL$_s$ technique, for a SM-like Higgs boson decaying to WW -> lνqq'. The 68% and 95% ranges of expectation for the background-only model are also shown with green and yellow bands, respectively. The solid horizontal red line at unity indicates the expectation for a SM-like Higgs boson. (Right) Comparison between exclusion limits in each bin category.

9.1 Electroweak singlet interpretation

The exclusion limit in the BSM electroweak singlet heavy Higgs has been investigated only in limited parameter space, where $C' \leq \sqrt{1 - BR_{new}}$, in order to remain in the region where the resonance width is narrower than the SM-like Higgs ones ($\Gamma \leq \Gamma_{SM}$). Since the maximum value considered for $BR_{new}$ is 0.5, $C'$ is constrained to be less than 0.7. In Fig.10, we show the 95% expected and observed exclusion contours for three times the expected theoretical signal strength in the [$m_H, C'^2$] plane.
Figure 10: Contours of the observed (expected) limits at the 95% CL in the \([m_{H}, C'^2]\) plane. The solid (dashed) lines depict the observed (expected) exclusion limits corresponding to \(\sigma/\sigma_{th} = 3\). The pink horizontal line is used to indicate the upper bound of \(C'^2\) and \(BR_{new}\) considered in the analysis.

Acknowledgments

We wish to congratulate our colleagues in the CERN accelerator departments for the excellent performance of the LHC machine. We thank the technical and administrative staff at CERN and other CMS institutes, and acknowledge support from: FMSR (Austria); FNRS and FWO (Belgium); CNPq, CAPES, FAPERJ, and FAPESP (Brazil); MES (Bulgaria); CERN; CAS, MoST, and NSFC (China); COLCIENCIAS (Colombia); MSES (Croatia); RPF (Cyprus); Academy of Sciences and NICPB (Estonia); Academy of Finland, MEC, and HIP (Finland); CEA and CNRS/IN2P3 (France); BMBF, DFG, and HGF (Germany); GSRT (Greece); OTKA and NKTH (Hungary); DAE and DST (India); IPM (Iran); SFI (Ireland); INFN (Italy); NRF and WCU (Korea); LAS (Lithuania); CINVESTAV, CONACYT, SEP, and UASLP-FAI (Mexico); MSI (New Zealand); PAEC (Pakistan); SCSR (Poland); FCT (Portugal); JINR (Armenia, Belarus, Georgia, Ukraine, Uzbekistan); MST, MAE and RFBR (Russia); MSTD (Serbia); MICINN and CFAN (Spain); Swiss Funding Agencies (Switzerland); NSC (Taipei); TUBITAK and TAEK (Turkey); STFC (United Kingdom); DOE and NSF (USA). Individuals have received support from the Marie-Curie programme and the European Research Council (European Union); the Leventis Foundation; the Alexander von Humboldt Foundation; the Belgian Federal Science Policy Office; the Fonds pour la Formation à la Recherche dans l’Industrie et dans l’Agriculture (FRIA-Belgium); the Agentschap voor Innovatie door Wetenschap en Technologie (IWT-Belgium); and the Council of Science and Industrial Research, India.

References


