Search for top-Higgs resonances in all-hadronic final states using jet substructure methods

The CMS Collaboration

Abstract

This note presents a search for heavy resonances decaying to top quarks and Higgs bosons in all-hadronic final states. The benchmark model is pair-produced vector-like T quarks with charge 2/3. High masses of the T quarks are assumed so that the decay products are highly Lorentz-boosted, such that jets can overlap and get merged. Dedicated methods such as top-tagging and Higgs-tagging algorithms are applied to resolve the substructure of boosted jets. Upper limits on the production cross section of a T quark with masses between 500 GeV/$c^2$ and 1000 GeV/$c^2$ are derived. If the heavy T quark has an exclusive branching ratio BR($T \rightarrow tH$) = 100%, the observed (expected) lower limit on the mass of the T is 747 GeV/$c^2$ (701 GeV/$c^2$) after analysing 19.7 fb$^{-1}$ of integrated luminosity of proton-proton collision data collected at $\sqrt{s} = 8$ TeV with the CMS detector.
1 Introduction

The discovery of the Higgs boson with a mass of 126 GeV motivates the search for exotic states involving the newly-discovered particle. While the mechanism that stabilises the mass of the Higgs particle is still a mystery, possible explanations are given by little Higgs models [1, 2], extra dimensions [3, 4] and composite Higgs models [3–5]. All of these models predict the existence of heavy vector-like quarks which may decay into top quarks and Higgs bosons. The left-handed and right-handed vector-like quarks transform in the same way under the Standard Model (SM) gauge groups $SU(3)_c \times SU(2)_L \times U(1)_Y$.

A chiral fourth generation of fermions, replicating one of the three generations of the Standard Model with identical quantum numbers, is disfavoured by the electroweak fits within the Standard Model framework [6] due to its large modifications of the Higgs production cross sections and branching ratios, if a single SM-like Higgs doublet is assumed. However, vector-like heavy quarks do not obtain their mass via the Yukawa couplings as the other quarks do, and are thus not similarly constrained, yet naturally resolve the hierarchy problem [7]. In this note, we report about a search for exotic resonances decaying into Higgs bosons and top quarks. We use the model of vector-like T quarks, which are produced in pairs by the strong interaction, as a benchmark for the sensitivity of the analysis.

Vector-like T quarks can decay into three different final states: $tH$, $tZ$ and $bW$ [7]. The assumption of exclusive decays with 100% branching ratio has been used in various searches by ATLAS and CMS [8–11]. However, realistic models allow mixed branching ratios, so that searches have been performed independently of the branching ratios as well [12]. In the present analysis, we optimise our event selection to be fully sensitive to an exclusive decay with $BR(T \to tH) = 100\%$. In addition, we quote the results as functions of the branching fractions.

In the Standard Model the Higgs boson decays predominantly into $b$-quark pairs with a branching fraction of 56% for a mass of 126 GeV, while the top quark decays almost exclusively into a $b$ quark and a $W$ boson, which itself decays hadronically in 66% of the cases. The main final state is therefore the all-hadronic final state $T \to tH \to b\bar{b}bjj$ where $j$ denotes the light-flavour jets of the $W$ boson decay and $b$ denotes the $b$-flavour jets from the top quark or Higgs boson decays. As the mass of the T quark can be very large, the decay products can be highly-boosted, leading to final states with overlapping and merged jets. In the extreme case, all top-quark decay products are merged into a single jet. A similar topology may arise for the Higgs boson decaying into $b$ quarks. In recent years, the technology of jet substructure analysis has proved to be very powerful in resolving such difficult boosted topologies [13–16]. For example, the analysis of high-mass $Z'$ resonances decaying into top quarks became feasible in the all-hadronic final state due to the application of jet substructure methods [17–19]. We follow a similar strategy in this analysis by applying top-tagging and Higgs-tagging algorithms in combination with $b$-tagging algorithms.

While searches for T quarks have been performed in leptonic final states [8–12], we present the first analysis that exploits the all-hadronic final state in the search for vector-like quarks. The development and commissioning of the jet substructure methods which are applied in this analysis are documented in [20] and [21]. In particular, the application of $b$-tagging algorithms in subjets has made the identification of boosted $H \to b\bar{b}$ possible and reduced the background to boosted $t \to bW$ significantly so that the all-hadronic final states are becoming accessible and may contribute to the overall combination with the leptonic final states to obtain the best possible sensitivity.
2 CMS detector

A detailed description of the CMS experiment can be found in Ref. [22]. CMS is a general-purpose detector that uses a silicon tracker, as well as finely segmented lead-tungstate crystal electromagnetic (ECAL) and brass/scintillator hadronic (HCAL) calorimeters. These subdetectors have full azimuthal coverage and are contained within the bore of a superconducting solenoid that provides a 3.8 T axial magnetic field. The tracker consists of silicon pixel and strip detector modules and is fully contained in the magnetic field that enables the measurement of charged particle momenta over the pseudorapidity range $|\eta| < 2.5$, where $\eta = \ln \tan(\theta/2)$ and $\theta$ is the polar angle of the track relative to the counterclockwise beam direction. The electromagnetic calorimeter consists of nearly 76,000 lead tungstate crystals which provide coverage in pseudorapidity $|\eta| < 1.48$ in a cylindrical barrel region and $1.48 < |\eta| < 3.0$ in two endcap regions. The hadronic calorimeter (HCAL) is made of brass and scintillators covering $|\eta| < 3.0$ to measure jets. Muons are identified in the range $|\eta| < 2.4$ by gas-ionisation detectors embedded in the steel return yoke. The first level of the CMS trigger system consists of custom hardware processors and uses information from the calorimeters and muon system to reduce the event rate down to 100 kHz. The high level trigger (HLT) processor farm further decreases the event rate to around 300 Hz before data storage.

3 Event Reconstruction

The events are reconstructed with the particle-flow algorithm [23, 24], which takes into account information from all subdetectors, including charged particle tracks from the tracking system and deposited energy from the electromagnetic and hadronic calorimeters. Given this information all particles in the event are classified into mutually exclusive categories: electrons, muons, photons, charged hadrons, and neutral hadrons. The resulting particle flow candidates are passed to jet clustering algorithms to reconstruct jets, using an implementation based on FastJet 3.0 [25].

Tracks are reconstructed using an iterative tracking procedure [26]. The primary vertex is reconstructed from all tracks in the event that are compatible with the beam spot, the location of the LHC beam in the $x, y$ plane. This is done by searching for clusters of tracks with similar $z$ coordinates at their point of closest approach to the beam-line using a deterministic annealing method [27]. The primary vertex with the highest $\sum (p_{T}^{\text{track}})^2$ is selected to define the primary interaction vertex, whose position is determined from an adaptive vertex fit [28].

The primary vertex and tracks are used in the analysis by the b-tagging algorithms. The algorithm for b-jet identification used in the analysis is the Combined Secondary Vertex (CSV) algorithm [29], which uses information based on reconstructed tracks and secondary vertices that are combined into a single discriminating variable. The medium operating point of the algorithm (CSVM) is used, corresponding to a minimum requirement on the CSV discriminator of 0.679.

4 Event samples

Simulated samples are used to determine signal selection efficiencies and the contribution of the background due to $t\bar{t}$ plus jets, $t\bar{t}H$ and $W/Z$ plus heavy flavour production in which the $W/Z$ bosons decay hadronically.

The $t\bar{t}$ background sample is generated with MADGRAPH 5 [30], interfaced with PYTHIA 6 [31] to simulate the parton shower and the hadronization. The signal $T\bar{T}$ samples are generated with
MADGRAPH 5, using the CTEQ6L1 parton distribution functions [32], interfaced to PYTHIA 6. The Z2star tune is used to simulate the LHC underlying event [33]. The allowed decay modes are $T \rightarrow tH$, $T \rightarrow tZ$, $T \rightarrow bW$. The signal mass points used in this analysis are listed in Table 1. The mass of the Higgs boson in these samples is set to 120 GeV/$c^2$. The branching ratios of the Higgs boson decays are corrected to the expected values for a Higgs boson with a mass of 126 GeV/$c^2$ using the recommendations from Ref. [34]. The decay into b quarks $H \rightarrow b\bar{b}$, which is the main search channel of this analysis, has a branching ratio of 56%.

The total inclusive cross sections for the $t\bar{t}$ and signal samples are calculated with exact NNLO and full NNLL soft gluon resummation [35] which are computed based on the Top++2.0 implementation using the MSTW2008nnlo68cl PDF and the 5.9.0 version of LHAPDF [36, 37]. The $t\bar{t}$ sample has been normalised to the NNLO cross section of 245.8 pb [36].

The background for QCD multijet production is determined from data. Simulated QCD samples are used to validate the method for the data-driven QCD estimation.

The data are collected at a centre-of-mass energy of $\sqrt{s} = 8$ TeV, corresponding to an integrated luminosity of 19.7 fb$^{-1}$.

5 Analysis strategy

The analysis first applies event selection criteria to reduce the large QCD multijet and $t\bar{t}$ background contributions. These event selection criteria make heavy use of novel jet substructure methods, which allow identification of hadronic decays of high-$p_T$ top quarks and Higgs bosons. Boosted top-quark jets are identified using a top-tagging algorithm, the HEPTopTagger [16]. Boosted Higgs boson jets are identified using a Higgs-tagging algorithm that is based on the identification of the subjet flavour (subjet b-tagging). These methods and the event selection are discussed in detail in Section 6.

Two main variables have been identified which further distinguish the signal from the background events after the event selection. These variables are the scalar sum of the transverse momenta of selected objects $H_T$ and the invariant mass $m_H$ of two b-tagged subjets. These two variables are used in the statistical analysis for setting limits on T production cross sections.

None of the available Monte Carlo simulation samples sufficiently reproduce kinematic distributions of QCD multijet events. In particular, we are interested in the normalisation and shape of the $H_T$ and the $m_H$ distributions. Thus, a data-driven strategy based on signal-depleted sideband regions is adopted. The sideband regions are defined using inverted jet substructure criteria. Closure tests are performed with simulated QCD events to understand whether the prediction for rates and shapes of $H_T$ and $m_H$ are statistically consistent. The background determination is discussed in detail in Section 8.

The analysis results are evaluated based on a combination of the $H_T$ and $m_H$ variables into one discriminator. This combination is done with a simple likelihood ratio method. Based on the expected distributions for the background and signal models for the $H_T$ and $m_H$, a discriminating quantity $L$ is calculated for each event, where

$$L = \ln \left( 1 + \frac{P_{signal}(H_T)}{P_{background}(H_T)} \frac{P_{signal}(m_H)}{P_{background}(m_H)} \right).$$

The $P$ variables represent the probability densities for the signal or background hypotheses. The $P_{background}$ values are obtained from distributions that are created from the expected $t\bar{t}$ background added to the expected QCD multijet background. For the signal hypothesis the
$P_{\text{signal}}$ values are obtained from simulated $H_T$ and $m_H$ distributions for each signal mass point. The $L$ variable can be evaluated in the same way for simulated events and for data.

For the statistical interpretation of the analysis results, two event categories are introduced to enhance the analysis sensitivity: a category with a single Higgs tag and a category with at least two Higgs tags. The categories are denoted as single and multi Higgs tag categories. They are statistically independent and are combined for the final limit setting. The procedure of the limit setting is discussed in detail in Section 10.

6 Substructure methods

The analysis is optimised to select events in which T quarks decay into a top quark and a Higgs boson where top quarks and Higgs bosons decay into fully-hadronic final states. Due to the expected large mass of the T quarks, these top quarks and Higgs bosons can be highly boosted. Therefore daughter particles of the top quarks leave tracks in the detector, which are spatially very close to each other. In many cases all of the top quark’s decay products are clustered into one large jet in the event reconstruction. The same is true for the decay products of the Higgs boson. Figure 1 shows the angular distance defined as $\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}$ between two $b$ quarks in boosted $H \rightarrow b\bar{b}$ decays depending on the transverse momentum of the Higgs boson. This illustrates that for increasing T quark masses the decay products become more collimated so that they cannot be separated anymore with classical jet reconstruction algorithms. To iden-

![Figure 1: Angular distance $\Delta R$ of the two generated $b$ quarks from Higgs boson decays versus the Higgs boson $p_T$. The signal mass hypotheses are 500 GeV/$c^2$ (left) and 1000 GeV/$c^2$ (right.)](image)

ify these so called top jets and Higgs jets, the analysis makes use of dedicated jet substructure tools, namely the HEPTopTagger algorithm and a brand new Higgs-tagging algorithm which relies on $b$-tagging algorithms for subjets. This analysis is pioneering the use of these cutting-edge substructure techniques at CMS.

The HEPTopTagger jets are clustered with the Cambridge-Aachen (CA) algorithm [38, 39] with a distance parameter $R = 1.5$, denoted as CA15 jets. The large cone size makes the algorithm suitable for the phase space with intermediate boosts as it is the case for this analysis. This means that a considerable fraction of signal events do not contain fully-merged jets from top quark or Higgs boson decays, especially for the lower mass points around 500 GeV/$c^2$. Such resolved events could in principle be reconstructed with classical methods using standard anti-$k_T$ jets [40]. The HEPTopTagger provides a seamless transition between the non-boosted domain and the boosted domain which simplifies the analysis.

For each jet, the substructure is analysed by stepping backward through the clustering history of the jet in an iterative procedure until the conditions for splitting are not fulfilled anymore
and the subjets are not split any further. For each combination of three subjets found, a filtering algorithm [13] is applied. The filtering algorithm reclusters the constituents with variable distance parameter $R_{filt} = \min(0.3, \Delta R_{ij}/2)$, where $i$ and $j$ are the closest subjets in $\Delta R$ in the subjet triplet. The five reclustered subjets with the largest $p_T$ are kept and the sum of their masses gives the invariant mass of the top candidate. The selected filtered association is based on the subjet triplet that gives an invariant mass closest to the nominal top quark mass. The constituents of the five leading reclustered subjets are then clustered again using the exclusive CA algorithm, which forces the jet to have exactly three final subjets. The HEPTopTagger uses these three final subjets to identify top jets by making selections on the invariant masses of pairwise and triplet subjet combinations using the top quark and W boson masses as reference. Details about the HEPTopTagger and its commissioning in CMS are given in [20].

The choice of using the HEPTopTagger and the jet size of $R = 1.5$ is driven by the studies in [20] which shows that the HEPTopTagger in combination with subjet b-tagging (described below) outperforms other top-tagging algorithms for the intermediate momentum range of $p_T < 600$ GeV/$c$ which is the case for the present analysis. Only for masses beyond $m_T > 1000$ GeV/$c^2$ would it be necessary to consider alternative algorithms such as the CMSTopTagger [20].

The top-tagging efficiency has been measured in data, and $p_T$-dependent scale factors to correct the efficiency in simulation to the efficiency in data have been derived [20]. As the present analysis uses the same top-tagging setup as in Ref. [20], the scale factors are directly applicable. These scale factors are applied to simulated $t\bar{t}$ and signal events. The HepTopTagger starts to be efficient for jets with $p_T > 200$ GeV and reaches efficiencies of about 35% to 40% for boosted jets with $p_T > 400$ GeV with misidentification rates from QCD of about 1-3%.

The sensitivity of top-tagging algorithms can be significantly improved by the use of b-tagging in boosted topologies, developed and studied in detail in [20] and [21]. These techniques are exploited in the present analysis both for identifying the b jet in boosted top-quark decays (subjet b-tagging) and for identifying two b jets in boosted Higgs boson decays (double subjet b-tagging or Higgs-tagging). The Higgs-tagging is applied to a filtered CA15 jet collections. The subjets are clustered with the Cambridge/Aachen algorithm applying a filtering with a distance parameter $R = 0.3$. Those filtered CA15 jets which have two b-tagged subjets and a di-subjet invariant mass larger than 60 GeV/$c^2$ are considered to be Higgs tagged.

To reconstruct the subjets of all decay products of a heavy object within one fat jet, the size of the fat jet should be as large as possible. On the other hand, large jet sizes lead to overlaps with nearby jets which decrease the selection efficiency, in particular for dense environments as is the case for the signal topology of this analysis. It has been verified that a jet size of $R = 1.2$ for the Higgs-tagging algorithm leads to a slightly increased selection efficiency for the high mass point with $m_T > 1000$ GeV/$c^2$ but to a decreased efficiency for the lower mass points.

The HEPTopTagger and subjet b-tagging used in this analysis are validated in [20] and [21]. In addition, the cut on the invariant mass of the two b-tagged subjets of the Higgs candidate is validated using a sample of semileptonic $t\bar{t}$ events. As no high-statistics sample of Higgs bosons can be obtained in data, the validation procedure is based on the selection of a pure sample of W bosons, which is assumed to be affected similarly by the mass cut, in terms of agreement between data and Monte Carlo. A muon is required and a b-tagged jet clustered using the anti-$k_T$ algorithm [40] with an R parameter of 0.5 (AK5). In addition, one filtered CA15 jet is required to be top-tagged by the HEPTopTagger algorithm and to have exactly one b-tagged subjet. The two subjets which are not b-tagged are used to calculate the invariant mass of a W boson candidate. The selection efficiencies in data and Monte Carlo simulation for a cut of 60 GeV on the invariant mass of the W boson candidate agree within statistical
Event selection

Events are selected online by a trigger algorithm which requires the scalar sum of the transverse momenta of reconstructed jets in the detector ($H_T$) to be greater than 750 GeV. Calorimeter jets, reconstructed exclusively from calorimeter information, with $p_T > 40$ GeV/$c$ are used to calculate the $H_T$ online. The offline analysis has a slightly different $H_T$ definition in which the $p_T$ of all subjets of fat jets is used. The trigger is nearly 100% efficient for events with offline $H_T$ values larger than 800 GeV/$c$. Figure 2 shows the trigger efficiency in simulation and in data. The $H_T$ threshold in the offline analysis is $H_T > 720$ GeV/$c$ where the trigger is not yet fully efficient. The simulation is corrected to match the data by weighting events based on the ratio between the trigger efficiency in data and in simulation. The systematic uncertainty due to this procedure is discussed in Section 9.

Candidate $tH$ events are selected by requiring at least two jets reconstructed from particle-flow objects. CA15 jets with $p_T > 150$ GeV/$c$ are kept for the analysis. At least two ($\geq 2$) of these jets are required, fulfilling the following additional criteria:

- At least one ($\geq 1$) of the CA15 jets must be top-tagged by the HepTopTagger algorithm AND must have a b-tagged subjet (by the CSVM b-tagging algorithm). Top-tagged jets must have $p_T > 200$ GeV/$c$. This jet is referred to as top candidate.
- At least one ($\geq 1$) of the CA15 jets must have $p_T > 150$ GeV/$c$ AND must be Higgs tagged (at least two subjet b-tags by the CSVM b-tagging algorithm). The invariant mass of the two b-tagged subjets has to be larger than 60 GeV. This jet is referred...
Figure 3: Left: Multiplicity of CA15 jets with $p_T > 150$ GeV. Events with at least two of these jets are selected. Right: Multiplicity of CA15 jets with $p_T > 200$ GeV which are selected by the HEPTopTagger algorithm. The coloured histograms show Monte Carlo simulation (tt in yellow and QCD in blue). The hatched regions indicate the statistical uncertainty of the simulated background due to finite size of Monte Carlo samples.

to as a Higgs candidate. The Higgs candidate jet may not be identical to the top-candidate, such that at least two CA15 jets have to be found in the event, one passing the top-tagging criteria and one passing the Higgs-tagging criteria.

Two event categories are used in the analysis. The category with a single Higgs-tag requires exactly one ($= 1$) Higgs-tagged jet, while the category with multiple Higgs-tags requires at least two ($\geq 2$) Higgs-tagged CA15 jets.

The number of reconstructed CA15 jets with $p_T > 150$ GeV/c is shown on the left hand side in Figure 3, while the right hand side shows the number of jets passing the top-tagging criteria.

The impact of subjet b-tagging is visible in Figure 4. The left hand side shows the number of top-tagged CA15 jets with a subjet b-tag, while the right hand side shows the number of Higgs-tagged jets for events which have at least one top-tagged CA15 jet with a subjet b-tag. These figures demonstrate the strong reduction of QCD multijet background due to the substructure criteria.

The number of selected events for each signal sample of the benchmark model and the selection efficiencies, derived from Monte Carlo simulation, are given in Table 1.

8 Background estimation

The tt background is evaluated from Monte Carlo simulation, corrected for known differences in efficiency due to b-tagging and the trigger efficiency difference described above. The uncertainties in the normalisation and shape of tt events are discussed in Section 9. Backgrounds due to ttH production are considered as well and are found to be very small, so that the impact of ttH events is ignored. In addition, backgrounds due to W plus heavy flavour and Z plus heavy flavour production, where the Z and W bosons decay hadronically are simulated and found to have no contribution at all.

The QCD multijet background is measured in data using a two-dimensional sideband extrap-
Figure 4: Left: Multiplicity of CA15 jets with $p_T > 200$ GeV which are tagged by the HEPTopTagger and contain a b-tagged subjet after requiring at least one jet per event to be selected by the HEPTopTagger algorithm. Right: Multiplicity of CA15 jets with $p_T > 150$ GeV containing two b-tagged subjets (Higgs-tagged jets) in events which have at least one top-tagged CA15 jets with a subjet b-tag. Events with three or more Higgs tags are rare and are included in the bin with two Higgs tags. The coloured histograms show Monte Carlo simulation (tt in yellow and QCD in blue). The hatched regions indicate the statistical uncertainty of the simulated background due to finite size of Monte Carlo samples.

Table 1: Cross section, expected number of selected events for 19.7 fb$^{-1}$ and the selection efficiencies for each signal sample of the benchmark model, derived from Monte Carlo simulation. The signal samples assume BR($T \rightarrow tH$) = 100%. The efficiencies are calculated with respect to an inclusive sample without restriction to any top quark and Higgs boson decay modes before application of any selection criteria.

<table>
<thead>
<tr>
<th>signal sample</th>
<th>production cross section</th>
<th>expected events</th>
<th>selection efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_T = 500$ GeV/$c^2$</td>
<td>0.59 pb</td>
<td>283.0</td>
<td>2.0%</td>
</tr>
<tr>
<td>$m_T = 600$ GeV/$c^2$</td>
<td>0.174 pb</td>
<td>152.0</td>
<td>4.4%</td>
</tr>
<tr>
<td>$m_T = 700$ GeV/$c^2$</td>
<td>0.0585 pb</td>
<td>69.3</td>
<td>6.0%</td>
</tr>
<tr>
<td>$m_T = 800$ GeV/$c^2$</td>
<td>0.0213 pb</td>
<td>30.3</td>
<td>7.2%</td>
</tr>
<tr>
<td>$m_T = 900$ GeV/$c^2$</td>
<td>0.0083 pb</td>
<td>12.1</td>
<td>7.3%</td>
</tr>
<tr>
<td>$m_T = 1000$ GeV/$c^2$</td>
<td>0.00336 pb</td>
<td>4.9</td>
<td>7.2%</td>
</tr>
</tbody>
</table>

In this method two uncorrelated criteria in the event selection are inverted to obtain sideband regions that are enriched with QCD multijet events and depleted in signal events. Inverting each criterion individually, as well as both at the same time, results in three exclusive sideband regions.

The sideband region B is obtained by inverting the selection criteria of the HEPTopTagger algorithm. The top mass window as well as all requirements on the pairwise subjet mass in the HEPTopTagger are inverted. Events in regions outside of the “A” shaped region (Figure 11 in [20]) are used to define the inverted HEPTopTagger control region, while the events inside define the signal region.

The sideband region C is obtained by inverting the Higgs-tagging algorithm. Only events with exactly zero Higgs tags are selected and the cut on the pairwise subjet mass is removed.
The sideband region A is obtained by inverting both the Higgs tagging and the top tagging algorithms as described above. The region D is the signal region in which all tagging requirements are applied.

The sideband regions are not completely free of \( t\bar{t} \) background. Therefore, the expected contribution of \( t\bar{t} \) is subtracted from the sidebands. The \( t\bar{t} \) contamination amounts to a maximum of 6% in sideband region C. A signal-injection test has been performed to evaluate the impact of a hypothetical signal on the background model. It has been found that the signal contamination in the sidebands leads to a small effect of less than 1.4% for \( m_T = 700 \text{ GeV}/c^2 \), so that the signal contamination in the sideband can be ignored in the analysis.

The event rates in the three sideband regions and the signal region are given in Table 2. The resulting measurement of the QCD multijet backgrounds are given in Table 3 for the two event categories with one and multiple Higgs-tags.

Table 2: Event rates in the signal and sideband regions as obtained from the two-dimensional sideband extrapolation in data for the two Higgs-tag categories. The \( t\bar{t} \) contamination is subtracted from the nominal yield in the sidebands. Also the prediction of the QCD multijet event rate in the signal region D is given. The quoted uncertainty on the prediction is purely statistical due to the sample sizes in the sideband regions.

<table>
<thead>
<tr>
<th>Single Higgs-tag category</th>
<th>Multi Higgs-tag category</th>
</tr>
</thead>
<tbody>
<tr>
<td>region A</td>
<td>region B</td>
</tr>
<tr>
<td>data</td>
<td>1152640</td>
</tr>
<tr>
<td>( t\bar{t} )</td>
<td>4174.19</td>
</tr>
<tr>
<td>data – ( t\bar{t} )</td>
<td>1148465.81</td>
</tr>
<tr>
<td>region C</td>
<td>region D</td>
</tr>
<tr>
<td>data</td>
<td>140911</td>
</tr>
<tr>
<td>( t\bar{t} )</td>
<td>8202.93</td>
</tr>
<tr>
<td>data – ( t\bar{t} )</td>
<td>132708.07</td>
</tr>
</tbody>
</table>

| prediction | 947 ± 11 | 140911 | 132708.07 | 947 ± 11 |

Table 3: Estimated background contributions in the signal region, for the two event categories with one and multiple Higgs-tags. The uncertainties are statistical only due to the finite size of the control regions and Monte Carlo samples.

<table>
<thead>
<tr>
<th></th>
<th>single Higgs-tag category</th>
<th>multiple Higgs-tags category</th>
</tr>
</thead>
<tbody>
<tr>
<td>QCD (measured in data)</td>
<td>947 ± 11</td>
<td>131 ± 4</td>
</tr>
<tr>
<td>( t\bar{t} ) (from simulation)</td>
<td>420 ± 5</td>
<td>44 ± 2</td>
</tr>
<tr>
<td>sum of backgrounds</td>
<td>1367 ± 12</td>
<td>175 ± 4</td>
</tr>
<tr>
<td>data</td>
<td>1355</td>
<td>205</td>
</tr>
</tbody>
</table>

In addition to the normalisation, also the shapes of the \( H_T \) and \( m_H \) distributions are obtained from sidebands. For both the \( H_T \) and \( m_H \) variables the sideband region B (inverted top-tagger) is used. The expected contribution from \( t\bar{t} \) events is subtracted from the sideband.

The closure of this method is verified with simulated events. As the method assumes the selection criteria defining the sideband regions to be uncorrelated, the following condition has to be fulfilled:

\[
\frac{R_A}{R_S} = \frac{R_C}{R_D},
\]

where \( R_A \) denotes the rate of events in sideband A. According to simulation the ratios are \( \frac{R_A}{R_S} = 185.1 \pm 5 \) and \( \frac{R_C}{R_D} = 185.2 \pm 17 \) for the single Higgs-tag event category, while the ratios are \( \frac{R_A}{R_S} = 1417 \pm 97 \) and \( \frac{R_C}{R_D} = 1203 \pm 250 \) for the multi Higgs-tag event category. The quoted
uncertainties are statistical due to the finite size of Monte Carlo samples. It can be seen that the closure is good within the statistical uncertainties. The largest uncertainties occur in the $R_{CC}$ ratios and they amount to about 10% for the single Higgs-tag category and to 20% for the multi Higgs-tag category. These uncertainties are applied to the estimate of the QCD background normalisation obtained with this data-driven method.

The closure is also verified for the shape of $H_T$ and $m_{H}$ distributions in the signal and sideband regions. Figure 5 shows a comparison of the $H_T$ and $m_H$ shapes in the sideband and signal regions for the single Higgs-tag event category. Figure 6 shows the same for the multi Higgs-tag event category. The agreement is very good within the statistical uncertainties due to the limited size of available Monte Carlo samples.

**Figure 5:** Comparison of $H_T$ (left) and $m_H$ distributions in the sideband region B and signal region for the single Higgs-tag event category for simulated QCD multijet events. Both distributions are normalised to unity for shape comparison.

**Figure 6:** Comparison of $H_T$ (left) and $m_H$ distributions in the sideband region B and signal region for the multi Higgs-tag event category for simulated QCD multijet events. Both distributions are normalised to unity for shape comparison.
9 Systematic uncertainties

As the analysis relies on MC simulation for the $t\bar{t}$ background prediction, a careful evaluation of uncertainties affecting both the normalisation and the shape has to be performed. The same is done for the simulated signal events.

The QCD multijet background is obtained with a data-driven method so that the associated uncertainties are of different nature. The rate and shape of the $t\bar{t}$ background has an impact on the measurement of the QCD background as well because the $t\bar{t}$ contamination in the sideband region is subtracted from data. As the $t\bar{t}$ contamination is small, these uncertainties are relatively small as well. The normalisation and shape of QCD events does not show any discrepancy between prediction and actual shape in the signal region, based on the closure test with simulated events as discussed in Section 8. Nevertheless, the statistical precision of the closure test is limited by the finite size of Monte Carlo samples. Therefore a systematic uncertainty in the normalisation of QCD multijet events is taken into account. The only systematic uncertainty in the shape of the QCD multijet background is the impact of the $t\bar{t}$ subtraction, which can be sizable as discussed in more detail below.

The detailed list of sources of systematic uncertainties is given in the following. Most of these uncertainties have an impact on both the shapes and normalisation of the sensitive variables $H_T$ and $m_H$. Some of them only affect the normalisation.

- **b-tagging scale factor uncertainties**: the agreement between simulation and measured efficiencies in data has been evaluated in Ref. [21]. Based on these measurements scale factors and scale factor uncertainties are applied to simulated samples. The scale factor uncertainties for the b-tagging efficiency depend on $p_T$ and $\eta$. The typical size of these uncertainties is between 1% and 2% while the mistag uncertainty is around 15%. The b-tag scale factor uncertainties affect both the normalisation and shape of $t\bar{t}$ background and signal events. Depending on the sample and signal mass point, the impact of the b-tagging scale factor uncertainty on the expected number of selected signal and $t\bar{t}$ events is 5% to 8% while the impact of the mistag scale factor uncertainty is 0.3% to 4%.

- **HEPTopTagger scale factor uncertainty**: the efficiency of the HEPTopTagger has been measured and compared to Monte Carlo simulation to derive scale factors. Results of this measurement are documented in Ref. [20]. The uncertainties of these scale factor measurements are between 3% and 6% and are parameterized as a function of $p_T$. These uncertainties affect both the normalisation and shape of $t\bar{t}$ background and signal events. The impact on the expected number of signal and $t\bar{t}$ events is 0.4% to 2.3%.

- **top-$p_T$ reweighting**: Monte Carlo events are reweighted using the generator level $p_T$ of the top and antitop quarks so that the simulated $p_T$ distribution matches the measured distribution in data. The systematic uncertainty due to the reweighting procedure is estimated by applying the correction twice or not at all (up/down variations). This uncertainty affects the normalisation and shape of $t\bar{t}$ background only. The impact of this uncertainty on the expected rate of $t\bar{t}$ events is 46.8% and -30.1% for the up/down variations respectively, while the impact on the estimation of the QCD multijet background is -3.2% and 2.1%.

- **Jet energy corrections**: dedicated energy corrections for CA15 jets are not available. Therefore, the energy corrections for AK7 jets have been used [41]. To show that these corrections are valid, the reconstructed jets in simulation are compared to the
corresponding generator level jets where exactly the same clustering and grooming
algorithms are applied. The ratio between reconstructed and generated $p_T$ is found
to be consistent with unity, with minor deviations within 4%. To evaluate the impact
of the uncertainty on the jet energy scale of CA15 filtered jets, the jet four-momentum
has been varied up and down by the jet energy scale uncertainties of AK7 jets, with
an additional 4% systematic uncertainty. The overall impact of this uncertainty is
small, below 0.5%. The knowledge of the subjet energy scale is assumed to be similar
to the energy scale uncertainty of AK5 jets. The impact on the expected number of
selected $t\bar{t}$ and signal events is between 0% and 5%.

- PDF Uncertainties: signal and $t\bar{t}$ events are weighted according to the uncertainties
  parameterised by the eigenvectors of CTEQ6 [32]. The shifts produced by the indi-
  vidual eigenvectors are added in quadrature in each bin of the observables distribu-
tions. The resulting uncertainties on the number of expected signal and $t\bar{t}$ events is
  2.4% to 8%.

- Scale and Matching Uncertainties: the effect due to the $Q^2$ scale uncertainties on the
  $t\bar{t}$ simulation, has been studied using $t\bar{t}$ MC samples generated with two different
  $Q^2$ scales (scale up and scale down). The effect due to the uncertainty on the extra
  hard parton radiation in $t\bar{t}$ simulation has also been taken into account, using $t\bar{t}$ MC
  samples generated varying the jet matching threshold. It has been verified that these
  uncertainties have no impact on the shapes of $H_T$ and $m_H$ distributions within the
  statistical uncertainties of available simulated samples. The resulting impact of the
  matching uncertainty on the expected number of selected $t\bar{t}$ events is 61%, that of
  the scale uncertainty 47%. These uncertainties have an impact on the estimation of
  the QCD multijet background of 4.3% and 3.4% respectively. Uncertainties on the $t\bar{t}$
simulation and the corresponding propagated uncertainties on the QCD prediction
  are treated as fully anti-correlated.

- QCD background normalisation: the statistical uncertainty due to the closure of the
  method for estimating the QCD multijet background (Section 8) is taken into account
  as systematic uncertainty. For the single Higgs-tag event category an uncertainty of
  10% is used, while 20% is used for the multi Higgs-tag event category.

- Trigger reweighting: a scale factor $SF_{\text{trig}}$ is applied to correct for the different be-
  haviour between data and simulation in the turn-on region of the trigger. A system-
  atic uncertainty in the scale factor is obtained by varying $SF_{\text{trig}}$ by $\pm 0.5 \cdot (1 - SF_{\text{trig}})$. This uncertainty does not affect the plateau region of the trigger, where $SF_{\text{trig}} = 1$
  and a perfect agreement between data and simulation is observed. This uncertainty
  is taken into account both as a shape and as a rate uncertainty. It only affects the first
  bin of the $H_T$ distribution which receives an uncertainty of about 18% for simulated
  $t\bar{t}$ and signal events. The other bins of the $H_T$ distribution are not affected so that the
  overall effect is very small.

- Luminosity: an uncertainty on the integrated luminosity of 2.6% is taken into ac-
  count [42].

10 Results

Figures 7 and 8 show the comparison between data and the expected background contributions
for the single and for the multi Higgs-tag event categories after all event selection criteria. The
QCD multijet background in blue has been derived from data as discussed in Section 8. Signal
samples at three different mass points are shown as well. In these plots only signal samples in
which all T quarks decay into a top quark and a Higgs boson are shown. The hatched error bands in the stacked plots show the quadratic sum of all systematic and statistical uncertainties on the background. In the ratio plot, the statistical uncertainties on the background are depicted by the darker grey central band, while the outer light grey band shows the quadratic sum of all systematic and statistical uncertainties.

Figure 7: $H_T$ (left) and Higgs-candidate mass (right) for the single Higgs-tag category. The QCD background (blue) is derived from data. The $t\bar{t}$ background is derived from simulation. The hypothetical signal is shown by the coloured lines for three different mass points: 500, 700 and 1000 GeV. The hatched error bands in the stack plots show the quadratic sum of all systematic and statistical uncertainties on the background. In the ratio plot, the statistical uncertainties on the background are depicted by the darker grey central band, while the outer light grey band shows the systematic uncertainties and statistical uncertainties added in quadrature.

From $H_T$ and $m_H$ a combined discriminator $L$ is calculated as discussed in Section 5. The distribution of this variable for data in comparison to the background prediction and signal hypotheses is shown in Figure 9 for the single Higgs-tag category and in Figure 10 for the multi Higgs-tag category. As the signal model is built into the discriminator, each signal mass hypothesis has its own definition of $L$. The mass points 500, 700 and 1000 GeV/$c^2$ are shown in these Figures only.

No signal-like excess is observed in data. We set upper limits on the production cross section of hypothetical T quarks. The limit setting is done with theTheta framework [43] using Bayesian statistics. The nuisance parameters are assigned to the sources of systematic uncertainties reported in Section 9, which are taken into account as global normalisation uncertainties and as shape uncertainties where applicable. The likelihood is marginalised in the fit and the shape uncertainties are taken into account applying vertical morphing. Figure 11 shows the expected and observed limits for the hypothesis of an exclusive branching ratio $BR(T \rightarrow tH) = 100\%$ after a combination of both the single and multi Higgs-tag event categories. T quarks of masses below 747 GeV/$c^2$ are excluded, while the expected exclusion limit is 701 GeV/$c^2$.

Expected and observed limits for different branching ratio hypotheses are listed in Table 4 and shown in Fig 12. The observed and expected limits on the production cross section for different branching ratios are given in Table 5 and shown in Fig. 13 and Fig. 14.
11 Conclusion

A search for heavy resonances decaying to top quarks and Higgs bosons has been performed. The benchmark model is a heavy vector-like T quark which is allowed to decay into $bW$, $tZ$ and $tH$. For the first time, a search is presented for vector-like quarks in the all-hadronic final state. The analysis makes use of jet substructure techniques such as top tagging, subjet b-tagging and Higgs tagging. If the heavy T quark has an exclusive branching ratio $BR(T \rightarrow tH) = 100\%$, the observed (expected) exclusion limit on the mass of the T quark is 747 GeV/$c^2$ (701 GeV/$c^2$) after analysing 19.7 fb$^{-1}$ of integrated luminosity of proton-proton collision data collected at $\sqrt{s} = 8$ TeV. In addition, limits have been derived for non-exclusive branching ratios.

Figure 8: $H_T$ (left) and Higgs-candidate mass (right) for the multi Higgs-tag category. The QCD background (blue) is derived from data. The $t\bar{t}$ background is derived from simulation. The hypothetical signal is shown by the coloured lines for three different mass points: 500, 700 and 1000 GeV. The hatched error bands in the stack plots show the quadratic sum of all systematic and statistical uncertainties on the background. In the ratio plot, the statistical uncertainties on the background are depicted by the darker grey central band, while the outer light grey band shows the systematic uncertainties and statistical uncertainties added in quadrature.
Table 5: Branching fractions (first three columns) and the (expected) observed upper limits on the cross section for different masses of the T.
Figure 10: Discriminating variable $L$ constructed from both $H_T$ and $m_H$ for the multi Higgs-tag category. The three different signal hypotheses, 500, 700 and 1000 GeV are shown on the left, middle and right, respectively. The QCD background (blue) is derived from data. The $t\bar{t}$ background is derived from simulation. The hypothetical signal is shown by the coloured line. The hatched error bands in the stack plots show the quadratic sum of all systematic and statistical uncertainties on the background. In the ratio plot, the statistical uncertainties on the background are depicted by the darker grey central band, while the outer light grey band shows the systematic uncertainties and statistical uncertainties added in quadrature.

Figure 11: Expected and observed limits determined from the combined variable $L$ for the combination of the single and multi Higgs-tag categories.
Figure 12: Branching fraction triangle with expected limits (left) and observed limits (right) for the T quark mass. Every point in the triangle corresponds to a particular set of branching fraction values subject to the constraint that all three add up to one. The branching fraction for each mode decreases from one at the corner labelled with the decay mode to zero at the opposite side of the triangle.

Figure 13: Branching fraction triangle with expected cross-section limits for three different T quark mass hypotheses: 500 GeV (left), 700 GeV (centre), 1000 GeV (right). Every point in the triangle corresponds to a particular set of branching fraction values subject to the constraint that all three add up to one. The branching fraction for each mode decreases from one at the corner labelled with the decay mode to zero at the opposite side of the triangle.
Figure 14: Branching fraction triangle with observed cross-section limits for three different $T$ quark mass hypotheses: 500 GeV (left), 700 GeV (centre), 1000 GeV (right). Every point in the triangle corresponds to a particular set of branching fraction values subject to the constraint that all three add up to one. The branching fraction for each mode decreases from one at the corner labelled with the decay mode to zero at the opposite side of the triangle.
References


