J. Phys. G: Nucl. Part. Phys. **34** (2007) N221–N250 doi:10.1088/0954-3899/34/5/N03

RESEARCH NOTE FROM COLLABORATION

Observability of Higgs produced with top quarks and decaying to bottom quarks

D Benedetti[1], S Cucciarelli[2], C Hill[3,6], J Incandela[3], S A Koay[3], C Riccardi[4], A Santocchia[1], A Schmidt[5], P Torre[4] and C Weiser[5,7]

[1] Dipartimento di Fisica, Università degli Studi di Perugia and INFN, Perugia, Italy
[2] CERN, Geneva, Switzerland
[3] University of California, Santa Barbara, USA
[4] Dipartimento di Fisica Nucleare e Teorica, Università degli Studi di Pavia and INFN, Pavia, Italy
[5] Institut für Experimentelle Kernphysik, Universität Karlsruhe (TH), Karlsruhe, Germany

Received 2 November 2006
Published 11 April 2007
Online at stacks.iop.org/JPhysG/34/N221

Abstract

The decay, $H \to b\bar{b}$, is dominant for a Standard Model Higgs boson in the mass range just above the exclusion limit of 114.4 GeV/c^2 reported by the LEP experiments. Unfortunately, an overwhelming abundance of $b\bar{b}$ events arising from more mundane sources, together with the lack of precision inherent in the reconstruction of the Higgs mass, renders this decay mode *a priori* undetectable in the case of direct Higgs production at the LHC. It is therefore of no small interest to investigate whether $H \to b\bar{b}$ can be observed in those cases where the Higgs is produced in association with other massive particles. In this note, the results of a study of Higgs bosons produced in association with top quarks and decaying via $H \to b\bar{b}$ are presented. The study was performed as realistically as possible by employing a full and detailed Monte Carlo simulation of the CMS detector followed by the application of trigger and reconstruction algorithms that were developed for use with real data. Important systematic effects resulting from such sources as the uncertainties in the jet energy scale and the estimated rates for correctly tagging b jets or mistagging non-b jets have been taken into account. The impact of large theoretical uncertainties in the cross sections for $t\bar{t}$ plus N jets processes due to an absence of next-to-leading order calculations is also considered.

1. Introduction

The dominant decay mode of the Standard Model Higgs boson is $H \to b\bar{b}$ in the mass range above the LEP exclusion limit of $m_H \sim 114.4$ GeV/c^2 up to $m_H \sim 135$ GeV/c^2. Direct

[6] Now at University of Bristol, UK.
[7] Now at Physikalisches Institut, Universität Freiburg, Germany.

0954-3899/07/050221+30$30.00

Higgs production is impossible to detect via this decay as a result of the combination of an overwhelming QCD cross section for continuum $b\bar{b}$ production and the inherent imprecision of the Higgs mass reconstruction. While the latter is still true in the case of Higgs production in association with other massive states, such as a $t\bar{t}$ or $b\bar{b}$ pair, such channels do entail substantially lower backgrounds.

The top quark decays almost exclusively to Wb in the Standard Model. As a result, $t\bar{t}H$ events in which the Higgs decays to bottom quarks contain four b quarks in the final state. The events can be further characterized by three salient topologies that are determined by whether or not the two W bosons decay hadronically or leptonically. Thus, in addition to the b jets, $\sim$46% of these events contain four hadronic jets (the 'all-hadron' channel), while $\sim$29% have two hadronic jets as well as a typically isolated electron or muon and missing E_t (the semi-leptonic channel) and $\sim$5% contain two oppositely charged leptons (each of which can be an electron or muon) and missing E_t (the di-lepton channel). The remaining $\sim$20% of events correspond to those cases in which one or both of the W bosons decay to a tau lepton and neutrino and are not easily distinguished as such, as a result of the rich decay repertoire of the tau meson. In fact, these events do contribute in small part to the three other classes of events in typical analyses.

Additional hadronic jets appear in these events and originate from initial and final state QCD radiation (IFSR). The variety and complexity of these events, and particularly the appearance of such a large number of high-energy jets and/or leptons, requires one to rely upon, and fully exploit, the performance of all components of the CMS detector. The tracking system is of particular importance, since the detection of b-jets is crucial to the identification and reconstruction of these events.

In this note, we present results of studies performed on Monte Carlo data samples for which the CMS detector has been fully simulated. Every attempt has been made to process the data as would be the case for real data acquired in proton–proton collisions at a centre of mass energy of 14 TeV. Thus, online triggers have been simulated, and all of the standard offline reconstruction algorithms as they currently exist have been employed in the processing of the data. This emphasis on realism has several important consequences and by-products. Firstly, the analyses are substantially more complicated than studies involving more simple approaches involving parametric detector modelling. This has the advantage of providing a more true picture of the complexity of the events under study, highlighting some of the limitations that might otherwise be overlooked. On the other hand, because these studies are undertaken at a time when no real data are available, it is also not possible to take advantage of the most important tools in the arsenal available to high-energy hadron collider experimentalists: the multitude of control samples resulting from the broadband nature of hadron collider physics.

The note has the following overall structure. Section 2 provides a brief overview of the CMS apparatus. In section 3, the event generation and simulation of signal and background samples are described. Event triggering is discussed in section 4. Section 5 details how the various signature objects (leptons, jets, b jets and missing E_T) of the final state are reconstructed, while a detailed description of the event selection for each of the main event topologies and the attempt to reconstruct the invariant Higgs mass for some of the channels is given in section 6. A brief calculation of expected selection efficiencies is presented in section 7. Effects of systematic uncertainties are discussed in section 8. Section 9 provides a summary of conclusions and future prospects. All results are for an integrated luminosity of 60 fb^{-1}.

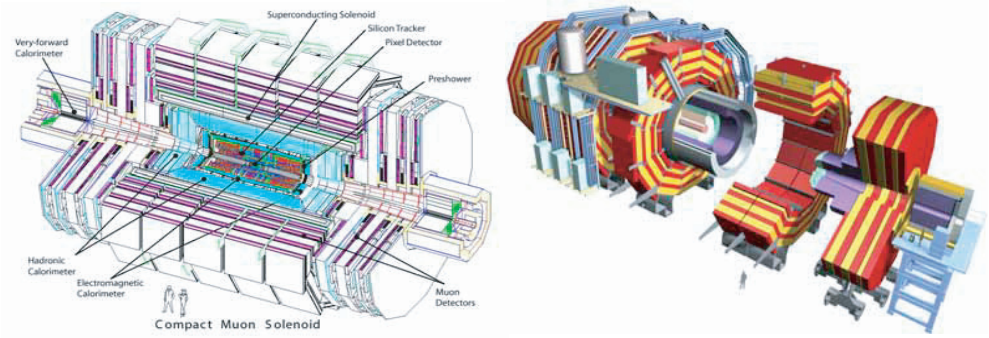

Figure 1. Schematic drawing of the CMS detector with a portion removed to reveal the various subsystems as indicated. The detector is divided into five barrel Yoke sections and three endcap yokes per end. The HF is mounted on the stand at far right.

(This figure is in colour only in the electronic version)

2. The CMS detector [1]

A schematic drawing of CMS is shown in figure 1. The total weight of the apparatus is 12 500 tons. The detector is cylindrical in shape with length and diameter of 21.6 m and 14.6 m, respectively. The overall size is set by the muon tracking system which in turn makes use of the return flux of a 13 m long, 5.9 m diameter, 4 T superconducting solenoid. This high field facilitates the construction of a compact interior tracking system and good exterior muon tracking. The return field saturates 1.5 m of iron into which are interleaved four muon tracking stations. In the central region (pseudorapidity range $|\eta| < 1.2$), the neutron-induced background, the muon rate and the residual magnetic fields are all relatively small, while in the forward regions ($1.2 < |\eta| < 2.4$) all three quantities are relatively high. As a result, drift tube (DT) chambers and cathode strip chambers (CSC) are used for muon tracking in the central and forward regions, respectively. Resistive plate chambers (RPC) with fast response and good time resolution but coarser position resolution are used in both regions for timing and redundancy. The solenoid is large enough to house the inner tracker and the calorimetry.

The sensing elements of the EM calorimeter (ECAL) are PbWO4 (lead tungstenate) crystals. The crystals can tolerate a radiation dose of 10 Mrad and have short radiation ($X_o = 0.89$ cm) and Moliere (2.2 cm) lengths which are ideal for the design of a compact calorimeter with fine granularity. They are also fast, emitting 80% of all scintillation light within the 25 ns spacing between proton bunch crossings at the LHC. The crystals and the avalanche photodiodes used to detect the signal require a temperature stability of 0.1 °C to take full advantage of the excellent inherent energy resolution of the crystals. A preshower system is installed in front of the endcap ECAL to help detect $\pi^o \to \gamma\gamma$.

The CMS central hadronic calorimeter (HCAL) is also inside the magnet coil. The absorber material is brass. A fraction of high-energy hadronic showers extend beyond the HCAL and are retrieved by a layer of scintillators that line the outside of the coil. The active elements consist of plastic scintillator tiles with embedded wavelength-shifting fibres. The HCAL has almost no uninstrumented cracks or dead areas in pseudorapidity. The endcap hadron calorimeter uses the same technology and covers the pseudorapidity region $1.3 < |\eta| < 3$ while the region $3 < |\eta| < 5$ is covered by the iron and quartz-fibre hadron forward calorimeter. Cerenkov light emitted in the quartz fibres is detected by fast photomultipliers. The HF technology is ideal for the dense jet environment typical of this

region as it leads to narrower and shorter hadronic showers. Calorimeter coverage to $|\eta| < 5$ is useful for reducing the uncertainty on missing transverse energy.

The CMS tracker occupies a cylindrical volume of length 5.8 m and diameter 2.6 m. The outer portion of the tracker is comprised of ten layers of silicon microstrip detectors and the inner portion is made up of three layers of silicon pixels. Silicon provides fine granularity and precision in all regions for efficient and pure track reconstruction even in the very dense track environment of high-energy jets. The three layers of silicon pixel detectors at radii of 4, 7 and 11 cm provide 3D space points that are used to seed the formation of tracks by the pattern recognition. The 3D points also enable measurement of the impact parameters of charged-particle tracks with a precision of order 20 m in both the r- and r–z views. The latter allows for precise reconstruction of displaced vertices to yield efficient b-tagging and good separation between heavy and light quark jets.

In regard to performance, the CMS experiment is designed for

- good muon and other particle tracking with good momentum resolution over a wide range of momenta in $|\eta| < 2.5$;
- relatively high efficiency heavy flavour and τ jet tagging with low rates for tagging light quark jets;
- very good e and μ energy resolution in the region $|\eta| < 2.5$ and good separation of γ's and e's from π's;
- the ability to determine the direction of photons and/or identify the relevant primary interaction vertex;
- good missing transverse energy and dijet mass resolution with fine lateral segmentation ($\Delta\eta \times \Delta\phi < 0.1 \times 0.1$) in HCAL.

3. Event generation and simulation

Because the identification of signal relies upon the presence of top quark decay products, one expects that the most significant backgrounds should be those associated with $\mathrm{t\bar{t}}$ events themselves. Indeed, the main backgrounds turn out to be $\mathrm{t\bar{t}jj}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}Z}$ with $\mathrm{Z} \to \mathrm{b\bar{b}}$.

These processes are studied in detail and are presented in this note. Secondary background sources include non-$\mathrm{t\bar{t}}$ QCD multijet events in the case of the all-hadron channel, and W/Z+jets or diboson+jets events in the case of the semi-leptonic and di-lepton channels. With the exception of QCD multijets, these processes have substantially lower production cross sections than $\mathrm{t\bar{t}}$ but similar topologies. Indeed, calculations (section 7 and [3, 4]) show that these backgrounds are negligible and so they are not considered further.

For the generation of the $\mathrm{t\bar{t}H}$ signal and the irreducible $\mathrm{t\bar{t}b\bar{b}}$ background, CompHEP [6] was used in combination with PYTHIA [7]. Though a leading order Monte Carlo, PYTHIA is known to do a very good job of reproducing IFSR and parton shower effects. For the $\mathrm{t\bar{t}}$ plus jets backgrounds, greater care must be exercised. PYTHIA alone cannot be expected to do a realistic job since the relevant processes are not leading order. On the other hand, there is currently no full next-to-leading order (NLO) MC for $\mathrm{t\bar{t}}$ plus jets. Estimation of the k-factors for $\mathrm{t\bar{t}}$ plus jets is also not possible at present. As a result, one uses higher order matrix elements that include additional radiated partons in conjunction with the parton showering of PYTHIA to produce the appropriate event topologies. This is not trivial because the soft QCD effects represented by the PYTHIA parton shower program are not completely distinct from the higher order perturbative diagrams. While there are jet energies for which the two are clearly distinct, they nevertheless represent two extremes in a continuum and so one is forced

Table 1. NLO signal cross sections and H → $b\bar{b}$ branching ratios for different Higgs mass hypotheses.

m_H	115 GeV/c^2	120 GeV/c^2	130 GeV/c^2
σ_{NLO} (pb)	0.747	0.664	0.532
BR(H → $b\bar{b}$)	0.731	0.677	0.525

Table 2. Leading order CompHEP cross sections and effective cross sections after generator-level filters [2] for the background processes under consideration.

	QCD $\hat{p}_t$ = 120–170 GeV/c	QCD $\hat{p}_t$ > 170 GeV/c	$t\bar{t}b\bar{b}$	$t\bar{t}Z$
σ_{LO} (pb)	3.82×10^5	1.05×10^5	3.28	0.65
$\sigma_{\mathrm{LO}} \times \epsilon$ (pb)	76.4	336.0	2.82	0.565

Table 3. Leading order ALPGEN cross sections for background samples of $t\bar{t}$ with various numbers of additional jets.

	Exclusive $t\bar{t}$+1j	Exclusive $t\bar{t}$+2j	Exclusive $t\bar{t}$+3j	Inclusive $t\bar{t}$+4j
σ_{LO} (pb)	170	100	40	61

to artificially place a boundary between them. This is facilitated by a process of matching final jets to initial partons.

For the present study, ALPGEN and PYTHIA are used for their matrix elements and parton showering, respectively, in order to produce more realistic $t\bar{t}$ plus n jets backgrounds. Matching is done in ALPGEN as discussed in [9]. In particular, all of the matrix elements for $t\bar{t}$ plus n additional hard partons are included and properly combined at each order, taking into account interference between amplitudes. These are interfaced to PYTHIA which then proceeds to generate parton showers and IFSR. The final generated event is then checked to see if the number of hard jets in the final state is in fact n for the case of production of exclusive samples of events with n additional partons. Events with more than n hard partons can occur as a result of the high-energy extremes of the parton shower program in PYTHIA. A comparison between CompHEP and ALPGEN for $t\bar{t}$ plus jets samples can be found in [2, 4]. QCD background events were generated with PYTHIA in the $\hat{p}_t$ ranges from 120 to 170 GeV/c and greater than 170 GeV/c. A noteworthy caveat here is the fact that in the absence of data, one cannot evaluate the accuracy of PYTHIA QCD multijet production at LHC energy scales. The QCD event samples used in the present study are therefore unlikely to match real data in all details. When real data become available, it will be possible to tune MC QCD generation to obtain a better correspondence between MC and data.

The signal cross sections calculated at NLO for different Higgs mass hypotheses are listed in table 1 together with the corresponding branching ratios for H → $b\bar{b}$ [8]. Leading order cross sections of background processes together with the effective cross sections after generator-level filters [2] are listed in tables 2 and 3.

The interaction of final state particles with the CMS apparatus was obtained using the CMS detector simulation program which is based upon a full GEANT simulation of the CMS detector. The detector response was digitized after the inclusion of pile-up events (proton–proton collisions per bunch crossing that occur in addition to the hard scattering process of interest).

4. Online trigger selections

It is assumed in what follows that events will be recorded by the CMS data acquisition after they have been accepted by the level-1 (L1) and high-level triggers (HLT) which are described in [10]. For the analyses described here, the cleaner signature of at least one isolated lepton in the final state is exploited whenever possible. The semi-leptonic channels thus require a single muon or single electron trigger, with transverse momentum (p_t) threshold of 19 GeV/c or 26 GeV/c, respectively. A logical OR of the single muon, single electron and single tau triggers is used by the di-lepton channel. Here, the same trigger setups as for the single-lepton streams were used, except that the p_T thresholds were lowered to a common threshold of 15 GeV/c to permit unbiased selection of 20 GeV/c leptons later in the offline analysis.

When there is no lepton, jet triggers are used to select all-hadron events. In particular, single-jet, three-jet and four-jet triggers are combined, using low luminosity E_t thresholds [10] of 572, 195 and 80 GeV, respectively.

The resulting efficiencies for the $t\bar{t}H$ signal samples are 63% for the single muon stream, 52% for the single electron stream, 76% for the di-lepton channel and 25% for the all-hadron channel. Note that as a result of the relatively small branching fraction of the di-lepton channel, it is important not to neglect the contribution to this sample that arises from single-lepton events in which a jet is misidentified as a second lepton. The converse situation, in which a lepton from a di-lepton event is misidentified as a jet, need not be taken into account since it has negligible effect on the single-lepton samples. Thus, with the exception of the di-lepton channel, the efficiencies presented above refer to exclusive samples that contain the appropriate types of W boson decays. The di-lepton efficiency is obtained for a sample in which either one or two leptonic decays are present. Inefficiencies in the lepton triggers are mainly the result of the threshold on the transverse momentum and the finite detector acceptance in pseudorapidity. Indeed, leptons within the detector acceptance are triggered with more than 90% efficiency [11].

The selection efficiencies for $t\bar{t}$ + jets background events are obtained from fully inclusive samples and are between 11% and 14% for the single-lepton channels and around 60% for the di-lepton channels. A detailed breakdown of these trigger efficiencies can be found in [2].

5. Reconstruction

5.1. Lepton reconstruction

The process of muon reconstruction begins in the muon chambers and is then extended to the tracking system, as described in [14]. For the studies presented here, additional selection criteria are applied to help distinguish muons coming from real W decays, (which will be referred to as *signal* muons), from those coming from other sources such as virtual W's appearing in heavy flavour decays or fake muons resulting from hadrons that are mistakenly identified as muons (which will be referred to as *background* muons).

To distinguish between the two classes of muon candidates, probability density functions (PDF) are constructed for a variety of observables associated with muons. The values of these observables differ statistically to varying degrees for the two muon types as measured in Monte Carlo data where a reconstructed true muon is identified to have originated from a W boson decay when the separation, $\Delta R \equiv \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$ in azimuth (ϕ) and pseudorapidity (η) space, satisfies $\Delta R < 0.01$. The observables that are employed are

- transverse momentum,
- track isolation,

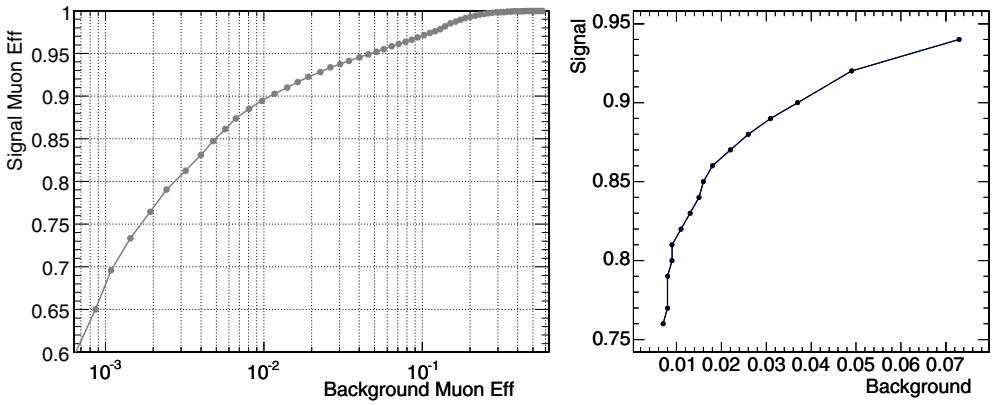

Figure 2. Performance of lepton likelihood discriminators for the semi-leptonic $t\bar{t}H$ channel. The plot on the left is for muons and the one on the right is for electrons.

- calorimeter isolation,
- track impact parameter significance (defined as the impact parameter divided by its uncertainty).

Track isolation measures the sum of transverse momenta of tracks in a cone around the muon candidate. A smaller value corresponds to a larger degree of muon isolation. Calorimeter isolation is completely analogous to track isolation and is based upon the energy collected in the electromagnetic and hadronic calorimeters in a cone around the muon candidate.

Electrons are treated in a very similar manner to that described above for muons except that additional observables are utilized. These include the ratio of the associated calorimeter energy to the track momentum and the ratio between the hadronic and electromagnetic energies associated with the track. Details regarding the calculation of the lepton likelihoods can be found in [2, 5].

The PDFs for both electrons and muons are obtained for a sample of $t\bar{t}H$ events with $m_H = 120$ GeV/c^2 in which one of the W bosons decays to a muon and neutrino and the other decays hadronically. In these samples, a signal lepton efficiency of 90% corresponds to selection of only 1% of background muons (figure 2, left) and ∼3.7% of background electrons (figure 2, right).

5.2. Jet and missing transverse energy reconstruction

Jets are reconstructed using an iterative cone algorithm applied to calorimeter data. A cone size of $\Delta R = 0.5$ is used when at least one W boson decays leptonically, while a smaller cone size was found to be better suited to the more dense jet environments associated with the all-hadron channel. A tower energy threshold of 0.8 GeV and a transverse-energy threshold of 0.5 GeV are used. Calorimeter towers that exceed 1 GeV are potential jet seeds. For the leptonic channels, the jet energy scale is obtained using calibrations obtained with Monte Carlo events [16].

The single-lepton analyses, as described in more detail below, make use of an event likelihood to help to isolate the $H \to b\bar{b}$ decay and subsequently calculate a $b\bar{b}$ invariant mass to associate with the Higgs. This is facilitated, in part, by taking advantage of the various invariant mass constraints associated with top quark decays. The corresponding likelihoods thus rely upon the resolutions that are obtained for the invariant masses of the hadronically decaying W boson and the two top quarks. To evaluate these resolutions, the invariant mass

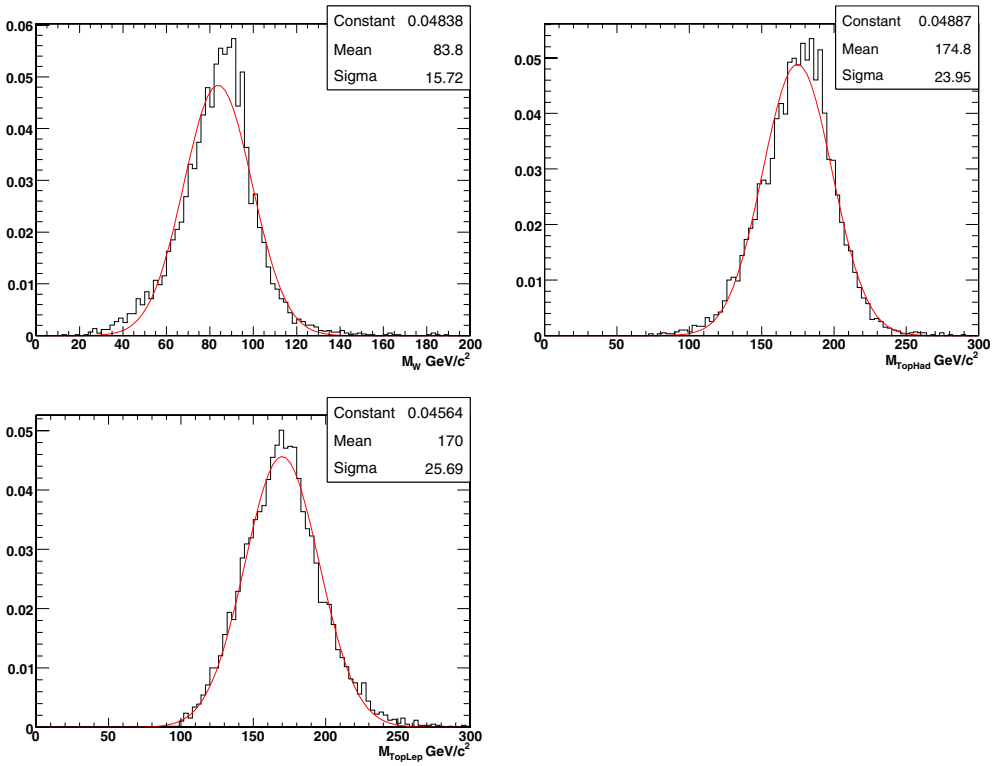

Figure 3. Invariant masses of the hadronically decaying W boson and the hadronically and leptonically decaying top quarks are obtained by means of jet–parton matching with $\Delta R_{j-p} < 0.3$.

distributions for the hadronically decaying W bosons, the hadronically decaying top quarks and the leptonically decaying top quarks are reconstructed using jet matching to generator-level parton information. A reconstructed jet is considered to be matched to the corresponding parton if their separation, ΔR_{j-p}, is less than 0.3. The relevant mass distributions are shown in figure 3.

The missing transverse energy of the event E_t^{miss} is computed as [13]

$$\vec{E}_t^{\text{miss}} = \sum_i \vec{E}_t^{\text{tower}} - \left(\sum_j \vec{E}_t^{\text{RawJet}} - \sum_k \vec{E}_t^{\text{CaliJet}} \right) + \sum_m \vec{E}_t^{\text{Muon}} \tag{1}$$

where the index i runs over the calorimeter towers, j runs over raw jets, k runs over calibrated jets (the number of calibrated jets is the same as the number of raw jets) and m runs over the reconstructed muons of the event. Equation (1) thus takes into account the corrections due to jet calibrations and the contributions of muons which deposit minimal energy in the calorimeter system. The distribution of the reconstructed missing transverse energy of semi-leptonic $t\bar{t}H$ events and the missing transverse-energy resolution are shown in figure 4.

In the case of the all-hadron channel, the choice of the jet reconstruction algorithm is an important step in the event selection optimization, because eight jets are present in the final state. For this reason, an optimization is obtained by means of a simple prototype analysis.

In this prototype analysis, the iterative cone algorithm is employed, and cone sizes ranging from 0.35 to 0.50 are considered. A simple analysis of events is then performed to extract

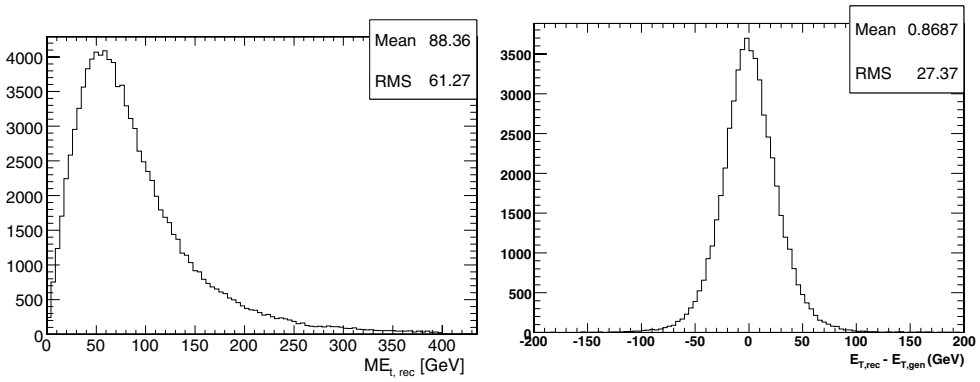

Figure 4. Left: distribution of the reconstructed missing transverse energy. Right: resolution of the reconstructed missing transverse energy in $t\bar{t}H$ events with semi-leptonic W decays.

a significance and signal-to-background ratio for each of a variety of cone sizes in order to obtain a relative comparison of their Higgs discovery potential.

The result of this analysis is that in the case of the all-hadron channel a cone size of 0.4 is the best choice. Details about this study are given in [2, 18].

5.3. b-Tagging

The identification of jets from b-quarks is performed using a combined secondary vertex tagging algorithm that uses track and secondary vertex properties to calculate a discriminator value that allows discrimination of b-jets from non b-jets. A detailed description of the algorithm is available in [17].

In the di-lepton and all-hadron $t\bar{t}H$ analyses, a fixed cut value for the b-tagging discriminator is applied and three or four jets are required to pass this cut in the di-lepton or all-hadron channels, respectively. The rates of misidentification of charm and light flavour jets as b-jets as a function of the b-tagging efficiency are shown in figure 5 for a $t\bar{t}$ plus jets sample. The efficiencies in the $t\bar{t}H$ signal sample are similar [2] to those for $t\bar{t}$.

For the semi-leptonic channels, two dedicated improvements to the secondary vertex tagging algorithm have been introduced. The first is an improved secondary vertex finding algorithm: 'Tertiary Vertex Track Finder' [19]. This algorithm exploits the fact that a b-hadron decay chain contains not only secondary vertices but also tertiary vertices from charm decays and so implements an improved treatment of tracks from tertiary vertices. This is accomplished by searching along the line-of-flight of the b-hadron for additional tracks that are compatible with a cascade decay. These tracks allow a more complete reconstruction of the vertex to be obtained. The second improvement is the inclusion of a soft lepton tagging algorithm that exploits the presence of leptons in b-hadron decays [20]. Depending on the jet flavour and the b-tagging working point, the overall relative improvement in non-b-jet rejection efficiency reaches 25% as discussed in greater detail in [3, 4].

In addition, the event selection and background suppression in the semi-leptonic channels is optimized via a likelihood ratio method. The likelihood ratio is constructed using the distributions of b-tagging discriminator values for the four jets with the highest discriminator values. In this way, the b-tagging information of these four jets is combined into a single discriminator L_{bTag} which simplifies the identification of the optimal b-tagging working point by avoiding the need to adjust four b-tagging cuts simultaneously. Information about non-b-

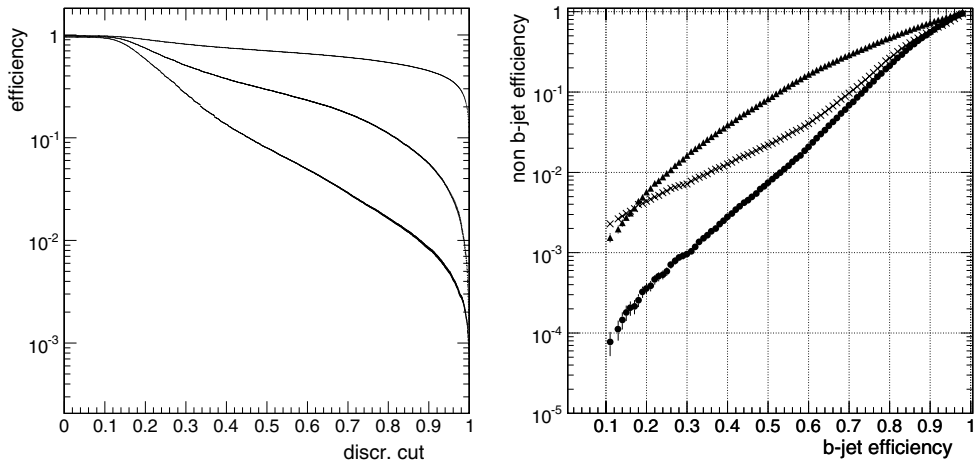

Figure 5. At left are shown the b-tagging efficiencies and misidentification rates versus b-tagging discriminator threshold. The upper curve shows the b-efficiency, while the middle and lower curves show the misidentification rates for charm and light flavour jets, respectively. At right are shown the non-b jet mistagging rate versus b-jet tagging efficiency for c-jets (triangles), uds-jets (stars) and gluon jets (crosses) in a $\mathrm{t\bar{t}}$ plus jets sample. The jets are required to have a minimum transverse momentum of 20 GeV/c. The true jet flavour is the flavour of the primary parton in the jet. Thus, jets originating from gluon splitting to heavy flavour quark pairs are included in the calculation of the gluon misidentification rates.

jets is also taken into account. The outcome is an improved overall performance relative to a fixed b-tagging discriminator cut applied to all jets.

Note that the ordered b-discriminator distributions of b-jets and non-b-jets were obtained for $\mathrm{t\bar{t}H}$ events only. One might expect a further improvement in discrimination by also taking into consideration the ordered b-discriminator distributions in $\mathrm{t\bar{t}}N\mathrm{j}$ background events. These distributions were in fact studied and they were found to be very similar to those presented above and so do not contribute additional discriminating information. More details about this b-tagging method can be found in [3].

6. Event selection and results

In this section, the event selections for the different channels under consideration are described. One can of course arrive at a set of cuts that optimizes the final significance of the signal over background for each channel independently. However, this approach is not ideal because it is important that the data samples are disjoint in order to be able to combine the results of all the $\mathrm{t\bar{t}H}$ search channels. This is of substantial interest, particularly in the time prior to a Higgs discovery. The exclusivity of the data samples used by the various channels can be easily achieved by simply coordinating how the high p_t *signal* electrons and muons from W's are either selected or vetoed by the different analyses. Thus, the di-lepton analysis requires at least two leptons satisfying the discriminator thresholds while the single-lepton channel requires one and only one, and the all-hadron channel requires there be none. Note that the discriminator values in this study were chosen to be those that optimize the signal-to-background significance obtained for the single-lepton channels. This provides a reasonable optimization for the combined significance for all channels, because the single-lepton channel has a substantially larger branching fraction than the di-lepton channel while the all-hadron

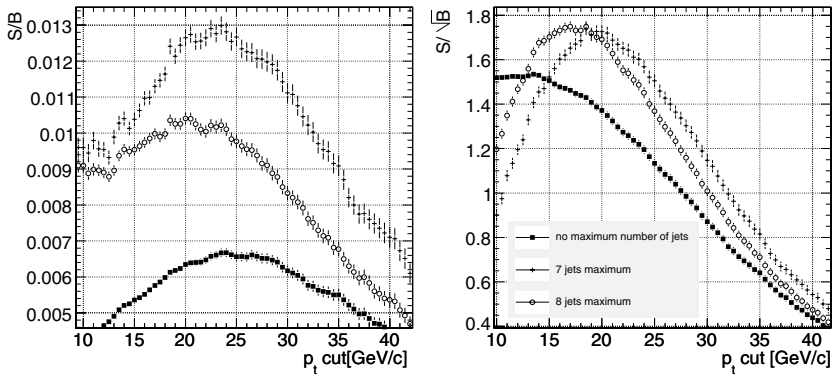

Figure 6. Purity S/B (left plot) and significance $S/\sqrt{B}$ (right plot) versus the cut on jet p_t after an integrated luminosity of 60 fb^{-1}. A simple b-tagging discriminator cut is applied. The squares indicate the requirement of at least six jets, while the crosses and circles further incorporate a cut on the maximum number of jets that pass the p_t cut. All relevant backgrounds ($\mathrm{t\bar{t}}N$j, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}Z}$) are taken into account.

channel is largely unaffected by variation of the discriminator values in the ranges that optimize the other channels. Detailed information regarding the selection criteria can be found in [2].

6.1. Semi-leptonic channel: $\mathrm{t\bar{t}H} \to \mathrm{b\bar{b}b\bar{b}qq'}\mu\nu_\mu$ *and* $\mathrm{b\bar{b}b\bar{b}qq'e}\nu_\mathrm{e}$

The strategy for selecting $\mathrm{t\bar{t}H}$ events with one isolated muon or electron in the final state proceeds along the following three steps:

- preselection,
- choice of jet pairing,
- selection.

The preselection requires the trigger stream for a single muon or a single electron, an isolated lepton using the likelihood information described in section 5.1 and 6 or 7 jets in the pseudorapidity region $|\eta| < 3.0$ with a calibrated transverse energy larger than 20 GeV. The latter is motivated by figure 6, which shows the event selection efficiency in terms of purity S/B and significance $S/\sqrt{B}$ as a function of the cut on the transverse momentum of jets. The figure also displays the dependence on the maximum number of jets required (any number, seven or eight), for the $\mathrm{t\bar{t}H}$ signal and all relevant backgrounds. An increased p_t threshold leads to a decrease in the number of jets passing the preselection requirement of at least six jets. A lowered p_t threshold, on the other hand, leads to more jets passing the p_t cut and therefore to more rejected events due to the requirement of a maximum number of seven or eight jets. The $\mathrm{t\bar{t}H}$ signal sample that has been used for these studies is an inclusive sample containing six jets plus additional jets from radiation. PYTHIA and CompHEP were used for signal generation because neither NLO nor ALPGEN $\mathrm{t\bar{t}H}$ plus N jets samples are available at this time. It is known that PYTHIA's showering algorithm does not model extra jets very well, and in particular does not generate sufficiently many high-energy additional jets in events. It is therefore not adequate for the $\mathrm{t\bar{t}} + N$ jets backgrounds because rare processes beyond $N = 1$ are those that will survive event selection requirements.

For the $\mathrm{t\bar{t}H}$ signal the presence of two b-jets from the Higgs assures that additional jets are not key to event selection and so it is the lowest order processes ($\mathrm{t\bar{t}H}$ with zero or one additional jet) that will dominate the final sample after event selection. Thus, the impact of a requirement of a maximum number of jets on the signal selection is expected to be realistically described for

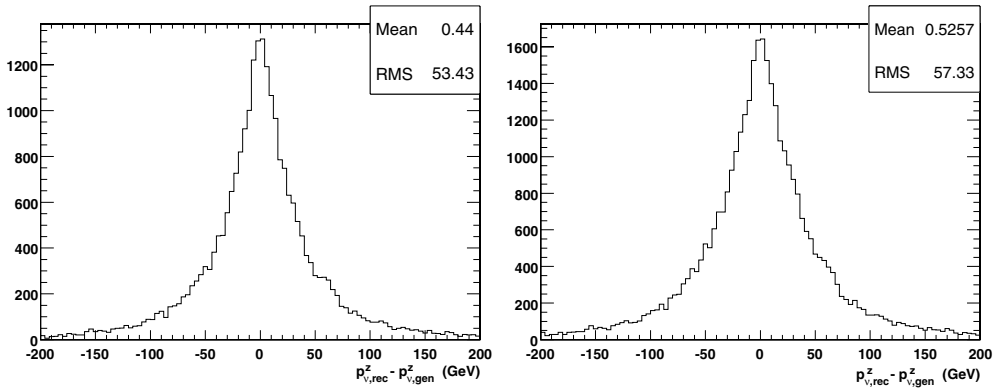

Figure 7. Longitudinal resolution of the neutrino: on the left, only those cases for which there are real solutions obtained for the W boson's 4-momentum constraint are included. On the right, the real plus collinear solutions as described in the text are shown.

the $t\bar{t}H$ signal by PYTHIA for which the parton shower model produces an additional radiated jet with roughly the expected frequency and roughly the expected kinematic properties.

To retain exclusivity with the di-lepton channel, a double muon, double electron and muon–electron veto are applied.

In order to perform a complete reconstruction of the event, the longitudinal component of the neutrino momentum has to be computed from 4-momentum conservation for the W boson: $m_W^2 = (E^\mu + E^\nu)^2 - (\vec{p}^\mu + \vec{p}^\nu)^2$. This equation gives two real solutions for p_z^ν in 66% of the cases, which are both evolved as a possible hypothesis during the subsequent jet pairing as described below. In the remaining 34%, the neutrino is taken to be collinear with the lepton: $p_z^\nu = p_z^l$. The resolution of p_z^ν before and after the collinear approximation is shown in figure 7. A small degradation in the longitudinal resolution is observed, but the reconstruction efficiency of the leptonic W boson decay is increased to 100%.

An important part of the analysis is the jet pairing. This is necessary in order to identify the two b-jets from the Higgs boson with the highest probability. Several strategies for assigning the jets to their originating partons have been developed. Among these are kinematic fits [2, 5, 15] and likelihood methods exploiting kinematic properties [4, 5], mass resonances and angular distributions [3, 4]. All of these methods achieve comparable efficiencies for correct jet assignment near 30%. Thus, one obtains an invariant mass peak for the Higgs boson that is smeared by the jet-pairing inefficiencies and detector resolution effects.

After the jet assignment is complete, additional criteria are applied to further reject background. The variable with the largest impact in this domain is the b-tagging likelihood value described in section 5.3. The expected significance $S/\sqrt{B}$ and purity S/B prior to consideration of systematic uncertainties is shown versus the cut on L_{bTag} in figure 8 for three different Higgs boson masses; 115, 120 and 130 GeV/c^2, and for an integrated luminosity of 60 fb^{-1}. The integrated luminosity of 60 fb^{-1} corresponds to three years of data taking at a luminosity of $2 \times 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The results shown in the plots are obtained without requiring that the reconstructed mass lie in any particular mass window. The significance reaches its maximum for a cut value between 0.125 and 0.225. The reconstructed invariant Higgs boson mass in the case of a generated value of $m_H = 115$ GeV/c^2 is shown on the left-hand side of figure 9 in comparison with the combinatorial background. Here 'combinatorial background' refers to events in which the two b-jets assigned to the Higgs boson are not within $\Delta R < 0.5$ of the generated b-partons from Higgs decay. The right-hand side of this figure shows the

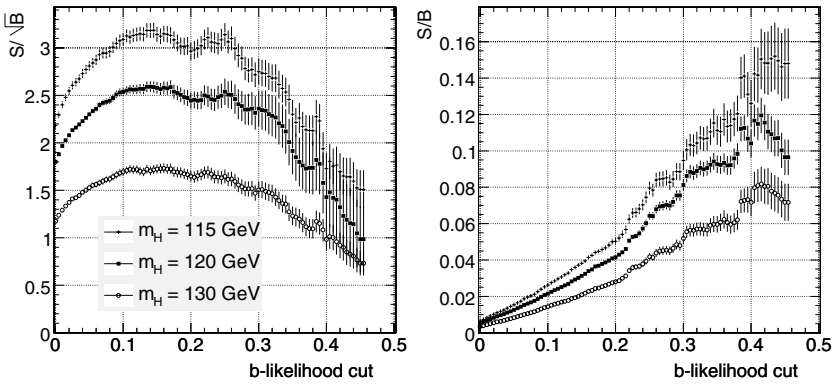

Figure 8. Significance (left) and purity (right), prior to inclusion of systematic uncertainties, for three different Higgs boson masses (115, 120 and 130 GeV/c^2) as a function of the cut on the b-tagging likelihood L_{bTag}, for an integrated luminosity of 60 fb^{-1} for the semi-leptonic (muon and electron) $\mathrm{t\bar{t}H}$ decay channel. No mass window requirement has been applied. The error bars indicate the statistical error due to the finite sizes of datasets. Bin-to-bin correlations also occur here because of the sliding cut on the b-tagging likelihood.

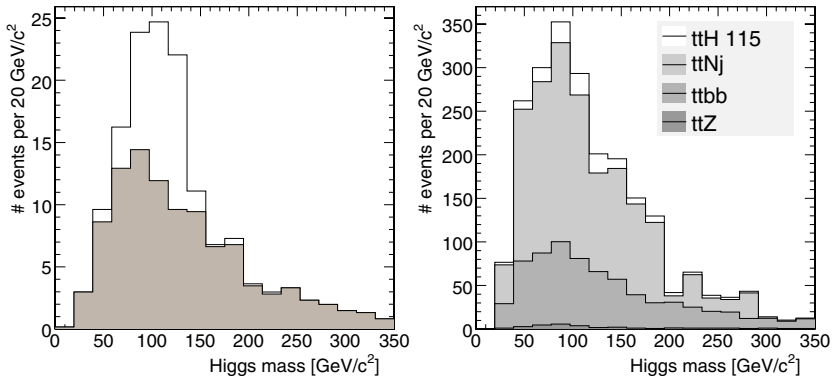

Figure 9. Invariant Higgs boson mass spectrum for an L_{bTag} cut of 0.225 and $m_H = 115$ GeV/c^2, after an integrated luminosity of 60 fb^{-1}. Only signal events are shown at left. The combinatorial background is shaded grey. The plot at the right adds all relevant physical backgrounds ($\mathrm{t\bar{t}Z}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}}N\mathrm{j}$) to the $\mathrm{t\bar{t}H}$ signal (including the combinatorial background). The contributions from all sources are stacked on top of each other.

physical backgrounds ($\mathrm{t\bar{t}Z}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}}N\mathrm{j}$) and the $\mathrm{t\bar{t}H}$ signal stacked on top of each other. Due to the limited amount of available Monte Carlo statistics for the $\mathrm{t\bar{t}}N\mathrm{j}$ background and the large scale factors that have to be applied to the remaining events, the statistical fluctuations in figure 9 are large.

The preselection and selection efficiencies, together with the corresponding numbers of expected events and signal significances, prior to consideration of systematic uncertainties, are reported in table 4 for the channel with a muon or an electron in the final state. The table presents results for two working points: a 'loose' working point, which optimizes $S/\sqrt{B}$, and a 'tight' working point, which optimizes the purity S/B.

Thus far, no mass window requirement has been used. If one applies the requirement $m_H < 150$ GeV/c^2, the purity can be increased by about 10%, while the significance does not change appreciably. This improvement is very minor because the shape of the Higgs

Table 4. Selection efficiency for $L_{\mathrm{bTag}} > 0.225$ ($\epsilon_{\mathrm{loose}}$) and for $L_{\mathrm{bTag}} > 0.350$ ($\epsilon_{\mathrm{tight}}$), number of expected events and signal significance in 60 fb^{-1} for the muon and electron $\mathrm{t\bar{t}H}$ channels. The signal datasets are labelled by the generated Higgs mass in GeV/c^2 (in parentheses). Also quoted are binomial errors arising from the finite sizes of processed datasets. No Higgs mass window has been applied. The last column of $\mathrm{t\bar{t}4j}$ gives the upper limit corresponding to a confidence level of 68% since no events are remaining after the cuts in this case.

	Number of events	$\epsilon_{\mathrm{loose}}$ (%)	$N^{\mathrm{ev}}_{\mathrm{loose}}$	$\epsilon_{\mathrm{tight}}$ (%)	$N^{\mathrm{ev}}_{\mathrm{tight}}$
$\mathrm{t\bar{t}H}$ (115)	55 395	1.60 ± 0.05	147 ± 5	0.5 ± 0.03	48 ± 3
$\mathrm{t\bar{t}H}$ (120)	191 133	1.55 ± 0.03	118 ± 2	0.52 ± 0.016	40 ± 1
$\mathrm{t\bar{t}H}$ (130)	44 595	1.70 ± 0.06	80 ± 3	0.54 ± 0.03	25 ± 2
$\mathrm{t\bar{t}1j}$	1297 064	0.0045 ± 0.0006	464 ± 60	0.00046 ± 0.0002	47 ± 19
$\mathrm{t\bar{t}2j}$	827 615	0.0089 ± 0.00103	536 ± 62	0.0011 ± 0.00036	65 ± 22
$\mathrm{t\bar{t}3j}$	108 778	0.014 ± 0.0035	331 ± 85	0.0028 ± 0.0016	66 ± 38
$\mathrm{t\bar{t}4j}$	114 054	0.0035 ± 0.0017	128 ± 64	0	<36
$\mathrm{t\bar{t}b\bar{b}}$	384 407	0.43 ± 0.01	734 ± 18	0.141 ± 0.006	239 ± 10
$\mathrm{Zt\bar{t}}$	94 706	0.104 ± 0.011	35 ± 4	0.029 ± 0.005	10 ± 2
Total background			2230		427
$S/\sqrt{B}$ (115)			3.1		2.3
S/B (115)			6.6%		11%
$S/\sqrt{B}$ (120)			2.5		1.9
S/B (120)			5.3%		9.3%
$S/\sqrt{B}$ (130)			1.7		1.23
S/B (130)			3.6%		5.9%

mass peak shown on the left-hand side of figure 9 is very similar to the shape of the physical background. It is possible that the contribution of the signal can be better distinguished from background once real data are available and a refined understanding of the background shapes is achieved. Currently, however, the limited separation of signal from background shapes leads one to conclude that the search for $\mathrm{t\bar{t}H}$ with $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$ has to be treated as a simple counting experiment, which then relies heavily upon ones knowledge of event rates and their corresponding systematic errors. The latter are evaluated in section 8.

NLO calculations for $\mathrm{t\bar{t}}$ plus two or more jets are not yet available and may not be available in time for early LHC operation. It will therefore be necessary to estimate all reducible backgrounds for this analysis from data. This will leave the irreducible backgrounds, such as $\mathrm{t\bar{t}b\bar{b}}$, which will be a serious limitation to observation of Higgs in association with $\mathrm{t\bar{t}}$.

6.2. $\mathrm{t\bar{t}H} \rightarrow \mathrm{b\bar{b}b\bar{b}l'\nu'l\nu}$

Di-lepton $\mathrm{t\bar{t}H}$ events are selected by requiring two reconstructed leptons (e, μ) accompanied by significant missing transverse energy and at least four but no more than seven jets, of which at least three have been b-tagged by the *combined secondary vertex* b-tagging algorithm.

Lepton identification is performed using the electron and muon likelihoods described in section 5. Events with more than one identified lepton that are vetoed by the single-lepton analyses are precisely the events that are selected by the di-lepton analysis.

Missing transverse energy is corrected for jet calibration and muon momenta according to equation (1), resulting in the distribution shown in figure 10. At present, the di-lepton analysis is a counting experiment and no effort has been made to assign the missing transverse energy to the two neutrinos from the hard event.

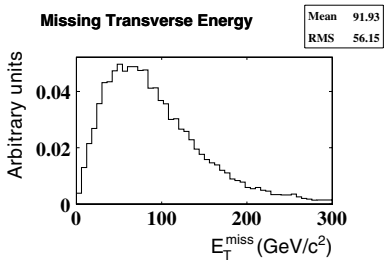

Figure 10. Missing transverse energy for non-exclusive $t\bar{t}H$ ($m_H = 120$ GeV/c^2) events.

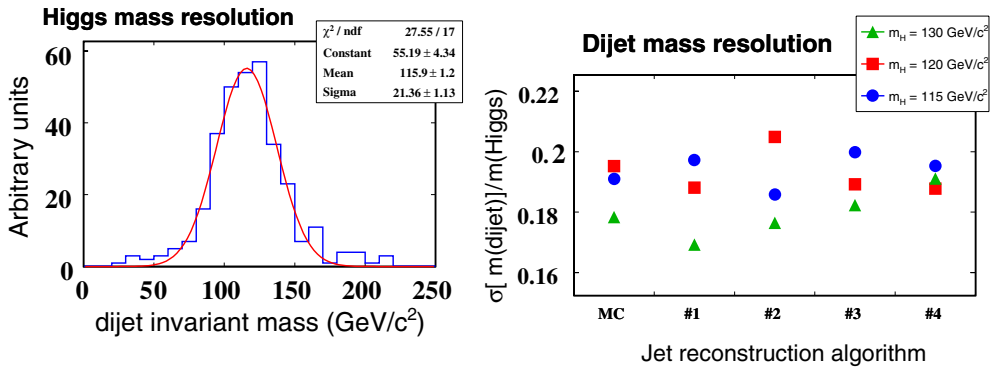

Figure 11. Left: invariant mass of pairs of jets matched ($R < 0.3$) to the Monte Carlo b partons from Higgs decay. A Gaussian fit is performed to obtain the 'best case' Higgs mass resolution σ. This example is for non-exclusive $t\bar{t}H$ with $m_H = 120$ GeV/c^2, with jets calibrated using the PTDR II recommendation 1 settings. Right: the relative 'best case' Higgs mass resolution σ/m(Higgs) for various Higgs masses and jet calibrations (Monte Carlo calibration or PTDR II recommendations 1, 2, 3 and 4). m(Higgs) is the mean reconstructed Higgs mass according to the Gaussian fit.

Jets are reconstructed using the iterative cone algorithm with a cone size of $\Delta R = 0.5$. The parameters and calibration used in the reconstruction follow the first of ten sets recommended in the CMS Physics Technical Design Report Volume II (PTDR II) [12]. The settings are detailed in [23]. A number of the other recommended settings were investigated and found to produce little variation in either the selection efficiencies or the 'best case' Higgs mass resolution (see figure 11 and associated caption).

As for the semi-leptonic electron channel, jets found in a $R < 0.1$ cone centred upon any selected electron are removed. This affects about 1% of all reconstructed 'jets' that are in actuality misidentified electrons.

The di-lepton $t\bar{t}H$ event selection criteria are the following:

- Two oppositely charged leptons (e, μ) passing discriminant thresholds: $-\mathrm{Log}(L_\mu) < 1.4$ for muons, $-\mathrm{Log}(L_e) < 1.2$ for electrons.
- Corrected $E_T^{\mathrm{miss}} > 40$ GeV.
- Four to seven jets with calibrated $E_T > 20$ GeV and $|\eta| < 2.5$.
- Three or four selected jets b-tagged with discriminator $D > 0.7$.

As for the semi-leptonic and all-hadron channels, a tighter set of selection cuts has been applied. It consists of 4–6 jets with calibrated $E_T > 20$ GeV and $|\eta| < 2.5$ and four jets passing the discriminator cut $D > 0.7$.

Table 5. Preselection efficiency ϵ_{pre}, selection efficiency ϵ (including branching fraction where applicable), and respective numbers of expected events N_{pre} and N in 60 fb^{-1}, for the di-lepton $\bar{\text{t}}\text{t}$H channel. Also quoted are binomial errors arising from the finite sizes of processed datasets.

	Analysed events	ε_{pre} (%)	N_{pre}	ε (%)	N
ttH ($m_H = 115$ GeV/c^2)	27 900	n/a	n/a	0.511 ± 0.025	168 ± 8.0
ttH ($m_H = 120$ GeV/c^2)	26 100	n/a	n/a	0.49 ± 0.025	132 ± 6.7
ttH ($m_H = 130$ GeV/c^2)	25 900	n/a	n/a	0.49 ± 0.025	82.2 ± 4.2
tt1j	280 000	5.1 ± 0.042	$520\,000 \pm 4200$	0.0125 ± 0.0021	1270 ± 220
tt2j	277 000	6.22 ± 0.046	$373\,000 \pm 2800$	0.0448 ± 0.004	2690 ± 240
tt3j	90 400	7.32 ± 0.087	$176\,000 \pm 2100$	0.0553 ± 0.0078	1330 ± 190
tt4j	120 000	10.2 ± 0.087	$374\,000 \pm 3200$	0.0716 ± 0.0077	2620 ± 280
ttbb	314 000	30.8 ± 0.082	$52\,100 \pm 140$	0.637 ± 0.014	1080 ± 24
ttZ	110 000	12.4 ± 0.099	$4\,200 \pm 34$	0.304 ± 0.017	103 ± 5.6
All backgrounds					9090

Table 6. Signal significance $S/\sqrt{B}$ of di-lepton $\bar{\text{t}}\text{t}$H channel.

	S/B	$S/\sqrt{B}$
ttH ($m_H = 115$ GeV/c^2)	0.0184	1.76
ttH ($m_H = 120$ GeV/c^2)	0.0145	1.39
ttH ($m_H = 130$ GeV/c^2)	0.009 04	0.862

The efficiency of the above selection has been studied for signal with Higgs mass values $m_H = 115$, 120 and 130 GeV/c^2 and for the dominant background processes. Signal samples were generated with the Higgs forced to decay to $\text{b}\bar{\text{b}}$. In addition, W^- was forced to decay leptonically (e, μ, τ), but the W^+ is allowed to decay freely. Such a 'non-exclusive' dataset incurs a branching ratio of 1/3, which has been factored into the selection efficiencies reported in table 5. This choice of dataset allows one to take into account the important contributions to this channel from semi-leptonic top decays which are mis-reconstructed as di-lepton events, and from leptonic tau decays as well as hadronic tau decays which are mis-reconstructed as e, μ. Obviously, the sample generation avoids all-hadron events which would not survive the di-lepton selection criteria in general.

The contributions from the background processes, $\bar{\text{t}}\text{t}\text{b}\bar{\text{b}}$, $\bar{\text{t}}\text{t}$Nj and $\bar{\text{t}}\text{t}\text{Z}$, are estimated from the samples described in section 3. The selection efficiencies for these processes are very small and so very large samples must be analysed. To make these samples more manageable, a loose preselection requiring at least 3 b-tags with discriminator $D > 0.7$ is applied before analysis.

6.2.1. Results. The preselection and selection efficiencies, with the corresponding numbers of expected events as well as the signal significance, are reported in table 5. The number of expected events is computed for an integrated luminosity of 60 fb^{-1}.

The significance of the di-lepton $\bar{\text{t}}\text{t}$H analysis versus Higgs mass is summarized in figure 12 and the accompanying table 6.

Since event selection is quite simple for the di-lepton channel, it is possible to formulate equations predicting the selection efficiencies. This is detailed in [2], where some back-of-the-envelope calculations are performed to estimate efficiencies for signal and backgrounds. The calculations include some of the most important backgrounds that were not taken into

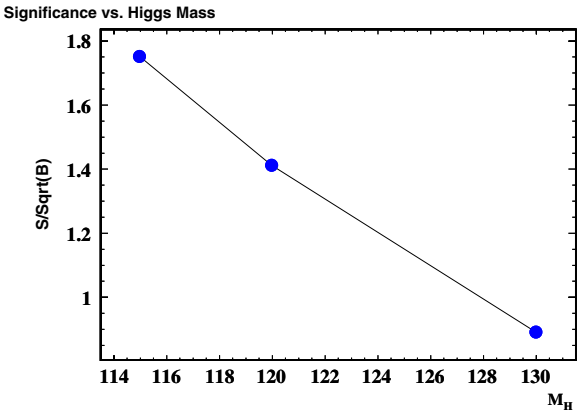

Figure 12. Significance versus Higgs Mass for di-lepton Analysis.

account in this analysis and show that their yields are expected to be negligible compared to those backgrounds that were considered.

6.3. All-hadron channel: $t\bar{t}H \rightarrow b\bar{b}b\bar{b}qq'q''q'''$

The all-hadron channel must address events with eight or more jets. Jet reconstruction is thus of critical importance. In view of this, a dedicated study of jet reconstruction and calibration was performed for this channel and published in [18]. The study resulted in an advanced calibration that corrects the jet energies by using the generated primary partons as reference.

For the task of jet pairing, a χ^2 method using the invariant masses of top quarks and W bosons was employed in the all-hadron channel. The following χ^2 variable is calculated for each possible jet combination:

$$\chi^2_{\text{mass}} = \left(\frac{m_{W^+} - m_{jj}}{\sigma(m_W)}\right)^2 + \left(\frac{m_{W^-} - m_{jj}}{\sigma(m_W)}\right)^2 + \left(\frac{m_t - m_{jjj}}{\sigma(m_t)}\right)^2 + \left(\frac{m_{\bar{t}} - m_{jjj}}{\sigma(m_t)}\right)^2. \quad (2)$$

The expected mass values and their resolutions (σ) are obtained by the same methodology as that employed in the semi-leptonic channel described earlier. The particular jet combination that yields the minimal χ^2 value is chosen for application of additional event selection criteria.

To optimize the signal selection relative to background rejection, the following variety kinematical variables have been studied:

- Transverse energies of the jets.
- Combined b-tagging discriminator.
- Event centrality, defined as $\sum_{i=0}^{8} E_T^i \big/ E^i$.
- Higgs centrality, defined as above, but only for the two jets assigned to the Higgs boson.

The thresholds applied to these variables were varied in steps in order to map performance over a very broad range of configurations. As an illustrative example, figures 13–16 show how significance $S/\sqrt{B}$ and purity S/B change upon varying one cut while keeping the other cuts fixed.

As for the other channels reported above, two sets of criteria were applied corresponding to 'loose' and 'tight' working points. The two sets differ mainly in the choice of the b-tagging discriminator threshold, since this has the largest influence on the suppression of light flavour backgrounds. The results are summarized in table 7.

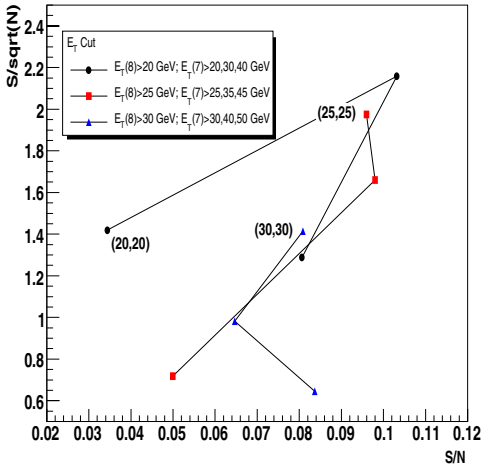

Figure 13. E_T thresholds for the 7th and 8th jets. Markers display the evolution of the purity and significance of the event selection as a function of the 7th jet E_T threshold as indicated in the legend while the 8th jet E_T threshold is kept constant [2].

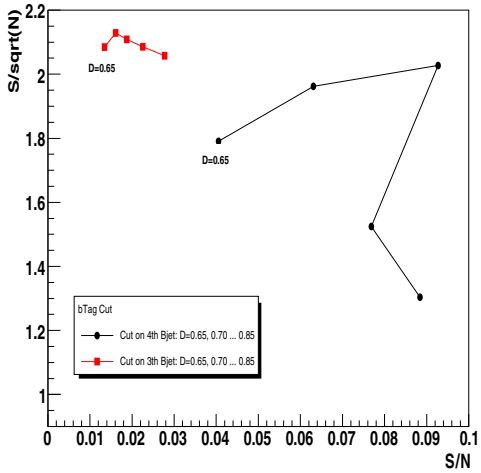

Figure 14. The evolution of the purity and significance of the event selection as a function of the threshold on the 'combined' b-tagging discriminator for the jets with the 3rd and 4th highest discriminator values, respectively [2].

Even though the *loose* working point gives a better result in terms of significance $S/\sqrt{B}$, the *tight* working point can be a better choice, once systematic errors are included.

7. Calculations of the expected selection efficiencies in the semi-leptonic channels

Compared to previous CMS studies [21, 22], in which significances of more than 5 have been reached, the $t\bar{t}$ plus light flavour jets background has proven to be dramatically more dominant (by more than a factor of 3) in the present study. This is the key to understanding the differences between results reported here and previous results regarding the observability of this channel.

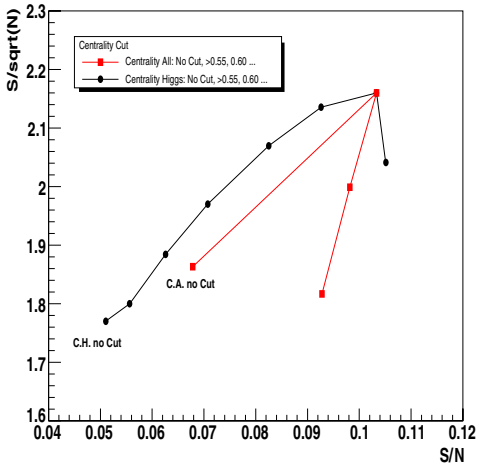

Figure 15. Evolution of significance and purity as a function of the Higgs centrality or total event centrality, respectively [2].

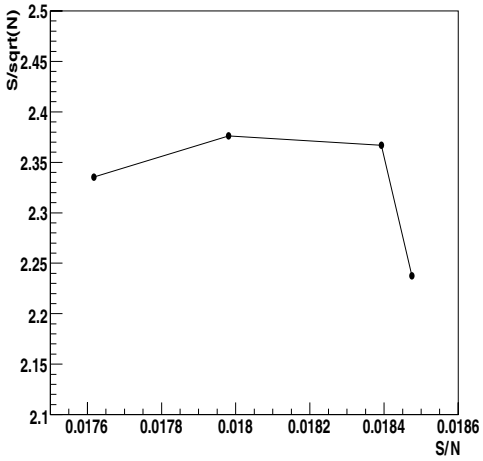

Figure 16. Evolution of significance and purity as a function of the maximum allowed absη of jets, in steps of $\Delta\eta = 0.2$, ranging from absη 2.4 to 3 [2].

As a cross check of the results obtained and reported in this note, a detailed analytic calculation of expected $\mathrm{t\bar{t}}N\mathrm{j}$ background rates to the single-lepton channel was performed. The calculation relies upon the b-tagging performance presented in section 5.3. A similar calculation for the di-lepton channel is reported in [2].

In order to calculate expected event rates corresponding to various b-tagging efficiencies and mistagging rates, the jet composition of the event samples has to be decomposed into specific flavours. This is seen in table 8 which lists the distributions of the jet flavours for the various samples.

In the detailed calculations, the tagging rates and jet flavour compositions are combined to evaluate suppression factors as indicated below.

The probability of b-tagging exactly i out of n_b b-jets is $C_i^{n_b}\varepsilon_b^i(1-\varepsilon_b)^{n_b-i}$, where $C_i^{n_b}$ is the combinatorial factor for the number of ways i jets out of n_b can be assigned independent

Table 7. Analysed events, selection efficiency, number of expected events and signal significance in 60 fb^{-1} for the all-hadron $t\bar{t}H$ channel for two working points ϵ_{loose} and ϵ_{tight}. Signal datasets are labelled by the generated Higgs mass in GeV/c^2 shown in parentheses. Also quoted are binomial errors arising from the finite sizes of processed datasets. The criteria applied for both working points are: $E_T^{8\text{th}} > 20$ GeV, $E_T^{7\text{th}} > 30$ GeV, χ^2 for W and top within 3σ of their expected values, and Higgs centrality > 0.55. For the *loose* working point there is also a requirement that the jet with the 3rd highest b-tagging discriminant value satisfies $D_3 > 0.8$. At the *tight* working point, the additional criteria are: $D_3 > 0.85$, $D_4 > 0.7$ and event centrality > 0.8.

	Number of events	ϵ_{loose} (%)	$N^{\text{ev}}_{\text{loose}}$ 60 fb^{-1}	ϵ_{tight} (%)	$N^{\text{ev}}_{\text{tight}}$ 60 fb^{-1}
$t\bar{t}H$ (115)	49 636	2.32 ± 0.07	347 ± 10	0.294 ± 0.015	44 ± 4
$t\bar{t}H$ (120)	163 494	2.55 ± 0.03	314 ± 5	0.366 ± 0.024	45 ± 2
$t\bar{t}H$ (130)	43 254	2.80 ± 0.08	214 ± 6	0.358 ± 0.029	27 ± 2
$t\bar{t}b\bar{b}$	203 135	0.702 ± 0.019	1190 ± 31	0.0645 ± 0.0056	109 ± 9
$t\bar{t}1j$	1031 551	0.0084 ± 0.0009	860 ± 92	0.0005 ± 0.0002	49 ± 22
$t\bar{t}2j$	559 111	0.0333 ± 0.0024	2000 ± 150	0.0009 ± 0.0004	54 ± 24
$t\bar{t}3j$	68 015	0.079 ± 0.011	1910 ± 260	0.0015 ± 0.0015	35 ± 35
$t\bar{t}4j$	97 334	0.182 ± 0.014	6660 ± 500	0.0021 ± 0.0015	75 ± 53
$Zt\bar{t}$	80 226	0.358 ± 0.021	121 ± 7	0.0312 ± 0.0062	11 ± 2
qcd170	264 310	0.0238 ± 0.0030	4810 ± 610	0.0004 ± 0.0004	76 ± 76
qcd120	55 128	0.0018 ± 0.0018	83 ± 83	0 ± 0	<95 (68%CL)
Total background			17 600		<505
$S/\sqrt{B}$ (115)			2.6		2.0
S/B (115)			2.0%		8.7%
$S/\sqrt{B}$ (120)			2.4		2.0
S/B (120)			1.8%		8.9%
$S/\sqrt{B}$ (130)			1.6		1.2
S/B (130)			1.2%		5.4%

Table 8. Percentage flavour composition of jets in the various data samples.

Sample	Light flavour	Charm	Bottom
Semi-leptonic $t\bar{t}H$	39.4	6.9	53.6
Semi-leptonic $t\bar{t}1j$	60.6	8.4	30.9
di-leptonic $t\bar{t}1j$	62.2	1.5	36.3
All-hadronic $t\bar{t}1j$	65.6	14	20.4
Semi-leptonic $t\bar{t}2j$	62.2	8.6	29.2
Di-leptonic $t\bar{t}2j$	63.9	2.1	33.9
All-hadronic $t\bar{t}2j$	67.7	12.5	19.8
Semi-leptonic $t\bar{t}3j$	63.9	8.7	27.4
Di-leptonic $t\bar{t}3j$	65.2	3.4	31.3
All-hadronic $t\bar{t}3j$	72.2	11.1	16.7
Semi-leptonic $t\bar{t}4j$	67.2	8.12	24.7
Di-leptonic $t\bar{t}4j$	68.6	4.5	26.8
All-hadronic $t\bar{t}4j$	75	6.8	18.2

of order:

$$C_i^N \equiv \frac{N!}{i!(N-i)!}. \tag{3}$$

Table 9. Selection efficiencies calculated according to equations (4) and (5) together with the efficiencies obtained from the single-lepton analysis as presented in table 4.

Sample	Calculated efficiency (%)	Monte Carlo selection efficiency (%)
$t\bar{t}H$	1.47	1.6
$t\bar{t}1j$	0.0078	0.005
$t\bar{t}2j$	0.012	0.01
$t\bar{t}3j$	0.015	0.014
$t\bar{t}4j$	0.0046	0.004

After including the mistagging probabilities, equation (4) gives the probability for b-tagging exactly n jets out of n_b b-jets, n_c charm jets and n_l light flavour jets:

$$\varepsilon^n(n_b, n_c, n_l) = \sum_{i=0}^{n} \sum_{j=0}^{n-i} \left[C_i^{n_b} \varepsilon_b^i (1-\varepsilon_b)^{n_b-i}\right]\left[C_j^{n_c} \varepsilon_c^j (1-\varepsilon_c)^{n_c-j}\right] \times \left[C_{n-i-j}^{n_l} \varepsilon_l^{n-i-j} (1-\varepsilon_l)^{n_l-(n-i-j)}\right]. \quad (4)$$

The fraction of jets of each type varies across samples so that the suppression factors must be calculated separately for each background sample.

The suppression factor for $t\bar{t}2j$ is then

$$\epsilon_{t\bar{t}2j} = \epsilon_{t\bar{t}2j}^{\text{pre}} (0.6 \cdot \varepsilon^{n=4}(2,1,3) + 0.4 \cdot \varepsilon^{n=4}(2,1,4)) = 0.012\%, \quad (5)$$

where $\epsilon_{t\bar{t}2j}^{\text{pre}}$ represents the preselection efficiency. The factor $\varepsilon^{n=4}(2,1,3)$ corresponds to equation (5) with $n_b = 2$, $n_c = 1$ and $n_l = 3$ which gives six jets in total. In the case of seven jets, $n_l = 4$ is used. These numbers are motivated by the values in table 8. The factors 0.6 and 0.4 in equation (5) correspond to the fraction of events with six jets (60%) or seven jets (40%). A more detailed presentation of these calculations is available in [3]. This result is in good agreement with the observed value at the *loose* working point in table 4. Similarly, the calculated efficiencies for signal and the other backgrounds summarized in table 9 are in very good agreement with those obtained for the analyses of Monte Carlo data presented above.

The excellent agreement that is found between the calculated and the measured Monte Carlo selection efficiencies provides an important cross check of the analysis.

Note that the $t\bar{t}Nj$ background data samples used in the study presented here were generated with ALPGEN which predicts smaller cross sections than CompHEP. The resulting rates, for identical selection efficiencies, differ by roughly a factor of 2 between ALPGEN and CompHEP samples. This also helps to explain some of the discrepancies that are found between the present study and some previous studies that relied on CompHEP.

The most important difference with previous published studies in CMS, however, stems from the fact that a fast parametric simulation of the CMS tracker was used in earlier studies. As shown in [4], the b-tagging performance obtained for the current version of the CMS fast simulation is not in good agreement with full simulation. The misidentification rate of light flavour jets in particular shows a difference of up to a factor of 5. This is significant since the $t\bar{t}$ plus light flavour background is the dominant contribution to the background in the present study. Moreover, as confirmed by other experiments such as the current Tevatron experiments, the light flavour misidentification rates are not easy to describe correctly with detector simulation programs.

Another difference with results from previous studies arises from the distribution of the jet flavours shown in table 8, where a substantial charm jet contamination is seen. These

jets arise from gluon splitting as well as from W boson decays. Therefore, a $t\bar{t}2j$ background cannot be simply understood as consisting of b- and light flavour jets only. Rate calculations based on parameterized tagging must take these jets into account properly, as they were in the calculations presented in this section.

8. Systematic errors

In this section, the systematic uncertainties relevant to the present understanding of the expected performance of the CMS detector will be evaluated. The following sources of uncertainties are taken into account:

- Jet energy scale (JES).
- Jet energy resolution.
- b-jet and c-jet tagging and mistagging rates, respectively.
- Light flavour mistagging efficiencies.
- Luminosity.

It should be noted that there are other sources of systematic errors arising from such things as trigger efficiencies that are not taken into account. For the treatment of the jet energy scale and resolution, the procedure follows commonly agreed upon CMS prescriptions [23]. The uncertainty due to the JES is implemented by shifting the jet energies systematically up or down by a relative percentage. For jets having a transverse momentum $p_t > 50\ \mathrm{GeV}/c$, the uncertainty is expected to be 3% because powerful calibration procedures such as the reconstructed mass of hadronic W boson decays in $t\bar{t}$ events [24] are relevant to this energy domain. In the low p_t region down to 20 GeV/c, where the W boson mass calibration is not pertinent, the energy scale is to be set by the photon-jet balancing [25] resulting in a linear increase of the uncertainty from 3% to 10% with decreasing transverse energy. Below 20 GeV/c, only single particle calibration methods are possible with an accuracy of 10%. This leads to the following functional form of the JES uncertainty:

$$\sigma_E^{\mathrm{jet}} \big/ E = \begin{cases} 10\% & p_t < 20\ \mathrm{GeV}/c \\ 10\% - 7\% \cdot (p_t - 20\ \mathrm{GeV}/c)/30\ \mathrm{GeV}/c & 20\ \mathrm{GeV}/c < p_t < 50\ \mathrm{GeV}/c \\ 3\% & p_t > 50\ \mathrm{GeV}/c. \end{cases} \tag{6}$$

To study the effect of jet energy resolution, the jet energy is smeared by an overall 10% according to a Gaussian distribution.

For the b-tagging systematics, the following relative uncertainties in the tagging efficiencies of jets of various flavours are assumed:

- 4% for b- and c-jets.
- 10% for u, d, s and gluon jets, where 'gluon' is now defined in such a way as to not include gluons splitting to charm or bottom jets. This definition yields a comparable mistagging rate to those for u, d and s-jets.

Heavy flavour b- and c-jets are treated identically, since they both have real secondary vertices and any systematic effect should be fully correlated. Light flavour jets have a higher systematic uncertainty because experience has demonstrated that the tagging rates for these jets are difficult to estimate. Even small oversights in the material traversed by a particle, and thus the degree of multiple scattering, can have significant impact upon the misidentification rate of light flavour jets.

In the all-hadron and di-lepton channels, the b-tagging uncertainties are taken into account by simply untagging 4% of the b-jets and by tagging a corresponding fraction of untagged

Table 10. Systematic uncertainties relative to final selection efficiencies (in %) for the semi-leptonic $t\bar{t}H$ channels. Σ is the quadratic sum of all changes in a given row. The last two columns show the absolute uncertainty (in number of events) at the two working points ϵ_{loose} and ϵ_{tight}. The $t\bar{t}4j$ line is given in brackets because this particular background does not give reliable results since the systematical variation is based on only four events remaining after all selection criteria are applied. This, a conservative upper limit of 40% for $t\bar{t}4j$ as estimated from the other backgrounds, is used.

	JES (%)	Jet res. (%)	bc-tagging (%)	uds-tagging (%)	Σ (%)	Number of events ϵ_{loose}	Number of events ϵ_{tight}
$t\bar{t}H$ (115)	5.4	4.4	23.8	0.2	24.8	36	12
$t\bar{t}H$ (120)	3.4	1.6	21.5	0.07	21.9	26	9
$t\bar{t}H$ (130)	3.3	1.1	23.1	0.3	23.3	19	6
$t\bar{t}1j$	23.7	8.5	25.4	0	35.8	166	17
$t\bar{t}2j$	4	5.4	37.8	2.7	38.5	207	25
$t\bar{t}3j$	26.7	6.7	26.7	0	38	127	25
($t\bar{t}4j$)	(175)	(100)	(50)	(0)	(207)	(266)	(0)
$t\bar{t}4j$					$\approx$40	$\approx$50	(0)
$t\bar{t}b\bar{b}$	6.4	1.4	25.3	0.12	26.2	192	62
$t\bar{t}Z$	6.1	2	28.3	1.01	29	10	3
Total background						753 (34%)	133 (31%)

b-jets to obtain the relative uncertainties downward and upward, respectively, in rates. The same procedure is applied for the mistagging rates. This is not possible in the semi-leptonic analysis where a complex likelihood method is applied to tag four b-jets simultaneously. In this case, there is no simple discriminator cut for each jet that can be passed or not. Thus, for the semi-leptonic channel, a different approach is utilized. First, the discriminator threshold required to obtain the tagging efficiencies used in the analysis is determined. Then, the b-tagging discriminator itself is shifted by a value corresponding to a shifted b-tagging working point. This modification can be applied at the very beginning of the analysis and is therefore easy to implement. The procedure is detailed in [3]. The estimation of the uncertainties is accomplished by simultaneously varying the discriminator for b- and c-jets according to the given percentages. The variation for light flavour jets is done independently. In both cases, the variation is performed 'upwards' and 'downwards', and the direction which gives the larger change in event yields is the one quoted in table 10 for the semi-leptonic channel. The change in final event numbers is also given at the two working points ϵ_{loose} and ϵ_{tight}. The relative uncertainties are calculated at the *loose* working points and it is assumed that the same uncertainties apply at the *tight* working point. This is justified by the fact that only the choice of the b-tagging working point differs between the two cases, and the mistagging efficiencies are roughly linear as a function of the b-discriminator cut. The propagation of the errors to the *tight* working point is necessary because of the small statistical significance of some of these calculations. For instance, only four events remain after all selection criteria at the *loose* working point in the $t\bar{t}4j$ sample, which leads to a relative statistical error of $\sim\sqrt{4}/4 = 50\%$. Obviously, the numbers obtained for this specific sample cannot be considered to be very meaningful. The statistical errors due to the finite sizes of data samples are given in table 4 for all samples in order to be able to judge the reliability of the numbers obtained. Fortunately, reliable numbers are available for all signal samples, and also for the $t\bar{t}1j$, $t\bar{t}2j$ and $t\bar{t}b\bar{b}$ and $t\bar{t}Z$ samples. From these samples, conservative estimations for $t\bar{t}4j$ are possible as indicated in table 10.

An interesting observation is the fact that the impact of the 10% uncertainty in the light flavour mistagging rate is generally below 3%. This confirms the observation that a large

Table 11. Systematic uncertainties relative to selection efficiencies (in %) for the di-lepton $t\bar{t}H$ channel. Σ is the quadratic sum of all changes in the given row, excluding the statistical uncertainties (provided for comparison only). Σ also includes the anticipated 3% uncertainty in the luminosity which is the same for all samples. The last two columns show the absolute uncertainty (in number of events) at the two working points ϵ_{loose} and ϵ_{tight}.

Di-lepton	JES (%)	Jet res. (%)	b- and c-tagging (%)	uds-tagging (%)	Σ (%)	Number of events ϵ_{loose}	Number of events ϵ_{tight}
$t\bar{t}H$ (115)	2.34	1.81	9.46	0.514	10.4	17	3
$t\bar{t}H$ (120)	3.65	1.77	9.38	0.651	10.7	14	2
$t\bar{t}H$ (130)	3.41	1.18	9.76	0.787	10.9	9	1
$t\bar{t}b\bar{b}$	5.26	0.851	8.62	0.572	10.6	114	17
$t\bar{t}1j$	25.7	3.81	10.7	9.71	29.9	380	13
$t\bar{t}2j$	12.1	1.13	11.0	5.97	17.7	477	15
$t\bar{t}3j$	6.0	7.33	10.7	7.71	16.5	220	5
$t\bar{t}4j$	16.3	5.81	6.51	7.85	20.3	532	19
$t\bar{t}Z$	6.87	3.16	8.96	3.58	12.6	13	2
Total background ϵ_{loose}	13.4	3.7	9.3	6.6	18.3	1660	
Total background ϵ_{tight}	11.2	2.8	9.0	4.8	15.7		66

Table 12. Systematic uncertainties relative to selection efficiencies (in %) for the all-hadron $t\bar{t}H$ channel. Σ is the quadratic sum of all changes in the given row, with the exclusion of the statistical uncertainties (provided for comparison only). The $t\bar{t}H$ signal row corresponds to $m_H = 120$ GeV. Σ also includes the anticipated 3% uncertainty in the luminosity which is the same for all samples. The last two columns show the absolute uncertainty (in number of events) at the two working points, ϵ_1 and ϵ_2.

All-hadron	JES (%)	Jet res. (%)	b- and c-tagging (%)	uds-tagging (%)	Σ (%)	Number of events ϵ_{loose}	Number of events ϵ_{tight}
$t\bar{t}H$ (120)	17.6	7.0	6.0	0.6	20.1	63.1	9.0
$t\bar{t}b\bar{b}$	14.2	5.9	6.4	0.7	16.9	201	19
$t\bar{t}1j$	43.7	9.2	2.3	4.6	45.1	388	22
$t\bar{t}2j$	23.1	10.8	5.9	5.9	27.0	539	15
$t\bar{t}3j$	18.5	3.7	3.7	3.7	19.8	377	7
$t\bar{t}4j$	5.6	0.6	6.8	4.5	10.4	689	8
$t\bar{t}Z$	17.8	8.4	7.0	1.7	21.2	26	2
qcd 170	52.4	15.9	7.9	4.8	55.6	2670	42
qcd 120	100	0.0	0.0	0.0	100	83	95
Total background ϵ_{loose}	24.7	7.1	6.4	4.4	27	4760	
Total background ϵ_{tight}	39.1	6.3	4.9	2.9	40		202

portion of the misidentified $t\bar{t}Nj$ events consist of events in which gluons split into real b-jets or a W boson decays to a charm jet.

The analogous numbers for the all-hadron and di-lepton channels are given in tables 11 and 12.

The impact of the systematic uncertainties on the final significance is given in tables 13–15. Under the assumption that these systematic errors follow a Gaussian distribution, the error on the number of background events $\mathrm{d}B$ has to be included quadratically, so that the appropriate significance (σ) is

$$\sigma = \frac{S}{\sqrt{B + \mathrm{d}B^2}}. \tag{7}$$

Table 13. Significance of the semi-leptonic channels before and after taking into account the uncertainty dB in the total number of background events due to systematic uncertainties. Results for the two working points ϵ_{loose} and ϵ_{tight} are obtained under the assumption that the systematic uncertainties are the same in both cases.

	S/B	$S/\sqrt{B}$	$S/\sqrt{B+\mathrm{d}B^2}$
	ϵ_{loose}		
$t\bar{t}H$ ($m_H = 115\ \text{GeV}/c^2$)	0.07	3.1	0.20
$t\bar{t}H$ ($m_H = 120\ \text{GeV}/c^2$)	0.053	2.5	0.16
$t\bar{t}H$ ($m_H = 130\ \text{GeV}/c^2$)	0.036	1.7	0.11
	ϵ_{tight}		
$t\bar{t}H$ ($m_H = 115\ \text{GeV}/c^2$)	0.11	2.3	0.35
$t\bar{t}H$ ($m_H = 120\ \text{GeV}/c^2$)	0.09	1.9	0.29
$t\bar{t}H$ ($m_H = 130\ \text{GeV}/c^2$)	0.06	1.2	0.19

Table 14. Significance of di-lepton channel before and after taking into account the uncertainty dB in the total number of background events due to systematics. The result is shown for two working points which correspond to different sets of criteria for number of jets and b-tagged jets, but assuming the same systematic uncertainties (as computed at the *loose* working point), for both.

Di-lepton	S	S/B	$S/\sqrt{B}$	$S/\sqrt{B+\mathrm{d}B^2}$
4–7 jets, 3–4 b-tagged (ϵ_{loose})				
$t\bar{t}H$ ($m_H = 115\ \text{GeV}/c^2$)	170	0.018	1.8	0.10
$t\bar{t}H$ ($m_H = 120\ \text{GeV}/c^2$)	130	0.015	1.4	0.08
$t\bar{t}H$ ($m_H = 130\ \text{GeV}/c^2$)	82	0.009	0.9	0.05
4–6 jets, 4–6 b-tagged (ϵ_{tight})				
$t\bar{t}H$ ($m_H = 115\ \text{GeV}/c^2$)	29	0.069	1.4	0.42
$t\bar{t}H$ ($m_H = 120\ \text{GeV}/c^2$)	19	0.045	0.9	0.27
$t\bar{t}H$ ($m_H = 130\ \text{GeV}/c^2$)	12	0.029	0.6	0.18

Table 15. Significance of the all-hadron channel before and after taking into account the uncertainty dB in the total number of background events due to systematic uncertainties. The result is shown for the two different working points as described in section 6.3.

Hadron	S	S/B	$S/\sqrt{B}$	$S/\sqrt{B+\mathrm{d}B^2}$
Softer b-tag discriminator cut (ϵ_{loose})				
$t\bar{t}H$ ($m_H = 115\ \text{GeV}/c^2$)	350	0.020	2.6	0.07
$t\bar{t}H$ ($m_H = 120\ \text{GeV}/c^2$)	310	0.018	2.4	0.07
$t\bar{t}H$ ($m_H = 130\ \text{GeV}/c^2$)	210	0.012	1.6	0.05
Harder b-tag discriminator cut and event centrality cut (ϵ_{tight})				
$t\bar{t}H$ ($m_H = 115\ \text{GeV}/c^2$)	44	0.087	2.0	0.22
$t\bar{t}H$ ($m_H = 120\ \text{GeV}/c^2$)	45	0.089	2.0	0.22
$t\bar{t}H$ ($m_H = 130\ \text{GeV}/c^2$)	27	0.054	1.2	0.13

Given the rather low anticipated significance values presented in tables 13–15, one can ask how precisely the backgrounds have to be known in order to reach a significance that would allow one to claim an observation of this channel.

Figure 17 shows the behaviour of the significance as a function of the background uncertainty for the semi-leptonic channel.

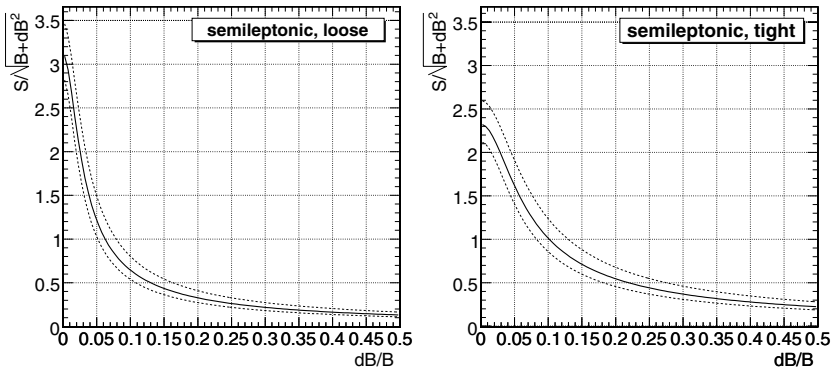

Figure 17. Significance $S/\sqrt{B + \mathrm{d}B^2}$ as a function of the fractional uncertainty $\mathrm{d}B/B$ in the total background at the *loose* and *tight* working points, assuming Gaussian errors, for a Higgs boson mass of $m_H = 115\ \mathrm{GeV}/c^2$ and an integrated luminosity of 60 fb^{-1}. The dashed line corresponds to a variation of the background cross section of 20% due to the theoretical uncertainty.

This figure also shows the uncertainty due to the theoretical knowledge of the background cross section which is varied by 20% up or down in the plot. The tight working point shows better results compared to the loose working point as soon as the background uncertainty reaches realistic values above 5%.

The uncertainty analysis presented above reflects current uncertainties on central values of such things as tagging rates. Once real data are available, control samples will allow the evaluation of the relevant quantities with smaller systematic uncertainties. Nevertheless, the main conclusion to be drawn from these studies is that the uncertainty in the background level will need to be much less than 10% before an observation of this channel is possible.

This is an enormous challenge. For some backgrounds, such as light flavour mistags, results from the Tevatron experiments lead one to conclude that an uncertainty below 10% is achievable. At the LHC one can take advantage of the abundance of $\mathrm{t\bar{t}}$ plus jets events as well as $b\bar{b}$ events obtained via the inclusive lepton trigger streams. However, for irreducible backgrounds such as $\mathrm{t\bar{t}b\bar{b}}$ this is not possible because the theoretical uncertainties on the cross section are currently estimated to be as high as 30–50%. Further improvement will therefore only be possible when the theoretical uncertainty is very substantially reduced by the inclusion of higher order corrections.

8.1. Combined significance

The event samples for the various channels studied in this note have no overlap and all of the analyses are 'counting experiments' therefore the individual results can be combined. Several approaches to combining significances have been investigated. The simplest approach is to add the individual signal and background yields of the three channels. This method is only optimal in the case where the event numbers are similar. Another simple approach would be to add the significances in quadrature.

An optimal combination is obtained via the more complicated likelihood method of Cranmer *et al* [27]. The significance values obtained by simply combining signal and background yields are compared to the results of the more optimal method of Cranmer in table 16 before and after inclusion of systematic uncertainties for the three Higgs mass hypothesis and for the loose and tight working points.

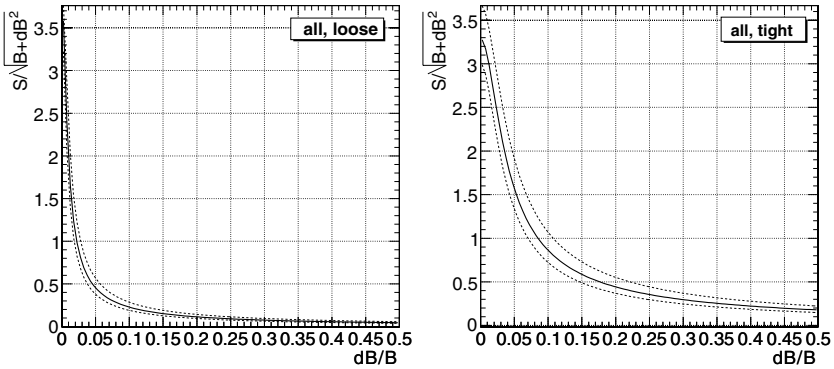

Figure 18. Combined significance (di-lepton + semi-leptonic + all-hadron) $S/\sqrt{B+\mathrm{d}B^2}$ as a function of the fractional uncertainty $\mathrm{d}B/B$ in the background at the *loose* (left) and *tight* (right) working points, assuming Gaussian uncertainties, for a Higgs boson mass of $m_H = 115$ GeV/c^2 and an integrated luminosity of 60 fb^{-1}. The dashed line corresponds to a variation of the background cross section of 20% due to the theoretical uncertainty.

Table 16. Combined significances of all analyses before and after inclusion of systematic uncertainties for the three Higgs mass hypothesis and for the loose and tight working points. Values under the headings of '$S/\sqrt{B+\mathrm{d}B^2}$' correspond to the result obtained by simply combining signals and backgrounds for the various analyses (which used exclusive datasets for this purpose). The values under the headings of 'likelihood' are those obtained with the more optimal likelihood method of Cranmer *et al* [27].

	No systematics		With systematics	
m_H (GeV/c^2)	$S/\sqrt{B}$	Optimal	$S/\sqrt{B+\mathrm{d}B^2}$	Optimal
	Loose working point			
115	3.89	4.40	0.130	0.0900
120	3.32	3.69	0.111	0.0612
130	2.21	2.48	0.0738	0
	Tight working point			
115	3.29	3.30	0.478	0.374
120	2.83	2.89	0.411	0.299
130	1.74	1.79	0.253	0.159

In order to compare the behaviour of the combined significance with figure 17, the method of adding event yields is used, which can be considered to be a sufficient approximation in this case. The solid central line in figure 18 shows the combined significance $S/\sqrt{B+\mathrm{d}B^2}$ and how it degrades as a function of $\mathrm{d}B/B$ for both the *loose* and *tight* working points. The *tight* working points shown in the right-hand plot of figure 18 give the best results after inclusion of systematics.

8.2. Prospects for improvements

A number of possibilities to improve the results remain to be implemented and tested and generally rely on an improved understanding of the performance of the CMS detector and improvements in analysis tools. For example, CMS jet reconstruction algorithms are in their initial stages and are expected to improve substantially over time, particularly with the inclusion

of tracking and muon detector information and taking into account the fine granularity of the electromagnetic calorimeter. To this end, an 'energy flow' project has been launched within CMS.

Additionally, during the operation of CMS, more complex triggers will likely be implemented. A dedicated $t\bar{t}H$ trigger could improve the signal selection efficiency significantly. For example, simply combining information about jets with single-lepton triggers would allow the lepton thresholds to be lowered to increase efficiencies.

Finally, and importantly, the exploitation of differences between signal and background event kinematics could be used to extract a clearer signal. However, current uncertainties in Monte Carlo event generation and detector modelling make it difficult to perform reliable performance estimates since small changes in any of the factors determining the final event characteristics can have a significant impact on distributions of kinematic variables. The important point to keep in mind, once again, is that the availability of large control samples from a wide variety of trigger streams will eventually enable fruitful use of event kinematics. One should not however underestimate the work and time required for such an endeavour.

9. Summary and outlook

A detailed Monte Carlo and full simulation study of the process $t\bar{t}H$ with $H \rightarrow b\bar{b}$ has been performed in order to evaluate its discovery potential for an integrated luminosity of 60 fb^{-1}.

The conclusions drawn from this analysis are substantially more pessimistic than those of previous studies [21, 22], which claimed significances in excess of 5. This change in viewpoint is due to a greater degree of realism that was made possible by more advanced tools for event generation, detector simulation and physics reconstruction that were not available for previous studies. These tools make it possible to obtain reasonable estimates for systematic uncertainties. For example, mistagging of light flavour jets cannot be reliably estimated without a full detector simulation, based upon a relatively detailed material description of the apparatus, followed by an equally detailed track reconstruction program. Mistagging of light flavour jets in $t\bar{t}jj$ events thus proved to be a substantially more serious problem than had been foreseen in earlier studies that made use of parameterized b-tagging. On the other hand, it has been noted that the availability of large control samples of $t\bar{t}$ events should enable b-tagging of jets with high transverse energy to be very well understood at the LHC. This would enable more precise estimates, and further suppression, of light quark and charm jet tagging relative to b-tagging. Experience with real data will also improve jet reconstruction and energy measurements which will then enhance the efficiency of many of the techniques described in this note. On the other hand, irreducible backgrounds, such as $t\bar{t}b\bar{b}$, will remain a serious problem for the observation of $t\bar{t}H$ with $H \rightarrow b\bar{b}$ until more precise calculations of the cross sections for these background processes are available. At present, the uncertainty in the production cross section for the irreducible $t\bar{t}b\bar{b}$ background is estimated to be between 30%–50% [26].

Acknowledgments

The authors would like to acknowledge R Cousins and K S Cranmer for their assistance in providing us with a more optimal method for combining significances. We also would like to thank our referees whose suggestions and comments have led to a much improved analysis and paper. The CMS experiment is the manifestation of many years of creativity and intense

effort by thousands of scientists, engineers, technicians and students worldwide. The study presented here, and others like it, would be neither possible nor meaningful without their contributions, made possible by the generous support of worldwide funding agencies who are gratefully acknowledged here.

References

[1] Additional information about all CMS subsystems is contained in their respective Technical Design Reports at http://cmsdoc.cern.ch/docLHCC.shtml
[2] Benedetti D, Cucciarelli S, Hill C, Incandela J, Koay S A, Riccardi C, Santocchia A, Schmidt A, Torre P and Weiser C 2006 Search for $H \to b\bar{b}$ in association with a $t\bar{t}$ pair at CMS *CMS Note 2006/119*
[3] Benedetti D, Hill C, Incandela J, Koay S A, Riccardi C, Santocchia A, Schmidt A, Torre P and Weiser C 2007 Refined analysis of $t\bar{t}H$ with $H \to b\bar{b}$ at CMS *CMS Analysis Note 2007/001*
[4] Schmidt A 2006 Search for $H \to b\bar{b}$ in association with a $t\bar{t}H$ pair in proton–proton collisions at $\sqrt{s} =$ 14 TeV *PhD Thesis* Institut für Experimentelle Kernphysik, Universität Karlsruhe (TH) (IEKP-KA/2006-20)
[5] Benedetti D Search for $H \to b\bar{b}$ in association with a $t\bar{t}H$ pair at CMS *PhD Thesis* Universitá Degli Studi Di Perugia
[6] Dubinin M *et al* 1998 CompHEP—a package for evaluation of Feynman diagrams and integration over multi-particle phase space *INP-MSU* **41** 542
[7] Sjostrand T, Lonnblad L and Mrenna S 2001 PYTHIA 6.2: physics and manual *Preprint* hep-ph/0108264
[8] Beenhakker W, Dittmaier S, Kramer M, Plumper B, Spira M and Zerwas P M 2003 NLO QCD corrections to t anti-t H production in hadron collisions *Nucl. Phys.* B **653** 151–203
[9] Mangano M L, Moretti M, Piccinini F, Pittau R and Polosa A 2003 ALPGEN, a generator for hard multiparton processes in hadronic collisions *J. High Energy Phys.* JHEP07(2003)001
Mangano M L, Moretti M and Pittau R 2002 Multijet matrix elements and shower evolution in hadronic collisions: W B Bbar + N jets as a case study *Nucl. Phys.* B **632** 343–62
Caravaglios F, Mangano M L, Moretti M and Pittau R 1999 A new approach to multijet calculations in hadron collisions *Nucl. Phys.* B **539** 215–32
[10] CMS Collaboration 2002 The TriDAS Project Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger *CERN/LHCC 2002/26*
[11] CMS Collaboration 2006 The Physics Technical Design Report: Volume I *CERN/LHCC 2006-001*
[12] CMS Collaboration 2007 The Physics Technical Design Report: Volume II *CERN/LHCC 2007-001*
[13] Haifeng Pi, Avery P, Green D, Rohlf J and Tully C 2006 Measurement of missing transverse energy with the CMS detector at the LHC *CMS Note 2006/035*
[14] James E, Maravin Y, Mulders M and Neumeister N 2006 Muon identification in CMS *CMS Note 2006/010*
[15] D'Hondt J, Lowette S, Buchmuller O, Cucciarelli S, Schilling F P, Spiropulu M, Mehdiabadi S P, Benedetti D and Pape L 2006 Fitting of event topologies with external kinematic constraints in CMS *CMS Note 2006/0023*
[16] Heister A, Kodolova O, Konoplianikov V, Petrushanko S, Rohlf J, Tully C and Ulyanov A 2006 Measurement of jets with the CMS detector at the LHC *CMS Note 2006/036*
[17] Weiser C 2006 A combined secondary vertex based B-tagging algorithm in CMS *CMS Note 2006/014*
[18] Santocchia A 2006 Optimization of jet reconstruction settings and parton-level correction for the ttH channel *CMS Note 2006/059*
[19] Müller T, Piasecki C, Quast G and Weiser C 2006 Inclusive secondary vertex reconstruction in jets *CMS Note 2006/027*
[20] Bocci A, Demin P, Ranieri R and Visscher S de 2006 Tagging b jets with electrons and muons at CMS *CMS Note 2006/043*
[21] Drollinger V, Muller Th and Denegri D 2001 Searching for Higgs bosons in association with top quark pairs in the $H \to bb$ decay mode *CMS Note 2001/054*
[22] Kappler S, Müller T, Quast G and Weiser C 2004 Progress report on studies of the channel $t\bar{t}H$ with $H \to b\bar{b}$ and $t\bar{t} \to WWb\bar{b} \to qq'\mu\bar{\nu}_\mu b\bar{b}$ for CMS in full simulation *CMS Internal Note 2004/048*
[23] Rohlf J and Tully C 2006 Recommendations for jet and missing transverse energy reconstruction settings and systematics treatment *CMS Internal Note 2006/025*
[24] D'Hondt J, Lowette S, Heyninck J and Kasselmann S 2006 Light quark jet energy scale calibration using the W mass constraint in single-leptonic $t\bar{t}$ events *CMS Note 2006/025*

[25] Konoplianikov V, Kodolova O and Ulyanov A 2006 Jet calibration using γ+jet events in the CMS detector *CMS Note 2006/042*
[26] Mangano M (CERN) Private communication
[27] Cranmer K S, Mellado B, Quayle W and Wu S L 2033 Challenges in moving the LEP Higgs statistics to the LHC *Preprint* physics/0312050