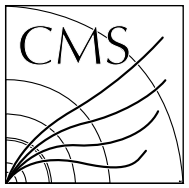
Available on CMS information server

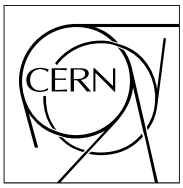
CMS AN -2012/115



The Compact Muon Solenoid Experiment

Analysis Note

The content of this note is intended for CMS internal use and distribution only



02 April 2012 (v4, 31 May 2012)

Search for the standard model Higgs Boson in the decay channel $H \to ZZ \to 2\ell 2q$ at CMS

CERN
CIEMAT, Madrid, Spain
INFN Sezioni di Napoli, Napoli, Italy
Purdue University, West Lafayette, IN, USA
Universidad Autonoma de Madrid, Madrid, Spain
Johns Hopkins University, Baltimore, MD, USA
University of Rochester
Panjab University, India
National Central University, Taiwan

Abstract

A search for the standard model Higgs boson decaying to two Z bosons with subsequent decay to a final state with two leptons and two quark-jets, $H \to ZZ^{(*)} \to (q\bar{q})(\ell^-\ell^+)$, is presented. Data corresponding to an integrated luminosity of about XXX fb^{-1} of LHC proton-proton collisions were collected and analyzed by the CMS experiment. The selection to discriminate between signal and background events is based on kinematic and topological quantities, which include the angular spin correlations of the decay products. The events are classified according to probability of the jets to originate from quarks of light or heavy flavor or from gluons. No evidence for a Higgs boson is found and upper limits on the Higgs boson production cross section are set in the range of masses between XXX and XXX GeV, and between 200 and 1000 GeV. Prospects for a Beyond the Standard Model boson exclusion are discussed.

Contents

1 Introduction

The Higgs boson is an essential element of the Standard Model (SM) of particles and their interactions explaining the origin of mass and playing a key role in the mechanism of electroweak symmetry breaking. Precision measurements of the W boson and top quark masses together with the Z pole variables constrain the mass of the Higgs boson which is a free parameter of the theory: currently the 95% one-sided confidence level upper limit on M_{H} is 158 GeV. Direct searches for the Higgs at LEP-II [1] set a 95% C.L. lower limit on M_{H} of 114.4 GeV, while the region between 158 GeV and 175 GeV has been excluded by the Tevatron experiments [2, 3]. The 95% C.L. upper limit on the Higgs mass increases from 158 GeV to 185 GeV when the results from the direct searches are included [4].

XXX NEED TO REVIEW THE RESULTS AND REFERENCE THE PROPER ONES. ALSO ADD 2011 results from Tevatron and LHC. XXX

The CMS collaboration is performing searches for the Higgs in a several decay modes. This comprehensive effort aims at gaining sensitivity over a large range of Higgs masses by combining many different analyses. We expect that this effort will finally explore the region with $M_{\mathrm{H}} \geq 2m_{\mathrm{Z}}$ which is not probed at the Tevatron. In this note we report a detailed study of the prospect for the search for the Higgs boson in $\rightarrow$ ZZ when one of Z decays as $\mathrm{Z} \rightarrow \ell^{-}\ell^{+}$ and the other as $\mathrm{Z} \rightarrow q\bar{q}$. The dominant background is Z+jets production. Other minor backgrounds are $t\bar{t}$ and diboson production. The data and Monte Carlo (MC) samples used in this analysis are described in Section 2. The analysis strategy including specific channel optimization and the enhanced signal-to-background selection are discussed from Sections 3 to 6. Contamination is estimated from background-enriched control samples. Methods to provide a data-driven derivation of the background is presented in Section 7. A detailed discussion of the systematic errors, including the impact of the pile-up events expected in the 2012 LHC run, as well as the modeling of the b-tagging algorithms in CMS, is described in Section 8. The strategy and the experimental were validated using 4.9 fb^{-1} data collected by the CMS experiment in 2011 as described in [**?** **?**]. The analysis mostly follows the analysis on 2011 data, with the following modifications and improvements: Lepton Id, b-tagging and Quark-Gluon discrimination are reworked according to POG recommendations; the analysis is extended to higher higgs masses; and Higgs production in VBF is treated separately from the gluon fusion process.

2 Samples

In this section we describe the data and MC samples that are used in the analysis.

2.1 Data

We use XXX fb^{-1} of data collected in 2012 and prompt-reco'ed or re-reco'ed with CMSSW_5_2_X The data samples are listed in Table 1. We use only lumisections that has been declared good for analysis by the central certification team.

2.2 Signal Simulation

We also use signal MC samples generated with POWHEG listed in Table 2.

2.3 Background Simulation

The dominant background in the $H \to ZZ \to 2l2q$ analysis is the inclusive Z production with jets, that is Z+jets background. For the case of low Higgs mass ($M_H < 2 \times m_Z$) the Z is off-shell (Z^*+jets). The primary samples that we use has been produced with MadGraph with the latest conditions and is listed in Table 3.

XXX exclusive samples? b-enriched samples? XXX

We also checked Sherpa simulation of the process, listed in the same Table. Sherpa implements a different "matching" procedure with the Parton Shower with respect to Madgraph. Also the Parton Shower simulation itself, is different between the two generators. The comparison between Sherpa and Madgraph is therefore a reasonable estimate of the uncertainties related with the MC modeling of Z+jets.

Further background arises from $t\bar{t}$, tW, ZZ, WZ production, and the samples used are shown in Table 4.

In all cases the Spring12 production with releases CMSSW_5_2_X is used.

3 Reconstruction and Base Selection

Two independent software packages are employed in the analysis of the channel $X \to ZZ \to 2l2j$: one makes use of the Physics Analysis Toolkit (PAT) [5], one runs directly on AOD objects. The analysis redundancy ensures an overall robustness, as one workflow may be employed to cross-check the other at any time. High level of agreement of better than 1% is achieved between the two complementary software packages.

Input objects to the analysis are GSF electrons [6], Global Muons [7], and Particle Flow jets [8, 9]. In order to avoid to double count the leptons inside the jets, a $\Delta R > 0.5$ cut is applied between the leptons chosen for the Z reconstruction and the jets.

3.1 Trigger

The results presented in this note are based on the SingleMu, DoubleMu and DoubleElectron datasets. Each of these datasets contain at least one un-prescaled trigger with looser requirements than our offline selections. The lowest threshold un-prescaled trigger is used and this trigger changes as instanteous luminosity rises. These triggers are HLT_IsoMu24 for the SingleMu dataset
HLT_Mu17_TkMu8 and HLT_Mu17_Mu8 for the DoubleMu dataset and
HLT_Ele17_CaloIdT_TrkIdVL_CaloIsoVL_TrkIsoVL_Ele8_CaloIdT_TrkIdVL_CaloIsoVL_TrkIsoVL for the DoubleElectron dataset. More details are available in the corresponding analysis note [10].

Since the level of precision of the trigger emulation in simulation is not well known, the trigger efficiency is computed from data (as reported in [11]) and the Monte Carlo samples are

Table 1: Data samples used in analysis.

/SingleMu/XXX/AOD
/DoubleMu/Run2012A-PromptReco-v1/AOD
/DoubleElectron/Run2012A-PromptReco-v1/AOD
/MuEG/Run2012A-PromptReco-v1/AOD

reweighted accordingly.

3.2 Lepton Selection

The electron candidates were reconstructed with the GSF algorithm. They were required to satisfy the following conditions to ensure good reconstruction performances [**?**]:

- the electron track is required to match the supercluster to within 0.007(0.009) in η in the barrel(endcap)
- the electron track is required to match the supercluster to within 0.8(0.1) in η in the barrel(endcap)
- the shower-shape variable $\sigma_{i\eta i\eta}$ is required to be lower than 0.01 (0.03) in the barrel(endcap)
- the HCAL energy behind the electrons cluster relative to the electron enerfy (H/E) is required to be lower than 0.12(0.10) in the barrel(endcap).
- the track has transverse impact parameter $d_0 < 0.4$ mm w.r.t. the primary vertex
- the longitudinal distance of the track w.r.t. the primary vertex is $d_z < 2$ mm
- the electrons must pass a cut of 15% on the relative PF isolation with a cone size of 0.3, using ρ·effective area corrections to reduce PU sensitivity.

A fiducial cut is applied to stay inside the ECAL acceptance. Electrons are rejected if $1.4442 < |\eta| < 1.566$ and $|\eta| < 2.5$.

Global muons were required to satisfy the following identification criteria:

- the muon passes global reconstruction ("isGlobalMuon")
- the muon is recognized by the PF reconstruction ("isPFMuon")
- the normalized χ^2 of the associated gloabl track is smaller than 10.
- at least one valid hit in the muon system is associated to the muon.
- muon segments in at least two muon stations are associated to the muon
- its tracker track has transverse impact parameter $d_{xy} < 2$ mm w.r.t. the primary vertex
- the longitudinal distance of the tracker track w.r.t. the primary vertex is $d_z < 5$ mm
- the assocaited tracker track must contain at least one pixel hit.

Table 2: Signal MC samples generated with POWHEG. Muon, electron, and tau final states in the leptonic Z decay are considered.

MC ID	name	mass	Γ (GeV)	$\sigma \times \mathcal{B}_{ZZ} \times \mathcal{B}_{2l2j}$ (fb)
2701	/GluGluToHToZZTo2L2Q_M-200_7TeV-powheg-pythia6/ Fall11-PU_S6_START44_V9B-v1/AODSIM	200	?	?
2707	/GluGluToHToZZTo2L2Q_M-300_7TeV-powheg-pythia6/ Fall11-PU_S6_START44_V9B-v1/AODSIM	300	?	?
2711	/GluGluToHToZZTo2L2Q_M-400_7TeV-powheg-pythia6/ Fall11-PU_S6_START44_V9B-v1/AODSIM	400	?	?
2715	/GluGluToHToZZTo2L2Q_M-500_7TeV-powheg-pythia6/ Fall11-PU_S6_START44_V9B-v1/AODSIM	500	?	?

Table 3: Spring12 Monte Carlo samples with Z+jets final state. A K-factor of 1.19 has been applied to Madgraph.

MC ID	name	σ LO(NLO) [pb]	lumi LO(NLO) [fb^{-1}]
5591	/DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball/ Summer12-PU_S7_START50_V15-v1/AODSIM	2950.0(3503.71)	4.9(4.1)

Table 4: Monte Carlo samples with $t\bar{t}$, tW, ZZ, WZ. The last column shows equivalent luminosity of the available MC sample.

MC ID	name	σ LO(NLO) [pb]	lumi LO(NLO) [fb^{-1}]
5502	/TTJets_TuneZ2star_8TeV-madgraph-tauola/ Summer12-PU_S7_START52_V5-v1/AODSIM	136.3(225.197)	8.83(5.34)
-1	/ZZ_TuneZ2star_8TeV_pythia6_tauola/ Summer12-PU_S7_START50_V15-v1/AODSIM	5.196(8.25561)	1544(972)

- the assocaited tracker track must contain hits from more than five layers of the tracker
- the muons must pass a cut of 0.12 on the combined, PU corrected isolation[1]

Muons are required to be in the pseudorapidity range $|\eta| < 2.4$.

The leptons are required to have transverse momentum $p_T > 20$ or 40 GeV, for the lower or higher momentum lepton for events that enter the high mass aprt of the analysis and $p_T > 10$ or 20 GeV for events that enter the low mass analysis. Some of the kinematic distributions for signal and background are shown in Section **??**. The di-lepton invariant mass is shown in Fig. 1. In the analysis the invariant mass of the $Z \to l^+l^-$ boson is required to be 70 GeV $< m_{ll} <$ 110 GeV for the high mass part of the analysis and 20 GeV $< m_{ll}$ GeV for the low mass part of the analysis. The upper bound on $m_{\ell\ell}$ come naturally from the kinematic constraint $m_{\ell\ell} < m_{ZZ} - m_Z$, where $m_Z \simeq 91.2$ GeV. Here the dijet invariant mass is constrained to the m_Z value when the m_{ZZ} mass is calculated, using the same kinematic fit approach.

XXX add low mass plot here XXX

3.3 Jet Selection

The PF jets are reconstructed with the `anti-k`$_\mathrm{T}$ algorithm [12] with radius parameter set to $R = 0.5$. Jet-energy corrections are applied to data and MC as explained in [13]. The jet corrections applied (excluding those for correcting foTo study these PU effects, signal events from a MC sample with 400 GeV Higgs mass which pass the preselection cuts listed in Section 3.7 are considered. The Fastjet algorithm has been applied to correct the PU energy in each jet.r pile-up described in Section **??**) were the L2 and L3 corrections.

Jets are required to be inside the tracker acceptance ($|\eta| < 2.4$) thus allowing high reconstruction efficiency and precise energy measurements using PF techniques. A very loose jets identi-

[1] PFIsoCorr= PF(ChHad PFNoPU) + Max ((PF(Nh+Ph) - ρ' EACombined),0.0)), where "ChHad PFNoPU" are PF charged hadrons after PU removal, "Nh" and "Ph" are PF neutral hadrons and photons respectively, ρ' is the average PU energy density in the central region after charged PU subtraction and EACombined is the effectefive area from the POG recommendation.

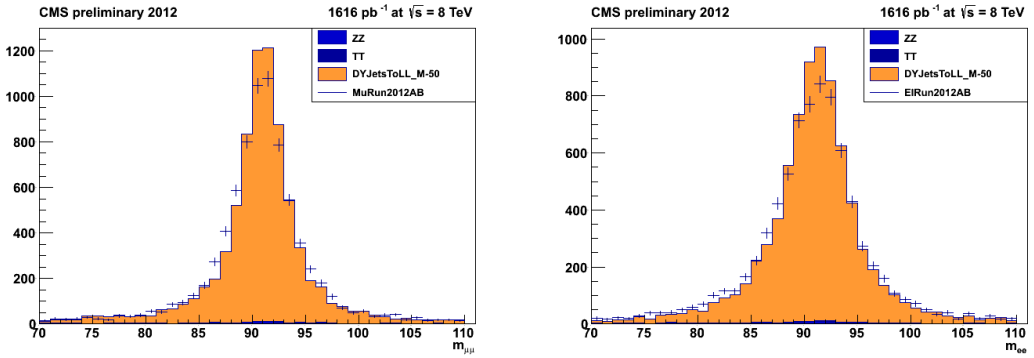


Figure 1: Dilepton invariant mass after loose selection requirements in the analysis, 2012 data and MC simulation with dominant contributions. Left: Muon channel. Right: Electron channl.

fication is applied to remove fakes due to calorimeter noise:

- fraction of energy due to neutral hadrons < 0.99;
- fraction of energy due to neutral EM deposits < 0.99;
- number of constituents > 1;
- number of charged hadrons candidates > 0;
- fraction of energy due to charged hadrons candidates > 0;
- fraction of energy due to charged EM deposits < 0.99.
- $\beta >= 0.2$ (see Subsection 3.5)

XXX are these upt o date?XXX

A preselection cut $p_T > 30$ GeV is applied to all jets since in signal jets are expected to come from the decay of a highly energetic Z boson. In the analysis a cut $75 < m_{JJ} < 105$ GeV (corresponding to $\sim 2\sigma$) is applied in order to reduce the dominant Z+jets background. The invariant mass of the 2 leading jets is shown in Fig. 2, the Z mass resolution is about 7 GeV.

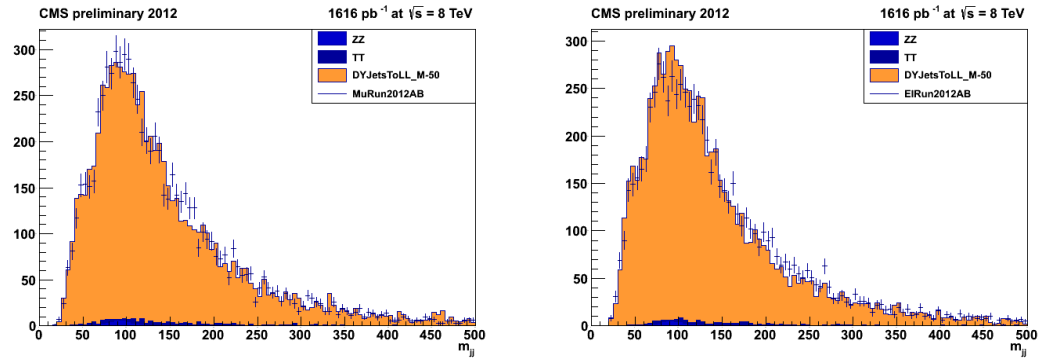


Figure 2: Dijet invariant mass . Left: comparison of 2010-2011 data and MC simulation with dominant contributions. Center: Right: comparison of signal MC with three different generated mass values and dominant background contributions.

3.4 B-Tagging of jets

Due to the relatively large branching fraction of the Z boson decaying into a pair of bottom-antibottom quarks, it is advantageous to exploit the possibility to identify jets originating from heavy-flavour quarks in order to enhance the sensitivity of the analysis.

However, no requirement on b-probabilities is imposed at this point in the analysis. Instead the b-tagging information is used to optimize the further optimize the analysis in categories of events with different likelihoods to contain heavy quarks. Details of the b-tagging algorithm are described in section 4 and its further use detailed in section 6.1

3.5 Studies relative to PU

The presence of additional interactions with respect to the primary one, known as Pile-up (PU), is expected to affect this analysis in the following ways:

- additional energy from PU get added to the jets from the main interaction
- additional low p_T jets fully composed of PU energy get added to the event
- tracks and calorimetric towers from PU energy deposits get added to the jets from the main interaction thus biasing their angles
- tracks and calorimetric towers from PU energy deposits get added to isolation variables of the leptons, reducing the isolation efficiency.

The amount of PU interaction per event in data and MC is shown in Fig. 3. In the following this distribution in MC has been re-weighted to match the data. In Fig. 4 we show consistency in the number of reconstructed vertexes between data and MC after MC samples have been re-weighted.

Various algorithms are available to correct for PU effects. The so-called Fastjet and L1-offset corrections remove the additional energy released in the event which is expected to come from PU interactions. The charged particles coming from PU can also be removed before the jet clustering requiring that all the tracks come from the primary vertex. In this case also the PU effect on jet angles is partially corrected. Alternatively, only jets with a sizable amount of tracks coming from primary vertex can be considered.

To study these PU effects, signal events from a MC sample with 400 GeV Higgs mass which pass the preselection cuts listed in Section 3.7 are considered. The Fastjet algorithm has been applied to correct the PU energy in each jet. In Fig. 5 the transverse momentum resolution and the angular resolution is shown for jets matched ($\Delta R < 0.5$) with particle-level jets (which do not contain PU). As expected, PU affect the angular direction of the jets as well as the transverse momentum but in the second case the Fastjet algorithm corrects for most of the effect (residual effects are further corrected by the kinematic fit described in Section **??**). In Fig. 6 the number of not-matched jets and their transverse momentum distribution is shown. Additional jets coming from PU are only partially corrected by Fastjet but this effect is only of the order of 1%. In Fig. 7 the transverse momentum of the two leading jets and the dijet invariant mass of the candidate nearest to the nominal Z mass are shown with and without PU and PU corrections. Clearly the jet from signal are energetic enough to be not sizably affected by the PU but the invariant mass is affected because of the PU bias on the jet direction (as already shown in Fig. 5).

In the rest of the analysis the PU effect is corrected using the Fastjet algorithm.

In addition, we compute one variable β defined as the sum of transverse momentum of all charged particles in the jet coming from the primary vertex, normalized to the total sum of

transverse momentums of all charged particles in the jet. In Fig. 8 (left)the β distribution normalized to unity is shown for different background and signal Monte Carlo samples. Jets well reconstructed and with most of its constituents pointing to the primary vertex will have values arround the unity, whereas jets with important fraction of constituents pointing to other vertices, or not well reconstructed, will populate low values of β. The picture shows that an important fraction Drell-Yann background jets are mostly not coming from the primary vertex. In Fig. 8(right) the efficiency of a cut in $\beta < 0.2$ in the selected jets is shown, as a function of the number of reconstructed vertices, for Drell-Yann background and different signal samples. Whereas the signal is not sensible to PU, an important fraction of background is rejected with this cut, increasing linearly with the number of vertices.

These pile-up jets are not relevant for the Physics analysis and it is important to reject them from the selection. In the rest of the analysis a cut in $\beta < 0.2$ in the selected jets is performed.

The final systematics due to PU on the signal acceptance is discussed in Section **??**.

plots/xxx.png

Figure 3: Number of PU interactions in data and MC. In the following the MC sample has been reweighted to match the data.

XXX add a plot and text to demspnstrate that lepton rho corrections are working XXX

3.6 Kinematic Fit

Finite resolution of the jet energy is the dominant source of uncertainty in both di-jet invariant mass m_{jj} and di-boson invariant mass m_{ZZ} for Higgs candidates. Therefore, the two variables become highly correlated, as can be seen in Fig. 9 (left). One could take this into account by introducing a correlated 2D selection algorithm which includes this correlation into account and maximizes the signal-over-background separation. However, it is more attractive to correct the jet energies taking into account a constraint that the di-jet invariant mass should correspond to a Z. This is effectively exploiting an additional information in signal events, and is therefore expected to improve the resolution on the Higgs invariant mass; as for background, the assumption introduces a constraint which is not correlated to the underlying physical process, and therefore has the effect of shuffling randomly the events in the final ZZ invariant mass spectrum.

In order to optimally scale the di-jet quadrimomentum to the Z boson mass, we use a kinematic fit to the two jets. The fit is provided with parametrizations of jet transverse momentum and

position resolutions as functions of transverse momentum and pseudorapidity, and therefore constrains the mass of the di-jet system to the value of the Z boson mass by modifying the jet quadrimomenta in accordance to their expected resolutions. This brings a further improvement in the resolution on the signal invariant mass, as is shown by the black curves in Fig. 10.

XXX the resolutions for the kinematic fit probably need to be updated for 2012 XXX

The kinematic fit to the di-jet system also removes the correlation between the di-jet and di-boson invariant mass in signal, as can be seen in Fig. 9 (right). This allows a straightforward definition of signal and sideband regions, through simple rectangular cuts.

As a final remark, while the constrain in the analysis is done by imposing the exact value of the Z boson mass, we have investigated the possibility of introducing a width in the mass constraint. This has been done in the fit by substituting the Dirac δ-function with a gaussian distribution, centered on the nominal Z boson mass. We have studied gaussian widths of 2 and 5 GeV, but no sensitive difference in signal efficiency has been observed.

3.7 Higgs Candidate Reconstruction

From all the possible combinations of $Z \to 2l$ and $Z \to 2q$, resonance candidates are constructed. There is certain fraction of events when more than one candidate is present in an event after all the final selection requirements. The typical multiplicity of candidates per event is 1.2 (XXX recheck with high PU XXX) and is predominantly due to more than one combination of jets satisfying selection requirements. The number of jets with p_T above the 30 GeV/c threshold in the pre-selected events is shown in Fig. 11. Multiple combinations happened in both signal and background. In analysis we do not necessarily need to pick a certain candidate.

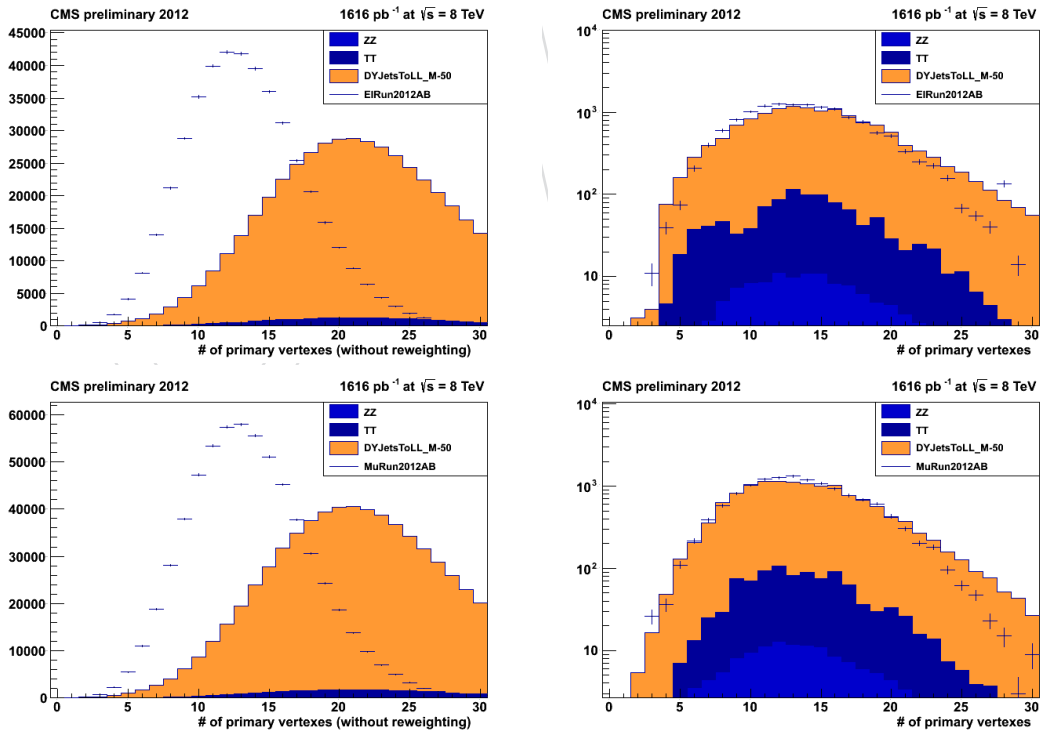


Figure 4: Number of reconstructed vertexes before (left) and after (right) the MC sample has been re-weighted. Pileup correction has been applied. Points with error bars show data after loose pre-selection, histograms show contribution of dominant background channels.

In the counting approach, we count the number of unique events after final selection, regardless of the number of combinations per each event. In the fit analysis, it is also possible to assign weight to each event. However, for simplicity of analysis we pick one unique candidate in each event passing all selection requirements. When the data are split in several b-tag categories (as discussed below), priority is given to candidates with the highest number of tagged jets. Among those candidates with the same number of tagged jets we pick the best candidate with the di-jet and the di-lepton invariant masses closest to the mean Z mass value [14].

The primary discrimination power of signal and background comes from the di-boson invariant mass. However, further sophisticated selection requirements are needed to discriminate signal from background, as shown in Fig. 11.

4 Performance of the B-tagging selection

Due to the relatively large branching fraction of the Z boson decaying into a pair of bottom-antibottom quarks, it is advantageous to exploit the possibility to identify jets originating from heavy-flavour quarks in order to enhance the sensitivity of the analysis.

For this purpose, the reconstructed jets in the event are checked for the possible heavy-flavour content and check whether it can be "tagged" as originating from a bottom quark. In a previous version of the analysis, this "b-tagging" was performed by using the so-called Track-Counting Discriminant [], but since that tagger is no longer recommended for physics analysis, in this generation of the anlaysis we have moved to the Jet-Probability (JP) Discriminant.

The JP tagger is based on track impact parameters: it entails computing the compatibility of a set of tracks to have come from the primary vertex. If the probability to come from the primary vertex is low, the jet is likely to be a b-jet. The jet probability estimation is performed through several steps: for each track, the probability to come from the primary vertex is computed and these probabilities are combined to provide the jet probability. The track probability distribution is calibrated by means of the distribution of track impact parameters with negative signs. The negative part of the impact parameter distribution is used for this purpose because it is mainly made up of primary vertex tracks. The advantage of this method with respect to track counting is the fact that a single discriminator is used and that information from all tracks is used at the same time.

The JP tagger has been validated in our selection using a control region that iscompletely dominated by background. This control region is defined by selecting 2 leptons and 2 jets (using the analysis object definition), requiring 70 GeV $< m_{ll} <$ 110 GeV, and one jet passing the tagger . Since we are mostly interested in the "medium" and "loose" version of the tagger, we have validated those version and confirm that the performance is similar to that of the Track-Counting tagger.

Tables 5 and 6 show the number of events selected with the "loose" working point in the control region using TCHE and JP taggers, and separated in flavor categories. The JP tagger provides similar prerformance tagging b-jets while reducing the amount of light jets in our major background, Z+jets production.

The addition of the tagging requirement is a fundamental part used to classify the final events

Figure 5: Resolution on transverse momentum (left), pseudorapidity (right) and azimuthal angle (bottom) in signal with Higgs mass 400 GeV with different PU conditions. Reconstructed jets are matched with particle-level jets ($\Delta R < 0.5$).

Muon Events - 4.6 fb^{-1}

	TCHEL			JPL		
	b-jets	c-jets	udsg	b-jets	c-jets	udsg
$t\bar{t}$	668.7	6.3	37.2	686.9	5.0	23.1
Z+jets	1647.8	1314.3	5482.42	1657.6	1317.6	4107.4
Total	2346.5	1359.3	5570.2	2375.2	1364.	4170.0
Higgs (400)	70.6	33.9	52.0	71.0	31.7	32.5

Table 5: Number muon events normalized to 4.6 fb^{-1} requiring one jet passing TCHE or JP "loose" working point. Events are break down in flavor categories and background samples, where total includes diboson events.

ELectron Events- 4.6 fb^{-1}

	TCHEL			JPL		
	b-jets	c-jets	udsg	b-jets	c-jets	udsg
$t\bar{t}$	598.1	8.2	39.6	607.2	6.6	22.1
Z+jets	1432.4	1131.2	4618.9	1455.7	1177.3	3408.4
Total	2056.9	1170.9	4701.5	2089.6	1217.1	3466.3
Higgs (400)	62.5	29.5	46.7	63.1	27.9	29.0

Table 6: Number electron events normalized to 4.6 fb^{-1} requiring one jet passing TCHE or JP "loose" working point. Events are break down in flavor categories and background samples, where total includes diboson events.

in order to use them in a more optimized way to enhance the sensitivity to signal. This is described in Section 6.1.

5 Optimization for VBF Higgs Production

A specific and well-motivated production channel for the Higgs boson is that mediated via vector bosons instead of the production involving quarks.

This production channel is especially interesting in models where the Higgs boson does not directly couple to the fermions, enhancing the branching fraction of decays into vector bosons, as in the analysis under investigation here.

If the Higgs is produced through Vector-Boson Fusion (VBF), the Higgs decay products are the same as in the standard direct (gluon-gluon) production, but in addition two additional partons appear in the final state, yielding two additional (spectator) jets in the final state. The presence of these jets is used to identify possible VBF production of the Higgs boson.

Therefore the analysis is performed in a similar way as in the standard selection: two leptons are used to identify a Z boson and two jets are used to reconstruct a second Z boson candidate. Additional two jets are required to "tag" the events as VBF candidate.

Thes two jets, being spectators of the Higgs production via the VBF are expected to present

Figure 6: Number of jets in signal with Higgs mass 400 GeV for different PU conditions: all jets (left) and jets not matched with particle-level jets ($\Delta R > 0.5$) (right). Transverse momentum of matched and not matched jets in the sample with PU corrections (bottom)

kienmatic properties that make them very different to the usual jets in events with a Z boson, produced with hard QCD radiation. They are expected to be very well separated and detected in the forward region of the detector.

Therefore a specific optimization has been performed in order to exploit these properties and enhance the sensitivity to the VBF events.

XXXX DETAILS OF THE OPTIMIZATION. XXXX

XXXX ALSO DETAILS ON HOW TO SELECT THE CANDIDATES XXXX

6 Final Selection

After the basic selection and the specific selections aiming to enhance the sensititivy for events with b jets and for VBF production, a final selection is performed to optimize the discrimination power for a Higgs signal.

6.1 Classification of events

Once the events are preselected as described in the previous sections, they are classified accordingly to the tagging-content of the jets associated to the Higgs candidate. Events are classified as "two tags" if the two jets are b-tagged with at least one medium-tag (see Section **??** for details) and one loose-tag. Events failing that requirement are classified as "one tag" if at least one of the two jets satisfy the loose-tag condition. Events failing this requirement are finally classified as "no-tag".

As these three categories present very different signal-to-background ratio and display some differences in the object description, they are treated separately in the final optimization described in this section. In addition, some differences in the cuts are also needed due to the different background composition, specially in the case of the "two tags" sample, in which the $t\bar{t}$ background starts begin noticiable.

On the other hand, since the optimization is based mostly in the intrinsic properties of the final state under investigation, there are also parts that are shared by the three categories. Also the methods used to estimate the backgrounds and obtain then final signal-background discriminant are common to all the categories.

6.2 Helicity Discriminant

There are several features in the signal $H \to ZZ \to 2l2j$ decay kinematics which discriminate it against background. We can exploit these kinematic differences to optimize selection and maximize signal significance or exclusion power. In the nominal approach we fully explore kinematics in the decay with five angles which characterize the decay.

It has been shown in Ref. [15, 16] that five angular observables fully describe kinematics in the decay $2 \to 1 \to 2 \to 4$ as in $ab \to X \to ZZ \to 2l2j$, and they are orthogonal observables to the three invariant masses of the X and the two Z. We should note that longitudinal and transverse momenta of the X are also additional orthogonal observables and could be used in analyses, but they typically have weaker discrimination power and rely on modeling of the PDFs and

Figure 7: Transverse momentum distribution of the leading jets (left) and of the second jet (right) and invariant mass of the dijet candidate nearest to the nominal Z mass (bottom) in signal with Higgs mass of 400 GeV with different PU conditions.

process dynamics. The above orthogonal observables are largely uncorrelated and are more attractive to be used in event selection rather than raw kinematic observables discussed in Section **??**.

In Fig. 12 we illustrate the angular distribution in the production and decay chain $ab \to X \to P_1P_2 \to p_{11}p_{12}p_{21}p_{22}$ with an example of the $ab \to X \to ZZ \to 4l$ or $2l2q$ (where quarks q hadronize to jets, which we refer to as $2l2j$ channel later) chain with two partons a and b, such as gg or $q\bar{q}$. The angular distribution can be expressed as a function of three helicity angles θ_1, θ_2, and Φ, and two production angles θ^* and Φ_1, as shown in Fig. 12. More details can be found in Refs. [15, 17], where parameterization of both signal and background distributions have been derived and implemented.

Here θ_i is the angle between the direction of the l^- or q from the $Z \to l^+l^-$ or $q\bar{q}$ (where the quark-antiquark pair produces two jets) and the direction opposite the X in the Z rest frame, and Φ is the angle between the decay planes of the two Z systems. The two Z's are distinguished by their decay type or, in case their daughters are the same type of particles, by an arbitrary convention. The production angle θ^* is defined as the angle between the parton collision axis z and the X decay axis in the X rest frame. The fifth angle can be defined as Φ_1, an angles between the production plane and the first Z decay plane.

A comparison of angular distribution in data and Monte Carlo can be found in Fig. **??**, where we find good agreement for background.

Previous work in Ref. [17] has concentrated on using angular information with a likelihood method for extracting signal and background yields from fits. Here we have chosen to adapt this information into a cut-and-count approach by, instead, building a likelihood discriminant from the angular distributions. While some statistical power is lost in reducing the MVA likelihood fit to a 1D discriminant, we gain in simplicity. For example, even if some parameterization of either signal or background effect is not perfect in the likelihood parameterization, the analysis is still not biased, it is only slightly less optimal than with the perfect description.

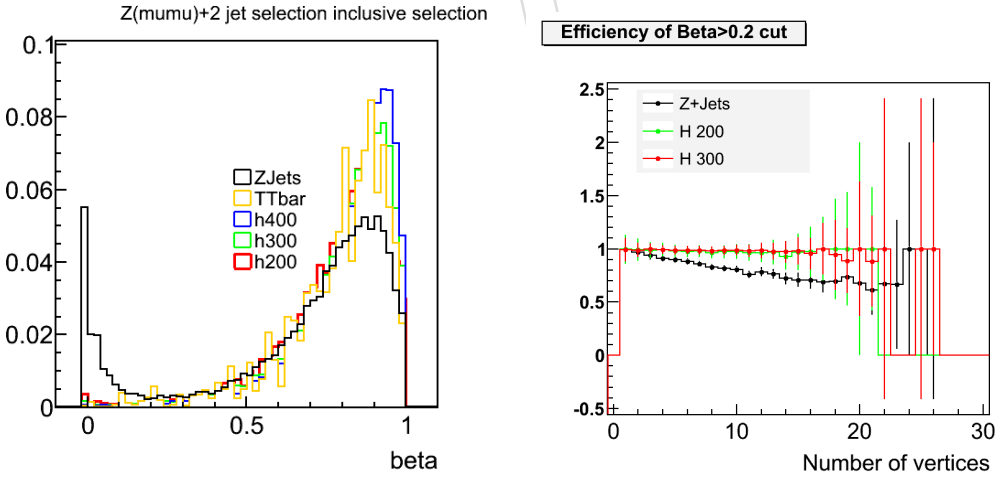


Figure 8: β fraction of the total transverse momentum of charged particles in a jet that is computed from particles pointing to the primary interaction for different background and signal samples (left) and efficiency of a cut in β <0.2 for signal with Higgs mass of 200 GeV and 300 GeV and Z+Jets background, as a function of the number of reconstructed vertices in an event (right)

Figure 9: Di-jet invariant mass vs. di-boson invariant mass for Higgs candidates (signal MC) after loose selection requirements. Left: before kinematic fit; right: after kinematic fit.

Figure 10: Comparison of signal MC with three different methods, raw invariant mass of two leptons and two jets, and with a kinematic fit and mass constraint. Left: $m_H = 250$ GeV; right: $m_H = 300$ GeV.

Assuming the probability distributions of the five helicity angles for both signal and background are known, P_{sig} and P_{bkg} respectively, the likelihood discriminant in given by the probability ratio

$$LD = \frac{P_{bkg}}{P_{sig}+P_{bkg}}.$$

This function has the feature that the signal is most likely to have values close to one and the background is most likely to have values close to zero. Events are then selected by requiring LD to be above certain threshold. This method of selection has been shown to provide at least similar results to those obtained using a set of optimized kinematic cuts. Furthermore, since the helicity angles are largely decoupled from the mass variables, by making selections based on the helicity likelihood discriminant one can better preserve the shape of the background's ZZ invariant mass distribution than with tight kinematic cuts. The method for obtaining such

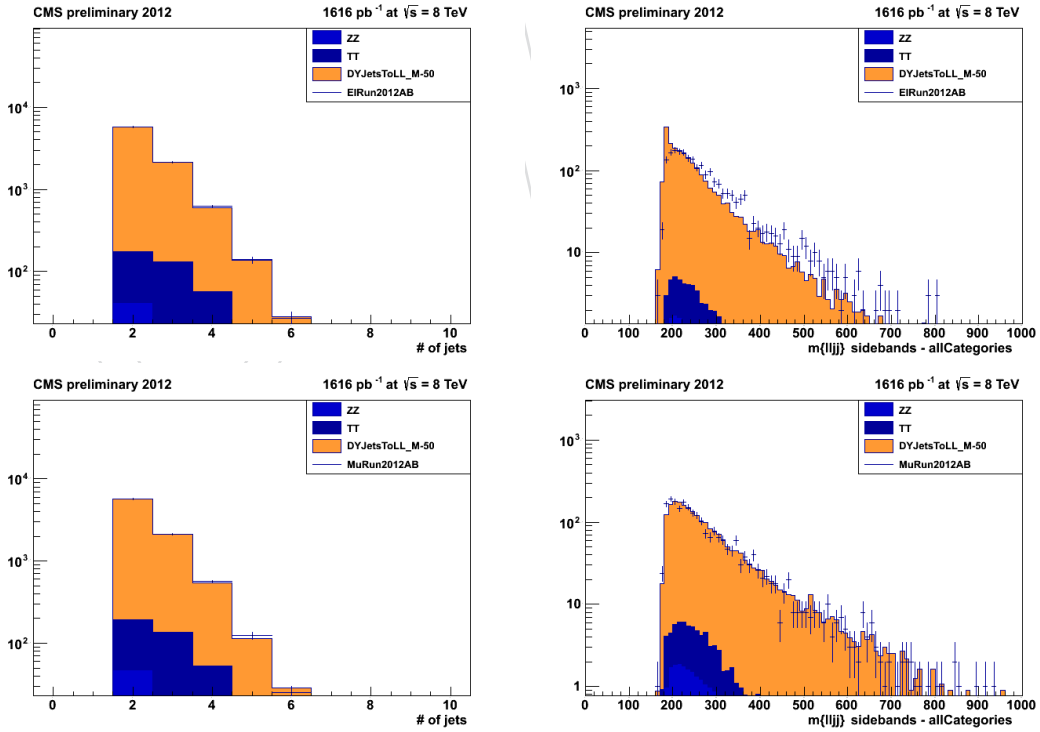


Figure 11: Left: number of jets pre-selected. Right: Di-boson invariant mass after loose selection requirements in the analysis. Comparison of 2012 data and MC simulation with dominant contributions is shown.

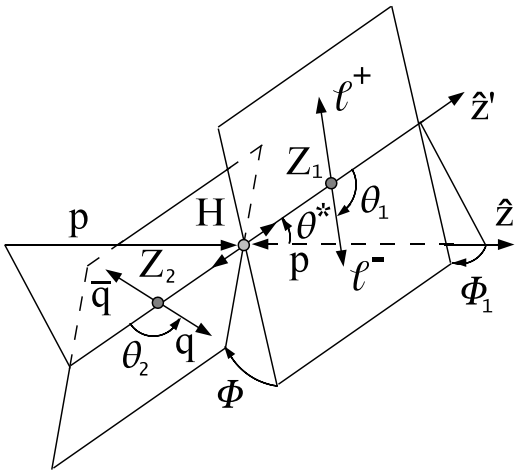


Figure 12: Diagram depicting the decay $X \to ZZ \to 2e2\mu$ and the 5 angles which describe such a decay.

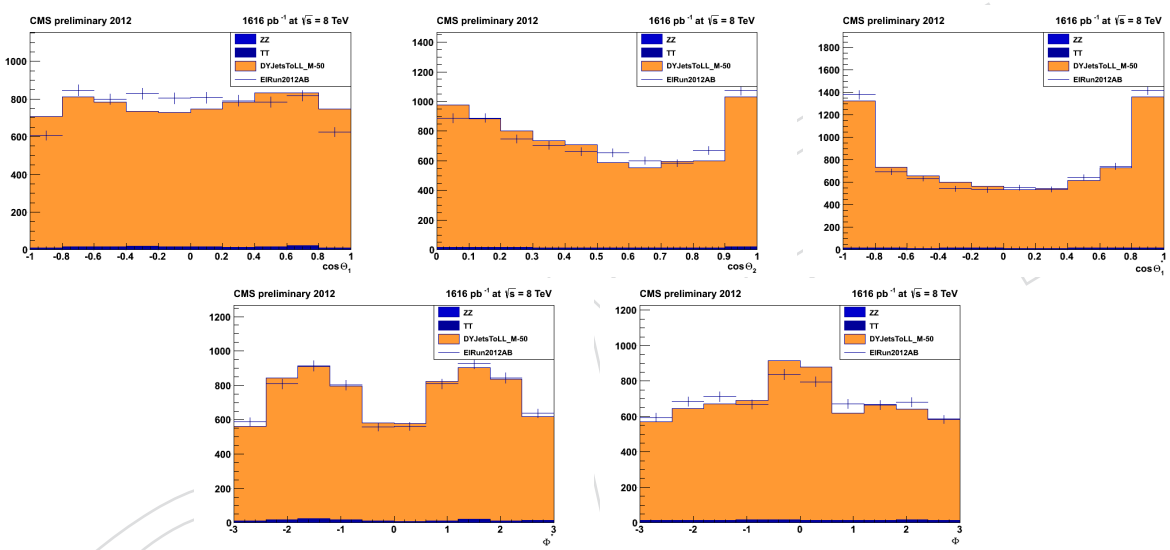


Figure 13: Five angular distributions of $\cos\theta_1, \cos\theta_2, \cos\theta^*, \Phi$, and Φ_1 for 2011 electron data (points) and Spring12 Monte Carlo samples (histogram).

a likelihood discriminant is described below.

The probability distribution function for signal was taken to be a product of the ideal, fully correlated, distribution which is derived in Ref. [15] and a set of four one-dimensional acceptance functions.

$$\mathcal{P}_{\text{sig}} = \mathcal{P}_{\text{ideal}}(\theta^*, \theta_1, \theta_2, \Phi, \Phi_1; M_{ZZ}) \cdot \mathcal{G}_{\theta^*}(\theta^*; M_{ZZ}) \cdot \mathcal{G}_{\theta_1}(\theta_1; M_{ZZ}) \cdot \mathcal{G}_{\theta_2}(\theta_2; M_{ZZ}) \cdot \mathcal{G}_{\Phi_1}(\Phi_1; M_{ZZ})$$

The four acceptance functions, $\mathcal{G}_{\theta^*}, \mathcal{G}_{\theta_2}, \mathcal{G}_{\theta_2}$, and $\mathcal{G}_{\Phi_1}$, have been obtained empirically from fits to Monte Carlo. Projections of $\mathcal{P}_{\text{sig}}$ can be seen in Fig. 15.

Where as the ideal function, $\mathcal{P}_{\text{ideal}}$, is naturally parameters with the ZZ invariant mass, the parameters of the four acceptance functions have all been re-parameterized in terms of m_{ZZ} only. This was done by fitting eight different Monte Carlo samples each corresponding to a different Higgs mass and then fitting the resulting parameters with either a linear or quadratic function of m_{ZZ}.

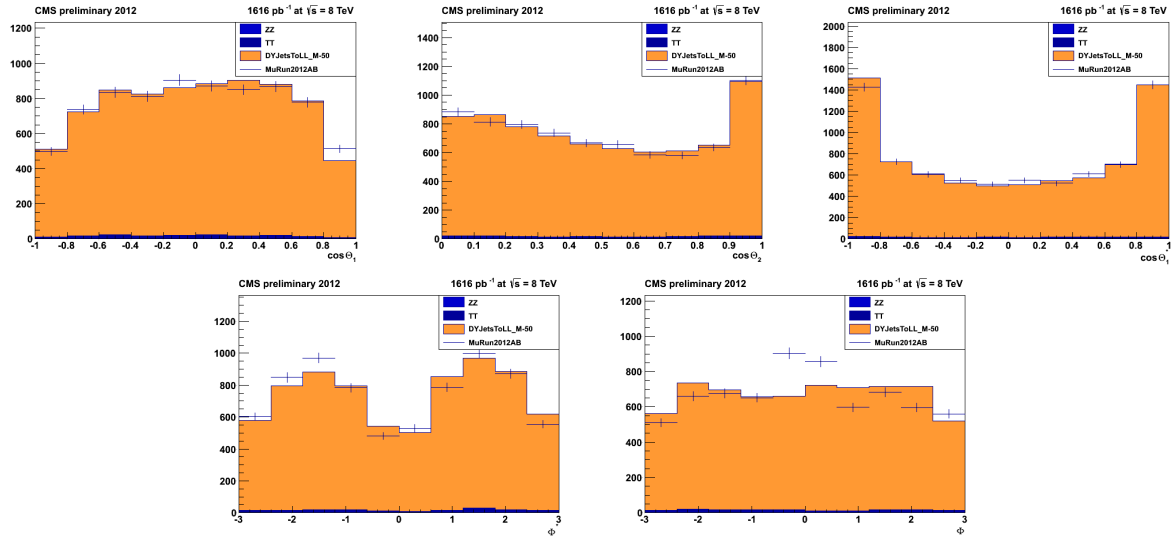


Figure 14: Five angular distributions of $\cos\theta_1, \cos\theta_2, \cos\theta^*, \Phi$, and Φ_1 for 2011 muon data (points) and Spring12 Monte Carlo samples (histogram).

The probability distribution function for the background was approximated with a product of five one-dimensional functions.

$$\mathcal{P}_{\text{bkg}}(\theta^*, \theta_1, \theta_2, \Phi, \Phi_1; M_{ZZ}) = \mathcal{P}_{\theta^*}(\theta^*; m_{ZZ}) \cdot \mathcal{P}_{\theta_1}(\theta_1; m_{ZZ}) \cdot \mathcal{P}_{\theta_2}(\theta_2; m_{ZZ}) \cdot \mathcal{P}_{\Phi}(\Phi; m_{ZZ}) \cdot \mathcal{P}_{\Phi_1}(\Phi_1; m_{ZZ})$$

All functions were obtained empirically from fits to Monte Carlo. Projections of $\mathcal{P}_{\text{bkg}}$ can be found in Fig. 16.

Similar to the case of $\mathcal{P}_{\text{sig}}$, the background Monte Carlo was divided into bins of m_{ZZ} and each bin was fit with $\mathcal{P}_{\text{bkg}}$. The parameters from each fit were then fit using either linear or quadratic functions of m_{ZZ}.

Combining $\mathcal{P}_{\text{sig}}$ and $\mathcal{P}_{\text{bkg}}$ into LD, we end up with the discriminant that is a function of the five helicity angles and parameterized by a given event's ZZ invariant mass. An example of the helicity likelihood discriminant is plotted in Fig. 17 for both background and signal Monte Carlo after loose kinematic selections .

There is a good agreement in the likelihood discriminant (LD) distribution between data and background MC shown in Fig. 18, as it is expected based on agreement of variables entering the LD calculation.

Figure 15: Distributions of $\cos\theta_1, \cos\theta_2, \cos\theta^*, \Phi$, and Φ_1 for a 500 GeV Higgs boson.

Figure 16: Distributions of $\cos\theta_1, \cos\theta_2, \cos\theta^*, \Phi$, and Φ_1 for background around 500 GeV.

Figure 17: Distribution of the helicity likelihood discriminant plotted for a 500 GeV Higgs signal (dashed lines) and background (solid lines) Monte Carlo.

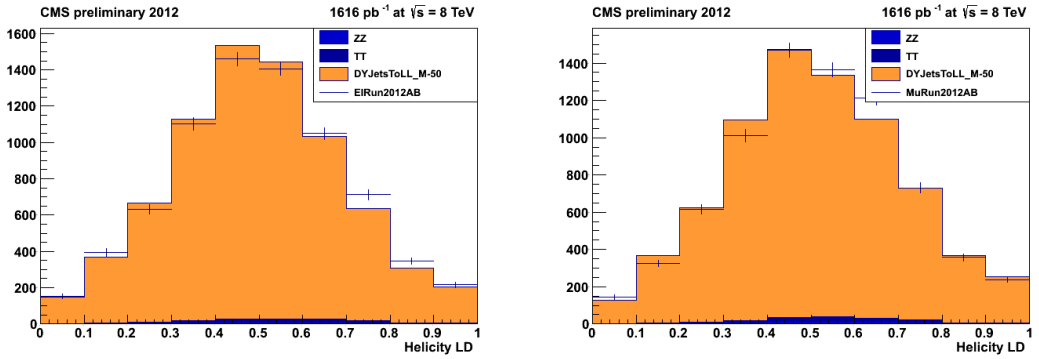


Figure 18: Comparison of helicity likelihood discriminant in 2012 Data (points) and Spring12 Monte Carlo (histogram); Left: electron channel, Right: Muon channel.

6.3 Quark-Gluon Discrimination

In the 0 b-tag category the dominant background originates from the Z+jets process with substantial contribution of gluon jets, while signal is dominated by hadronic Z decays to light quarks. Therefore identification of gluon jets helps to further reduce background. On the other hand the 1 b-tag category still has large fraction of Z decays to heavy flavor quarks in signal. Topology of the heavy flavor jets and gluon jets is similar, and therefore we do not attempt to achieve quark-glue discrimination in the 1 and 2 b-tag categories.

Gluons have a more intense coupling to the strong field than quarks, therefore their hadronization favors the production of a larger number of stable particles. This translates, in the detector, in the observation of wider, high-multiplicity jets, when compared to those generated by final state quarks. Furthermore, the phenomenon of 'gluon-splitting', if occurring at the beginning of hadronization, may give rise to jets made a number of collimated sub-jets.

These structural differences between gluon and quark hadronization may be exploited to derive a likelihood based discriminant. A more detailed description of this method may be found in [18].

The Particle Flow jet reconstruction method exploits all CMS subdetectors to their maximal granularity. It therefore offers the most detailed information on jet composition at particle level. The following variables are considered:

- **charged multiplicity**: number of charged hadron PFCandidates present in the jet;
- **neutral multiplicity**: number of photon and neutral hadron PFCandidates present in the jet;
- **transverse momentum distribution** among candidates ($P_{\mathrm{T}}D$), defined as:

$$P_{\mathrm{T}}D = \sqrt{\frac{\sum p_{\mathrm{T}}^2}{(\sum p_{\mathrm{T}})^2}}$$

  where the sums are extended over all PFCandidates in the jet.

We expect gluon jets to present higher values of charged and neutral multiplicities with respect to quark jets, and lower values of $P_{\mathrm{T}}D$, as is shown in Figs. 19 for a representative jet transverse

momentum bin.

These three variables can be combined into a likelihood, defined as a simple product of the single-variable distributions. The probability density functions are taken from MC QCD events, in which jets are matched to a parton in order to define their flavor. The expected likelihood distribution, for quark and gluon jets in the same jet transverse momentum bin, is shown in Fig. 20.

XXX get new plots for 2012 parameterization! from Fracnesco? XXX

Figure 19: Multiplicity of charged PFCandidates in quark (red) and gluon (blue) jets. All distributions are normalized to unity.

Figure 20: Multiplicity of charged PFCandidates in quark (red) and gluon (blue) jets. All distributions are normalized to unity.

The expected likelihood distributions, for the leading and subleading jets passing the aforementioned kinematic selections, are shown in Figs. 21, 22, and 23 for three Higgs masses ranges. Separation power improves at higher masses (with harder jets).

Data-MC comparisons of the discriminant, for leading and subleading jets passing preselection cuts, are shown in Fig. 24. There is a visible discrepancy in the distribution of the subleading jet. As will be shown in Section 8.3.5, we have no indication that the discriminant is performing differently from expected on quark jets. Further, even if we are limited in statistical power by the available data, there is no significant discrepancy yet for what concerns the leading jet. We therefore assume that the origin of the discrepancy lies in an incorrect description in the MC of the flavor of the subleading jet parton, i.e. it is more frequent to have it originating from a quark parton in the data than what modeled in the simulation.

The data-based validation of the quark-gluon discrimination likelihood is detailed in Section 8.3.5 and in [18].

6.4 MET

Without further selection, in the 2 b-tag category the dominant background originates from $t\bar{t}$ decay chain which contains two true b quark jets. We further reduce this background with a particle flow MET significance requirement of less than 10.

The distribution of PF MET significance for three Higgs mass hypotheses (300, 400, and 500 GeV) and the $t\bar{t}$ background are shown in Fig. **??**. We have also considered selection based on mass-dependent requirement on MET directly, as illustrated in the left plot of Fig. **??**. However, we find MET significance to be more physically motivated quantity.

Efficiency of the MET significance requirement of less than 10 varies between 99.0% at 250 GeV to 97.4% at 500 GeV for the signal Higgs candidates. This rather loose requirement is expected to be robust even in the presence of PileUp.

XXX Add low mass plot XXX

Figure 21: Quark-gluon likelihood distributions for leading jet (left) and subleading jet (center) passing selection, and for the two combined (left). Higgs at 300 GeV and Z+jets background in the range of 300 GeV are shown.

Figure 22: Quark-gluon likelihood distributions for leading jet (left) and subleading jet (center) passing selection, and for the two combined (left). Higgs at 400 GeV and Z+jets background in the range of 400 GeV are shown.

Figure 23: Quark-gluon likelihood distributions for leading jet (left) and subleading jet (center) passing selection, and for the two combined (left). Higgs at 500 GeV and Z+jets background in the range of 500 GeV are shown.

6.5 Optimization

Optimizing selection based on a large number of kinematic observables is a complicated task and we use the TMVA tool [19] to find an optimal point in a multidimensional space of observables. We employ several complementary approaches. In the default approach we use angular and mass variables and optimize requirement on the likelihood discriminant which is a function of the reconstructed mass m_{ZZ}. In Table **??** we show optimized selection requirements in the three b-tag categories. We remove glue-tag category from further consideration in analysis because it is fully dominated by background.

XXX reoptimize LD and btag categories? don't touch mass windows. XXX

XXX Final cut table XXX

6.6 Blinding policy/Unblinding strategy

In accordance with the Higgs Group proposal in 2012 about keeping the analyses "blind" to sensitive selections, we have implemented a blinding policy and set a well-defined unblinding strategy after the preapproval.

Basically the only distribution that could bias our approach to the analysis is the full invariant mass of the two Z, which is directly related to the Higgs mass. Therefore we have decided that the distribution itself is to be kept blind as long as possible.

By general motivation, we consider also sensitive to use the evens in the hadronic Z window (see Section**??**) for any study. However, due to the nature of our background estimation, the distribution of the full mass for those events is needed to actually validate the expectation so it would be required to be "unblinded" before the final calculations are done. It should be

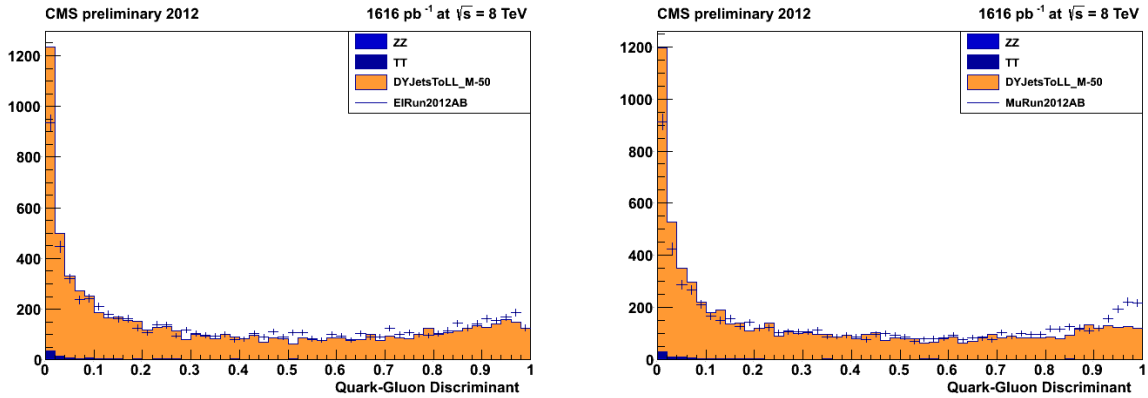


Figure 24: Quark-gluon likelihood distributions for leading jet (top-left) and sub-leading jet (top-right), and combined likelihood (bottom; left: electron channel, right: muon channel). Points with error bars show data after loose pre-selection, histograms show contribution of dominant background channels with MC.

noted that reproducing the global shape of that distribution is not really something that can bias our optimization that is not based on that distribution. We just use that distribution for defining the signal-region events for which the angular analysis is performed.

On the other hand, only the "no tag" category contains enough statistics and can be considered "signal insensitive" due to the overwhelming background. Therefore the validation of the sideband method will be performed in this region only. The others will remain blind until the end.

As described in Section 6.5, the final discriminant is not the full invariant mass, but some multivariate analysis of kinematics variables obtained from events in the specific signal regions. These signal regions are defined by a mass window for the hadronic Z and, for each assumption of the Higgs mass, a mass window for the full invariant mass.

In this respect, both, the distributions of the events AND the definition of the signal region for each Higgs mass under study will be considered sensitive to signal and therefore will be 'blind" in the analysis. The events in the deciced signal regions will not be looked at until the unblinding is approved (at preapproval). And of course the exact definition of the mass windows will be defined a priori as part of the blind optimization process. In fact they will be also be defined before the events in the signal region are used for anything, specifically for validating the background estimation from the sidebands.

The validation of the final variables and of the optimization process will be done using the events in the sidebands that enter in the signal region, when MC is not available for getting the background estimation. Therefore we do no need to use the data in the signal region (i.e. the selected window for the hadronic Z and the window in the full invariant mass for each assumption of the Higgs mass) for anything and will be blind regarding this purpose.

Those events will be kept blind until the analysis is preapproved. At that moment nothing will be changed in the analysis.

7 Background Determination from Data

After the final selection, discussed in previous sections, we have three m_{ZZ} distributions corresponding to the three b-tag categories 0 b-tag, 1 b-tag, and 2 b-tag. These distributions are the same for any of the Higgs mass hypothesis analysis. Therefore, they do not change while we change the signal prediction according to a hypothesis under investigation. These three distributions with the currently available data are shown in Fig. 26. We must note that the analysis has been performed blind, that is signal region in data has never been looked at while selection requirements were optimized. Once selection is "frozen," we investigate the m_{JJ} sidebands first, also shown in Fig. 26. Only when we observe an agreement in the sideband, we study the signal region.

7.1 Estimation of Z+jets sidebands

XXX TO BE REVISED... COPIED FROM PREVIOUS VERSION XXX

In order to minimize systematic uncertainties from MC predictions, we normalize background to sidebands. The procedure is applied independently in each b-tag category since background composition varies between categories. In all cases, the dominant backgrounds include Z+jets with either light or heavy flavor jets and top background. All these backgrounds equally populate the m_{JJ} signal region and m_{JJ} sidebands, as can be seen from Fig. 26. The di-boson background amounts to less than 5% in the 0 and 1 b-tag categories and about 10% in the 2 b-tag

category. Even very conservative uncertainty of 30% on MC prediction of the di-boson production would result in only 1-3% uncorrelated contribution to the total background prediction error. Therefore, we have developed a procedure which extrapolates the m_{JJ} sideband distribution of background, as shown in Fig. 26, to the signal m_{JJ} region. This extrapolation accounts for the 5-10% contribution of di-boson production not modeled in sideband using MC prediction. Uncertainties in this procedure result in uncertainties in both normalization and shape parameterization of the m_{ZZ} distributions. It is easy to show that the typical normalization uncertainties due to statistics in the sideband is about 5% in the 0 and 1 b-tag categories and 17% in the 2 b-tag category (expected with 1/fb of data). If we were to rely on MC prediction of background, uncertainties in the b-tag efficiency alone would result in about 20% uncertainty in the 2 b-tag category. Observe good agreement of m_{JJ} distributions in the three b-tag categories and in different m_{ZZ} mass ranges for the range $60 < m_{JJ} < 130$ GeV as shown in Fig. 27.

Figure 26: The m_{ZZ} distribution in the three b-tag categories from top to bottom: 0 b-tag (top), 1 b-tag (middle), and 2 b-tag (bottom). left: distribution in the signal m_{JJ} range; right: distribution in the sidband m_{JJ} range. Points with error bars show data after final selection, histograms show MC prediction with the dominant contributions shown separately.

Figure 27: Left column: The m_{jj} distribution in the three b-tag categories from top to bottom: 0 b-tag (top), 1 b-tag (middle), and 2 b-tag (bottom). Right column: The m_{jj} distribution in the three m_{ZZ} mass ranges categories from top to bottom: low (top), middle (middle), and high (bottom).

Figure 28: Histograms of $\alpha(m_{ZZ})$ for 0 btag (top left), 1 b-tag (top right), and 2 b-tag (bottom) categories using all relevant MC samples. Two options for Z+jets MC are used: ALPGEN and MadGraph.

We can use parameterization of the m_{ZZ} mass spectrum in the the full background MC cocktail in the m_{JJ} signal region and m_{JJ} sideband, as shown in Fig. 26. We create a ratio $\alpha(m_{ZZ})$ between the two, which predicts how sideband data can be scaled to obtain background prediction in the signal region. This is shown in Fig. 28. Since the full cocktail is used, this scale factor accounts for the di-boson production as well, with the small uncertainty discussed above. Then background expectation $N_{\rm bkg}(m_{ZZ})$ is obtained from distribution of events in the sideband $N_{\rm sb}(m_{ZZ})$ as follows:

$$N_{\rm bkg}(m_{ZZ}) = N_{\rm sb}(m_{ZZ}) \times \frac{N^{\rm MC}_{\rm bkg}(m_{ZZ})}{N^{\rm MC}_{\rm sb}(m_{ZZ})} = N_{\rm sb}(m_{ZZ}) \times \alpha(m_{ZZ}) \quad (1)$$

The advantage of the above approach is that most systematic uncertainties cancel in the ratios while the remaining factor $\alpha(m_{ZZ})$ reflects small kinematic differences between the signal region and sideband, which is mostly independent from the theoretical prediction of cross-sections. This procedure also provides automatic normalization of background and makes

Figure 29: Fits, blue lines, to extrapolated sideband distributions, black points, in 0 b-tag (top left), 1 b-tag (top right), and 2 b-tag (bottom) categories. Typical systematic variations of the shape are also shown (red and green curves).

any needed adjustments to the shape of the m_{ZZ} mass spectrum should there be any discrepancy. However, we would still like to note that there is very good agreement between data and MC prediction of the m_{ZZ} and m_{JJ} distributions with preselection selection requirement as discussed in previous sections.

To estimate the shape of background in data, the events in the sideband region, $60 < m_{jj} < 130$ GeV excluding the signal region, have been rescaled according to the corresponding $\alpha(m_{ZZ})$ shown in Fig. 28. These distributions are then fit using parameterization derived from MC. The result of such fits are shown in Fig. 29. The errors on the parameters are also shown with alternative in Fig. 29. Variations of the shapes come from uncertainties on the width of the core of the CB parameterization and position of the tail. These errors were obtained from a correlated fit of the two parameters, but then they are included in an uncorrelated manner. This is done for two reasons: (1) this approach is more conservative and may cover other potential uncertainties, and (2) the current version of Higgs combination tools does not allow correlations different from +-100% or zero. The plots show changes in the shape when both parameters are changed by $+1\sigma$ and -1σ at the same time.

The main systematic uncertainties in the above procedure are statistical in nature and scale with the expected size of the sideband data samples. Table 7 shows the expected yield from applying the above procedure and observed yield in each of the 6 channels. The errors on these expectations are on the order of $\sim 5\%$ for the 0 and 1 b-tag categories and $\sim 20\%$ for the 2 b-tag category. Errors on the shape parameters are taken from fits to the sideband.

7.2 $t\bar{t}$ Background Determination from Data

The $t\bar{t}$ background is estimated from the data using $e\mu$ events passing the same cuts as the signal. This method accounts for other small backgrounds (as WW + jets, $Z \to \tau\tau$ + jets, single top, fakes) where the lepton flavour symmetry can be invoked as well.

In this study we use the Powheg + Pythia $t\bar{t} \to 2l2\nu X$ Monte Carlo sample. Other top MC samples as the Madgraph $t\bar{t}$ inclusive sample or Powheg + Herwig $t\bar{t} \to 2l2\nu X$ produce consistent results.

Top-pair Monte Carlo studies show that the $e\mu$ vs. $ee + \mu\mu$ symmetry works very well at the level of the shapes of the distributions of all considered variables. Also, the relative event normalization is consistent with one, within the MC finite statistical errors. For instance, in the case of Powheg + Pythia top MC, the $e\mu/(ee + \mu\mu)$ relative event normalizations are 0.999 ± 0.002 after selection and kinematical cuts, and 1.006 ± 0.009 after b-tagging.

Figure 30 shows a comparison of the $ee + \mu\mu$ and $e\mu$ top MC distributions of several relevant variables, for events with at least two leptons and two jets passing selection cuts. Only the hardest-$\sum P_T$ dilepton combination, and the dijet combination with largest TCHE discriminator values are considered. The selection step is specified in each plot. The category "$\geq$ 1 b-tag" includes events with at least one jet tagged using TCHEM prescription. The category "2 b-tag" includes events with one jet tagged using TCHEM prescription and one jet tagged using TCHEL. The normalization is arbitrary.

Table 7: Observed and expected yields with 1.00 fb^{-1} of data after sequential preselection and all selection requirements. The yields are quoted in the range $183 < m_{ZZ} < 800$ GeV and are not intended for signal extraction without further analysis of the m_{ZZ} spectrum. The expected background is quoted from the sideband procedure (data) and from simulation (MC). The errors on the expected background from simulation include only statistical uncertainties. The expected Higgs signal yield is combined for the two lepton flavors.

		preselection		
p_T and η		31368		
m_{ll}		24641		
m_{jj}		5451		
		selection		
		0 b-tag	1 b-tag	2 b-tag
$\mu^-\mu^+jj$				
observed yield		359	396	25
exp. background (data)		345.7 ± 17.8	376.4 ± 19.3	24.3 ± 3.7
exp. background (MC)		351.5 ± 5.6	371.2 ± 5.9	23.7 ± 1.6
e^-e^+jj				
observed yield		307	352	30
exp. background (data)		286.4 ± 16.2	334.7 ± 18.2	20.3 ± 3.1
exp. background (MC)		304.3 ± 5.4	332.6 ± 5.8	23.6 ± 1.7
signal expectation (MC)				
Higgs	200 GeV	2.57 ± 0.37	3.43 ± 0.49	0.70 ± 0.19
	250 GeV	5.34 ± 0.75	4.75 ± 0.67	1.46 ± 0.38
	300 GeV	5.84 ± 0.83	4.97 ± 0.69	1.75 ± 0.46
	350 GeV	6.45 ± 0.94	5.61 ± 0.80	2.15 ± 0.56
	400 GeV	5.33 ± 0.76	4.83 ± 0.67	1.96 ± 0.50
	450 GeV	3.54 ± 0.53	3.38 ± 0.48	1.43 ± 0.35
	500 GeV	2.21 ± 0.34	2.21 ± 0.32	0.97 ± 0.24
	550 GeV	1.36 ± 0.22	1.41 ± 0.22	0.63 ± 0.16
	600 GeV	0.83 ± 0.21	0.86 ± 0.21	0.39 ± 0.15

The 2011 $e\mu$ data yields are compared to the sum of top MC prediction and other small backgrounds in Table 8, while distributions of relevant variables are superimposed in Figure 31. The cross section of the top MC has been increased by 4% to match the overall normalization of the data. Events selected contain at least two leptons and two jets passing selection cuts. Only the hardest-$\sum P_T$ dilepton combination and the dijet combination with largest TCHE discriminator values are considered. Pile-up corrections have been applied. Other extra cuts are detailed where appropriate.

The table and figure above include an estimation of WW, $Z \to \tau\tau$, and single top contributions from Monte Carlo. The fake component is estimated from $e\mu$ data; the yield of events with one or two non-isolated leptons (in the combined relative isolation region 0.25 - 0.85), is extrapolated into the isolated lepton region assuming a flat distribution in the combined relative isolation variable. Changing the size of the non-isolation region changes the fake prediction by at most 10%. The $e - \mu$ symmetry holds in reasonable approximation for the non-isolated lepton data.

The sample composition before b-tagging is 86% $t\bar{t}$, 6% fakes, and 8% other small backgrounds.

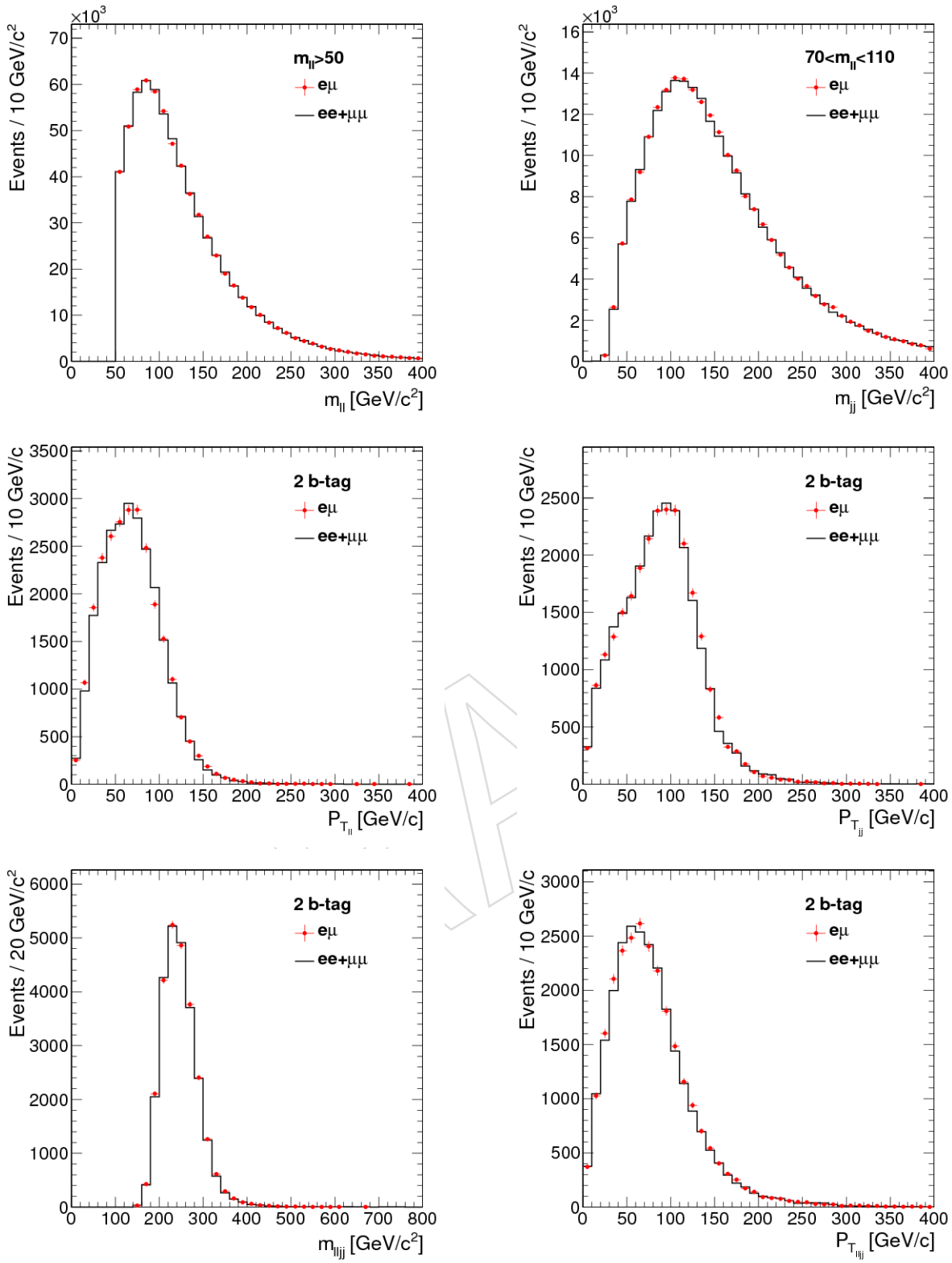


Figure 30: Powheg + Pythia top MC $e\mu$ to $(ee + \mu\mu)$ comparison for several variables after different step of the selection, as specified in the legends. Top: di-lepton invariant mass (Left) and di-jet invariant mass. Center: di-lepton (left) and dijet transverse momentum (right). Bottom: di-lepton + di-jet "Higgs" invariant mass (left) and transverse momentum (right).

Cuts	Total MC	$e\mu$ data
$M_{ll} > 50$ GeV/c^2	5947.4	5949
70 GeV/$c^2 < M_{ll} <$ 110 GeV/c^2	2147.6	2162
70 GeV/$c^2 < M_{jj} <$ 110 GeV/c^2	481.6	484
≥ 1 TCHEM b-tagged jet	346.9	369
≥ 1 TCHEL & ≥ 1 TCHEM	196.5	205
MET Significance < 10	64.7	61

Table 8: Comparison of 2011 $e\mu$ data to Powheg + Pythia top MC event yields, corresponding to an integrated luminosity of 4.9 fb^{-1}. "Total MC" contains the top, WW, $Z \to \tau\tau$, single top, and fakes contributions. Every cut in a line assumes all cuts in lines above.

After requiring 1 TCHEM tag (1 TCHEM and 1 TCHEL tags) the relative fractions change to 89%(94%) $t\bar{t}$, 6%(3%) fakes, and 5%(3%) other small backgrounds.

Now, we test the $e\mu$ vs. $ee + \mu\mu$ symmetry using a top-enriched subsample of the data. Figure 32 shows the MET significance distribution after requiring 1 TCHEM tag (left) and 1 TCHEM and 1 TCHEL tags (right). For values sufficiently large of MET significance the number of events of the $e\mu$ and $ee + \mu\mu$ samples are equal within statistical errors. In order to test the agreeement in shape, Figure 33 contains the di-jet invariant mass and "Higgs" invariant mass distributions after requiring 1 TCHEM tag, MET significance > 6, and $|M_{ll} - M_Z| > 30\,\text{GeV}/c^2$. One can observe agreement on both the normalization and shape of the $e\mu$ and $ee + \mu\mu$ distributions.

A first comparison of the 2012 $e\mu$ data to the 2011 data is performed in Figure 34. The agreement is good, within the still considerable statistical errors of the 2012 sample.

In summary, the $e\mu$ data sample has been understood. We plan to use it in the future for a data-driven estimation of the top background (plus other smaller components) in our analysis.

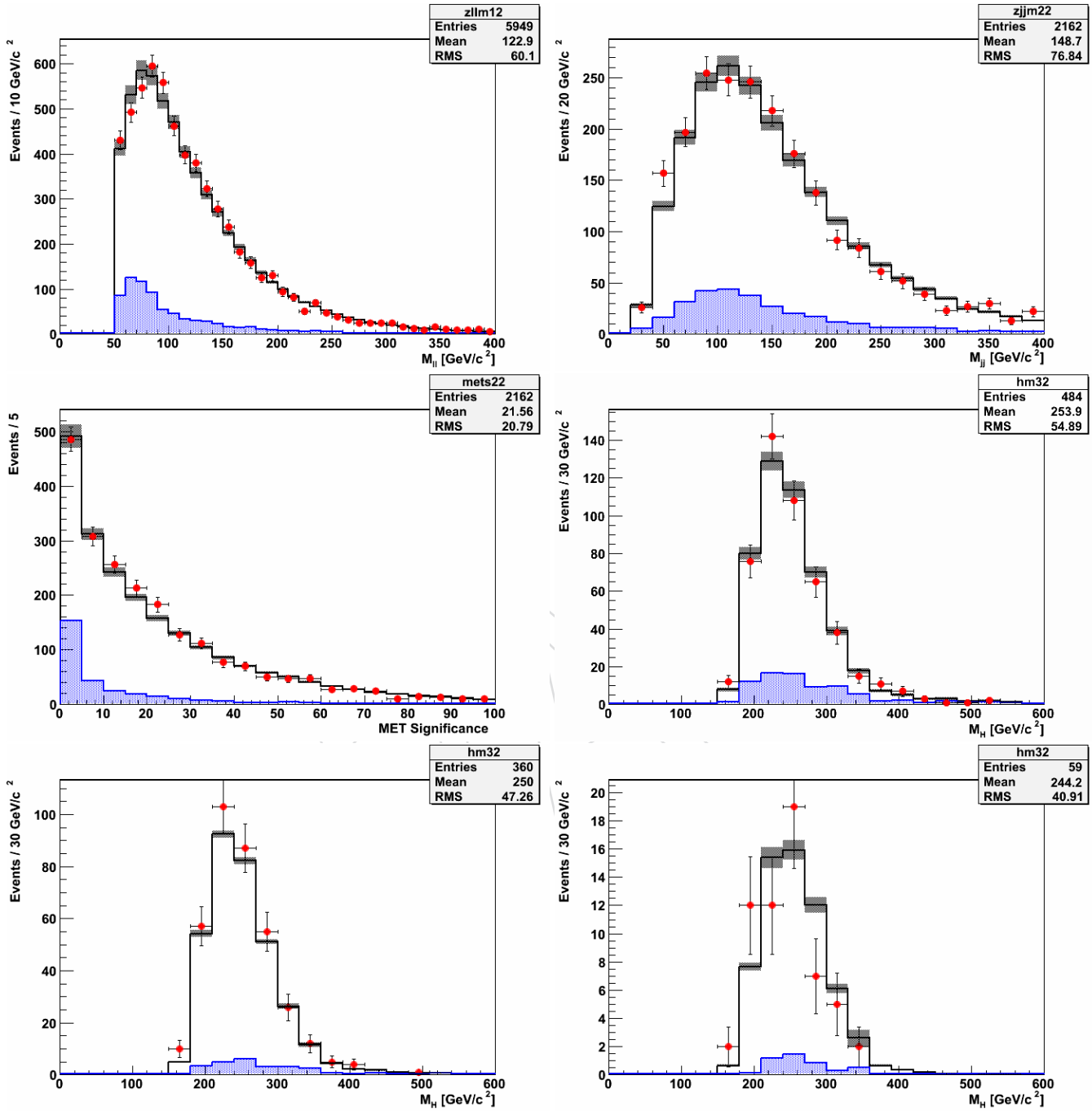


Figure 31: Comparison of 2011 $e\mu$ data to Powheg + Pythia top MC, corresponding to an integrated luminosity of 4.9 fb^{-1}. Red dots are $e\mu$ data; white histogram top Monte Carlo; blue histogram other small backgrounds. Top: dilepton invariant mass (left) and dijet invariant mass (right). Center: MET significante (left) and "Higgs" invariant mass (right). Bottom: "Higgs" invariant mass for events with 1 TCHEM b-tag (left), and 1 TCHEM +1 TCHEL b-tags and MET significance < 10.

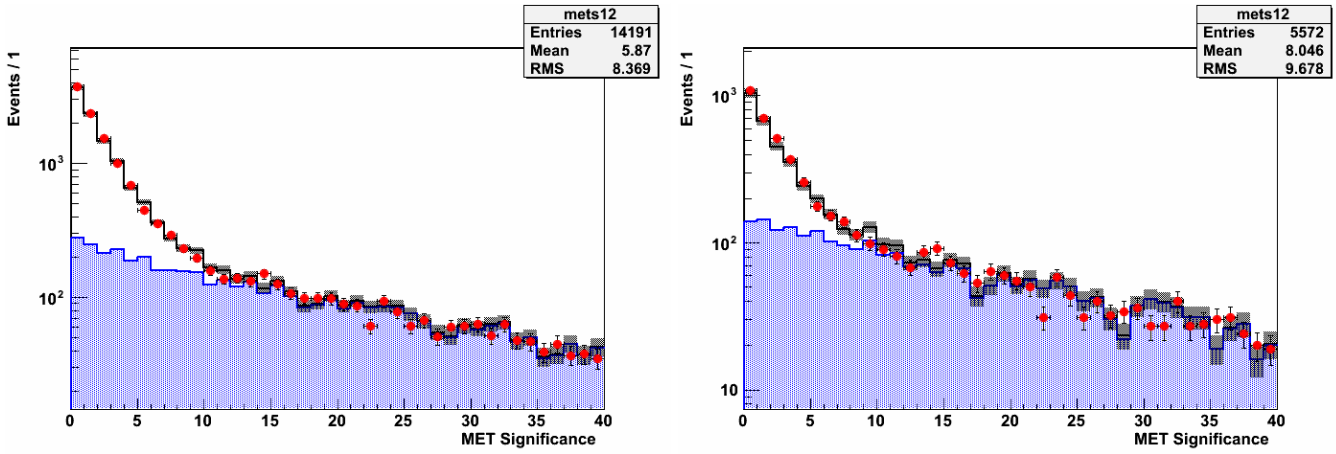


Figure 32: MET significance distribution for dilepton data compared to the sum of Drell-Yan Monte Carlo plus $e\mu$ data after 1 TCHEM b-tag (left) and two 1 TCHEM + 1 TCHEL b-tags (right). Red dots are $ee + \mu\mu$ data; white histogram Drell Yan Monte Carlo; blue histogram $e\mu$ data (plus other small backgrounds).

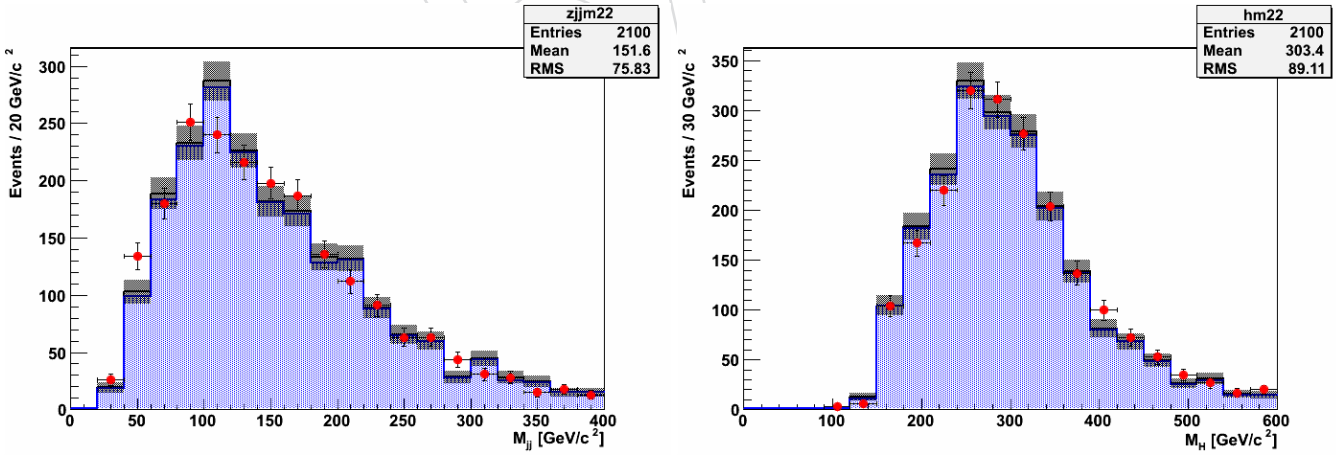


Figure 33: Dijet (left) and "Higgs" (right) invariant masses for $ee + \mu\mu$ and $e\mu$ data outside the leptonic Z mass window. Other cuts are detailed in the text. Red dots are $ee + \mu\mu$ data; white histogram Drell Yan Monte Carlo; blue histogram $e\mu$ data (plus other small backgrounds).

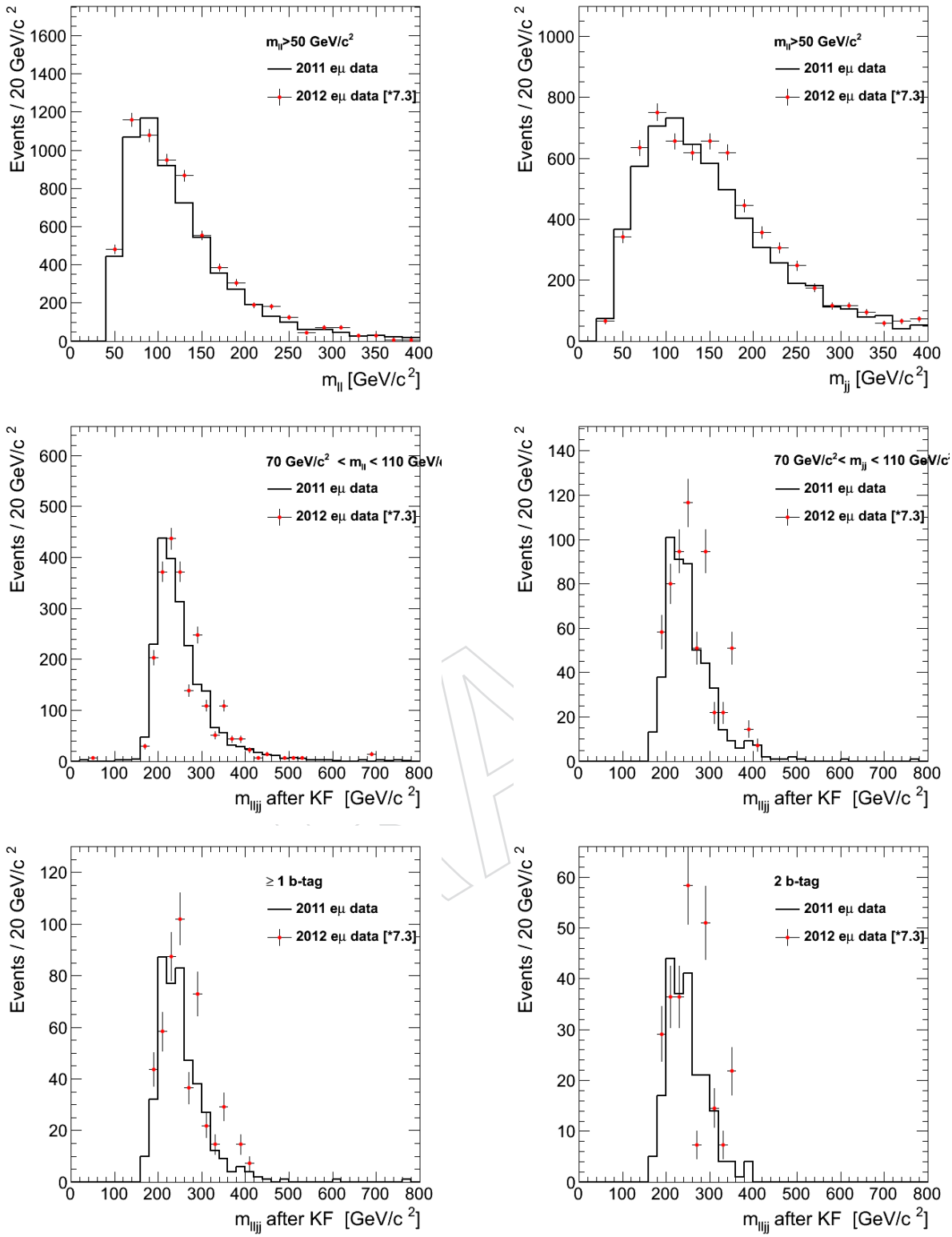


Figure 34: Comparison between 2011 $e\mu$ data and 0.624 fb^{-1} of 2012 $e\mu$ data projected to a luminosity of 4.9 fb^{-1}. Top: di-lepton and di-jet invariant mass after base selection. Center: "Higgs" invariant mass in the the leptonic Z mass window (left) and the hadronic Z mass window (right). Bottom: "Higgs" invariant mass after 1 TCHEM b-tag (left), and 1 TCHEM + 1 TCHEL b-tags (right).

7.3 Mistags

TO BE REVISED... COPY FROM AN-11-125

Another source of background consists of jets that, although initiated by u,d,s quarks or gluons[2], pass all selection cuts, including b-tagging requirement. The contribution from this background is estimated as the efficiency for tagging light jets, denoted as *mistag rate* [20]. In simulated events, it can be directly measured as:

$$\epsilon_{MC}^{\text{mistag}} = \frac{N^{\text{TAGGED light}}}{N^{\text{ALL light}}} \tag{2}$$

Jet flavour identification considers all partons generated within a cone ΔR <0.3 around the jet axis. If neither a b nor a c quark is found, the flavor *light* is assigned to the jet, if the highest energetic parton is either an u,d or s quark or a gluon.

However, mistag rates estimation must be cross-checked using data-driven techniques, so that it does not rely completely on the Monte Carlo modelling of jet-flavor content. At CMS, the negative taggers method [20] is used. It uses the fact that impact-parameter (IP) of b-jets is mainly positive, whereas for light jets is symmetric around 0 due to uncertainties in the track measurements. The tagger discriminant, given by the IP significance, can also be defined as positive or negative accordingly. Given its symmetry around zero, the number of tagged jets with negative tagger is a good estimate of the number of light jets tagged on the positive side. Further details are found in [20].

The working points for b-taggers at CMS are defined for different b-tagging efficiency and mistag rates. Correction factors $\epsilon_{DATA}^{\text{mistag}}/\epsilon_{MC}^{\text{mistag}}$ are available as functions of jet E_T, η in the database, and should be applied to simulated events before the comparison with the data.

Here, we use the same negative taggers method to investigate if the Higgs selection imposes any bias on the mistag rates, in which case the correction factors provided by the bTag POG group would not be appropriate. The data consists on 17 pb^{-1} of 2011 run period, and for Monte Carlo, the sample of Higgs signal for mass 400 GeV (see Table **??**) was used. Only muons were considered. Because of the small statistics – to be updated as 2011 data becomes available – neither the cut on the invariant mass of the muon pair nor the muon-jet separation was imposed. Only taggeable jets are considered, both in the numerator and the denominator of Eq. 2. All jets are required to have neutral EM energy fraction <1, neutral HAD energy fraction<1, charged HAD energy fraction >0, charged EM energy fraction<1; number of charged particle tracks >0 and number of charged+neutral particles >1. A charged particle track is defined as a track fulfilling the cuts: Number of pixel hits>1, number of silicon strip+pixel hits >7, $p_T > 1$ GeV, normalized χ^2 <5, transverse impact parameter < 0.2 cm and longitudinal impact parameter < 17 cm.

Results are presented in Tables 9 and 10. In spite of the large uncertainties, due to low statistics, the results are in very good agreement with the official bTag POG numbers, given in [20]. Therefore, the correction factors available in CMSSW database can be directly used in our analysis.

The contribution from charm-initiated jets, in addition to the mistags from light jets, is also shown.

In summary, the mistag estimation in our analysis region using the negative tag method is in

[2] In this note, we refer to jets initiated by either u,d,s quarks or gluons as *light jets*

Jet E_T (GeV)	**Mistag**	**Mistag+charm**
Data (TCHEM)		
>30	0.009 ± 0.005	0.019 ± 0.010
30-45	0.012 ± 0.018	0.028 ± 0.038
45-80	0.014 ± 0.016	0.032 ± 0.032
>80	0.012 ± 0.019	0.024 ± 0.035
Monte Carlo (TCHEM)		
>30	0.012 ± 0.004	0.026 ± 0.006
30-45	0.007 ± 0.006	0.015 ± 0.008
45-80	0.009 ± 0.006	0.021 ± 0.008
>80	0.015 ± 0.015	0.031 ± 0.020

Table 9: Mistag rates and rates of light+c flavor jets using TCHEM tagging, in bins of jet E_T.

Jet E_T (GeV)	**Mistag**	**Mistag+charm**
Monte Carlo (TCHEL)		
>30	0.086 ± 0.011	0.110 ± 0.012
30-45	0.059 ± 0.018	0.075 ± 0.019
45-80	0.070 ± 0.016	0.091 ± 0.017
>80	0.102 ± 0.039	0.127 ± 0.041
Data (TCHEL)		
>30	0.051 ± 0.012	0.065 ± 0.014
30-45	0.060 ± 0.031	0.076 ± 0.038
45-80	0.027 ± 0.012	0.035 ± 0.015
>80	0.133 ± 0.080	0.166 ± 0.094

Table 10: Mistag rates and rates of light+c flavor jets using TCHEL tagging, in bins of jet E_T.

good agreement with the estimations provided by the BTV POG. During the analysis official numbers from the BTV POG shown in Appendix **??** are used.

7.4 Diboson background

8 Systematics

In this section the systematic uncertainties affecting the analysis, the method used to estimate them and their estimated values are described.

8.1 Luminosity uncertainty

The latest recommendation for the 2012 datasamples is the uncertainty on LHC luminosity of 6%.

8.2 Higgs cross-section and branching fractions

The Higgs production cross-section uncertainty depends on production mechanism, either gluon fusion or weak boson fusion (WBF). However, since the gluon fusion mechanism dominates, it drives the total uncertainty. We use gg and WBF errors separately and for each mass point according to Yellow Report prescription. The total weighted error is in the range 13.4–18.0%. We note that this uncertainty is relevant only for the measurement of the ratio to SM

expectation R, while it does not affect the absolute cross-section measurement.

8.3 High-Mass selection

Uncertainty on background is considered separately and is one of the dominant effects on the exclusion limits. Systematic uncertainties on background are discussed in Section **??**. Uncertainty on the signal shape parameterization arise from two sources: theory uncertainty in the BW width (Γ) parameterization, such as its mass-dependence, and experimental uncertainty in the CB resolution function. The former is taken from comparison of the mass dependent and mass-independent width parameterization used in the POWHEG and JHU generators. The latter is estimated from various sources discussed below, such as jet energy uncertainty, resolution, etc. In general signal shape uncertainties are not the dominant sources of systematics with expected signal yield of several events.

The main systematic uncertainties on signal normalization are summarized in Table 11, and are discussed in more detail below. Lepton efficiencies are evaluated with a tag-and-probe approach when one lepton from an inclusive sample of Z decays serves as a tag and efficiency for the other lepton is calculated. Effects of jet reconstruction are evaluated with variation of the jet energy and resolution within calibration uncertainties. Effects of pile-up are taken as a difference between reconstruction efficiency with pileup below and above the average expected value, otherwise distributed according to observed values in data. Requirement on the MET significance translates into about 3% inefficiency and the resulting uncertainty does not surpass this value. Uncertainty on the b-tagging has been evaluated with inclusive sample of b-jets. Uncertainty on quark-glue LD selection efficiency was evaluated with predominantly quark jets in the γ+jets sample. Uncertainties in the production mechanism affect both longitudinal momentum of the Higgs, due to PDFs, and transverse momentum of the Higgs, due to QCD initial-state radiation effects. We follow the PDF4LHC recommendation to estimate uncertainty due to PDF knowledge and calculate uncertainty on signal acceptance. We rescale the transverse momentum distribution of the Higgs using HQT as a reference and take the full change in efficiency as systematic uncertainty. Uncertainties on the Higgs cross-section are taken from the Yellow Report which includes uncertainties from QCD renormalization and factorization scales, PDFs, and α_s. These uncertainties are separated between the gluons fusion and VBF production mechanisms, but gluon fusion uncertainties dominate in the total production cross-section.

8.3.1 Lepton energy scale, resolution, selection, and trigger

Lepton trigger and selection is common among several $H \to ZZ$ analyses and we benefit from common study based on tag-and-probe techniques. In particular, recent studies within the framework of Ref. [11] indicate systematics of 1.0% due trigger, 0.5(3.3)% due to muon (electron) identification, 0.2(0.8)% due to muon (electron) isolation mostly independent of the mass hypothesis, 1.0(2.0)% due to muon (electron) momentum/energy scale.

8.3.2 Jet Energy Scale and Resolution

The main uncertainty in jet reconstruction comes from jet energy scale (JES) uncertainty, while the uncertainty on the resolution contributes a much negligible effect to the total uncertainty. Preliminary estimate with 2010 data show that jet energy uncertainty could be kept within about 4% and resolution within 10%. For more details see Refs. [21, 22]. Since the background is extracted from sidebands in data, the systematics due to JES and jet resolution uncertainty affects only signal efficiency and potentially m_{ZZ} distribution.

Our preliminary estimates show that JES variation by $\pm 1\sigma$ changes reconstruction efficiency of a 400 GeV Higgs by about 5%. In Fig. 35 the effect of a possible JES bias ($\pm 1\sigma$) is shown on some fundamental variables. The effect on the jets transverse momentum and dijet invariant mass is sizable and it drives the bias on the acceptance while the effect is very small on the Higgs candidate invariant mass, thanks to the m_{JJ} kinematical fit. The small effect on the angular LD is due to the bias in the boosts applied to compute the angles in the Higgs and Z reference frames. The bias on the QG discriminant and b-tag categorization is negligible, as expected. Detailed study as a function of Higgs mass hypothesis is provided in Table 12.

Figure 35: Distribution of leading (top, left) and subleading (top, center) jet transverse momentum, dijet invariant mass (top, right), angular LD (bottom, left) and Higgs candidate mass (bottom, right) after final selection requirements with nominal Jet Energy Scale (JES) and with JES changed by $\pm 1\sigma$.

A bias on the JES can also affect the signal shape. The Higgs mass shape is defined in Sec. **??**. In Tab. 13 the values of the Higgs shape parameters for different JES conditions are listed for various Higgs masses. In Fig. 36 the Higgs mass distribution and its fit are shown for different JES conditions.

Figure 36: Reconstructed Higgs mass distributuions and its fit in different JES conditions for 200 GeV (left) and 600 GeV (right) Higgs mass hypothesis, integrated over all b-tag categories.

The effect of the jet resolution uncertainty on the signal was evaluated by applying an additional smearing to the jets and comparing to the same sample without additional smearing. As the background is evaluated directly from the data it is not expected that the jet energy resolution will have a significant effect. The additional smearing applied to the jets is the nominal jet resolution at the given p_T and η multiplied by the difference observed between resolution in data and simulation as given in [22] (differential in η). The applied factor is typically quite small, in the low percent to sub-percent range, due to the good data-MC agreement. The effect on the selection efficiency is very small ($< 1\%$) due to several effects: For all but the lowest

Table 11: Summary of systematic uncertainties on signal normalization. Most sources are multiplicative errors on the cross-section measurement, except for expected Higgs cross-section (which is relevant for the measurement of the ratio to SM expectation R). See text for more details.

source	0 b-tag	1 b-tag	2 b-tag	comment
muons reco	2.7%			tag-and-probe study
electrons reco	4.5%			tag-and-probe study
jet reco	1%–8%			JES uncert., JER uncert. negligible; correlated between c
pileup	3-4%			correlated between categ
b-tagging	2-7%	3-5%	10-11%	anti-correlated between categ.
glue-tagging	4.6%	–	–	loose requirement ⇒ expected small, studies on-goir
MET	–	–	3-4%	loose requirement
production mechanism (PDF)	3%			PDF4LHC, acceptance only
production mechanism (HQT)	2%	5%	3%	only for $M_H = 200$ GeV, $< 1\%$ for $M_H = 200, 400$
production mechanism (VBF)	1%			
luminosity	6%			same for all analyses
Higgs cross-section (for R)	13–18%			detailed table from YR available

Table 12: Cut acceptance for different jet energy scales. The uncertainty is due to the available MC statistics. Only the muon channel is considered and the acceptance is defined on the full m_H spectrum.

Sample	cuts	central value	JES $+1\sigma$	JES -1σ	systematics: $+1\sigma, -1\sigma$
signal (mh200)	preselection cuts	0.134 ± 0.002	0.141 ± 0.002	0.126 ± 0.002	
	LD and QG cut	0.058 ± 0.001	0.061 ± 0.001	0.054 ± 0.001	+5%, -7%
signal (mh300)	preselection cuts	0.282 ± 0.002	0.286 ± 0.002	0.275 ± 0.002	
	LD and QG cut	0.195 ± 0.002	0.198 ± 0.002	0.192 ± 0.002	+1%, -2%
signal (mh400)	preselection cuts	0.381 ± 0.002	0.380 ± 0.002	0.378 ± 0.002	
	LD and QG cut	0.266 ± 0.002	0.264 ± 0.002	0.266 ± 0.002	-1%, +0%
signal (mh500)	preselection cuts	0.417 ± 0.003	0.410 ± 0.003	0.417 ± 0.003	
	LD and QG cut	0.276 ± 0.002	0.269 ± 0.002	0.280 ± 0.002	-2%, +1%
signal (mh600)	preselection cuts	0.416 ± 0.003	0.408 ± 0.003	0.419 ± 0.003	
	LD and QG cut	0.255 ± 0.002	0.248 ± 0.002	0.260 ± 0.002	-3%, +2%

mass working points, jet energies are high enough above the threshold of 30 GeV that the additional smearing has little effect on the selection. Distribution of m_{JJ} falls very softly, so that the number of events migrating into the acceptance roughly compensates the number of events migrating into it. The LD variable is indirectly affected by the smearing via boosts into the H and Z center of mass systems. However, this is counteracted by the kinematic fit. Similarly, the reconstructed Higgs line shape does not significantly broaden as shown in Fig. 37.

Figure 37: Reconstructed Higgs mass distributuions with (red) and without (black) additional jet-p_t smearing for several masses, integrated over all b-tag categories.

8.3.3 Pile-up

The PU is corrected as reported in Section **??**. Since the background is extracted from sidebands in data, the main systematics due to PU residual effects after corrections can be the effect on the signal efficiency and m_{ZZ} distribution.

The MC samples are re-weighted to match the PU distribution measured in data and the main source of systematics may come from a mis-modeling of the PU in MC with respect to data (Pythia Z2 tune is used for the PU simulation) or from the uncertainty on the measurement of the amount of PU in data.

As systematics we quote in Table 14 the difference between the average efficiency and the efficiencies we get with more or less PU events than the mode of the data (7 PU events). This procedure is quite arbitrary but surely conservative.

To find a more motivated estimate of the systematics, the two sources of uncertainty can be separated:

- The expected number of pileup events in a given collision is affected linearly by the luminosity uncertainty.
- The events overlayd in the simulation originate from a minimum bias simualtion and may not perfectly mirror the topology of overlaying events in the data.

The first point was studied by shifting the N_{PU} distribution extracted from data by the luminosity uncertainty. This leads to a change of $\sim 0.5\%$ in signal acceptance, approximately inde-

Table 13: Value of the parameters for the signal higgs mass fit for different JES conditions. The last column is an estimation of the statistical error of the fit.

Sample	parameter	central value	JES $+1\sigma$	JES -1σ	fit error
signal (mh200)	CB mean	0.7	0.5	0.6	0.2
	CB sigma	3.0	3.0	3.0	0.4
	CB α_1	0.9	1.0	0.9	0.2
	CB n_1	2.7	1.8	3.0	0.5
	CB α_2	1.1	1.0	1.1	0.1
	CB n_2	1.2	1.3	1.2	0.1
signal (mh300)	CB mean	4.2	4.5	4.0	0.3
	CB sigma	6.6	7.0	7.2	0.7
	CB α_1	1.3	1.4	1.5	0.2
	CB n_1	1.4	1.4	1.3	0.2
	CB α_2	1.0	1.0	1.2	0.1
	CB n_2	2.4	2.5	2.0	0.3
signal (mh400)	CB mean	15.1	15.9	14.5	0.5
	CB sigma	11	12	11	1
	CB α_1	1.8	1.8	1.8	0.2
	CB n_1	0.9	0.8	0.9	0.2
	CB α_2	1.4	1.3	1.4	0.2
	CB n_2	2.2	2.3	2.1	0.4
signal (mh500)	CB mean	17	18	16	1
	CB sigma	15	15	14	3
	CB α_1	2.3	2.4	2.2	0.4
	CB n_1	0.3	0.3	0.5	0.3
	CB α_2	1.8	1.6	1.8	0.4
	CB n_2	1.2	1.4	1.2	0.5
signal (mh600)	CB mean	20	22	20	1
	CB sigma	25	23	26	3
	CB α_1	2.3	2.3	2.2	0.2
	CB n_1	0.1	0.1	0.2	0.2
	CB α_2	5	5	5	6
	CB n_2	10	9	8	8

pendent of higgs mass and b-tag category. The second point is much harder to address as the quality of the minimum bias simulation used to emulate pileup cannot be easily quantified. We use the distribution of ρ as defined in the Fastjet algorithm (an estimation of the energy spread around in the detector due to PU and underlying event) as a general indicator of the quality of the simulation. We compute the change in efficiency when reweigting this distribution in the signal simulation to the spectrum observed in data. Note, that the signal is dominated by quark jets, while the background dominated data contains mostly gluon jets, which are more likely to produce soft unclustered hadrons. Thus the reweighting may not be completely accurate. The observed effect of this reweighting is $< 1\%$. Until these studies on the PU event topology have been further consolidated, we use the conservative estimate discussed above.

To show which distributions are affected by the PU and how, we compare no-PU and PU corrected samples in signal MC. The distributions of the angular likelihood discriminant and of the QG discriminant are shown in Fig. 38 in different PU conditions for all the Higgs candidates,

passing the preselection cuts listed in Sec. 3.7, in a signal MC sample with 400 GeV Higgs mass. The distributions of the Higgs mass and the b-tag categorization are shown in Fig. 39 for the same sample after all the cuts have been applied and the best Higgs candidate has been chosen. The PU effect on the LD distribution and the Higgs mass is negligible. Moreover the PU may impact the b-tag performance moving events from one b-tag category to another.

Figure 38: Angular (left) and QG (right) likelihood discriminant in signal with Higgs mass of 400 GeV for different PU conditions. *The QG plot has not been updated yet to the latest robust implementation of the QG LD, but the study includes the latest implementation*

Figure 39: Categories of b-tags and reconstructed Higgs candidate mass in signal with Higgs mass of 400 GeV for different PU conditions.

While the above method properly estimates the residual effects of pileup after corrections, it is overly conservative as the residual effects of the pileup are at least partly accounted for in the simulation.

8.3.4 Heavy quark flavor tagging uncertainty

For more specific b-tag studies we refer to Ref. [11]. Preliminary estimates show uncertainties $\sim 20\%$ in 2 b-tag category and $\sim 5\%$ in 0 and 1 b-tag categories.

8.3.5 Quark-glue tagging

The quark-gluon discriminant is founded on general assumptions on the structure and couplings of the QCD Lagrangian, yet does rely on the modeling of hadronization done in the generator. Mismodelings of (light) quark hadronization, which affect the chosen observables (jet multiplicities and transverse momentum distributions) would alter the predicted signal efficiency of the selection. It is therefore important to verify that the performance of the discriminant on quark jets is similar to expectations.

A control sample is identified by photon+jet events, in which the leading jet originates from light quarks in more than 90% of the cases. In order to contrast the dominant background, constituted by QCD dijet events in which one of the two jets has fragmented mainly into a particle capable of creating a large energy deposit in ECAL (such as a neutral pion), a stringent photon identification is needed. We make use of the photon identification described in Ref. [**?**]. In order to ensure the absence of jets originated from b-quarks, events in which the leading jet has a positive loose TCHE b-tag have been vetoed. The expected photon+jet purity of this selection is of the order of 90% at high transverse momenta, and significantly lower at low transverse momenta (reaching $\sim 50\%$ at 20 GeV). It must be noted that a background infiltration does dilute the quark component, but not dramatically, as about 40% of jets in QCD events are originated from quark partons.

Table 14: Summary of efficiency differences between low/high PU subsamples and the average.

M_H	$M_H = 200$ GeV			$M_H = 400$ GeV			$M_H = 600$ GeV		
	0 b-tag	1 b-tag	2 b-tag	0 b-tag	1 b-tag	2 b-tag	0 b-tag	1 b-tag	2 b-tag
Δ_{eff}	6.5%	1.7%	3.2%	1.6%	$< 1\%$	$< 1\%$	3.7%	3%	2.2%
total	3.7%			1%			3.1%		

Figure 40: Distributions of the quark-gluon discriminant in photon+jet events in three transverse momentum ranges. The Monte Carlo distributions are normalized to the shape of the data.

Figure 41: Expected distributions of the quark-gluon discriminant for quark jets in three transverse momentum ranges. The gluon contribution has been subtracted by accessing the MC truth. The Monte Carlo distributions are normalized to the shape of the data.

The shape of the quark-glue discriminant obtained on photon+jet events, in three different transverse momentum bins, is shown in Figures 40. The data is compared to the simulation, and the latter is normalized to the signal shape. The available amount of data decreases at lower transverse momenta because of the presence of high prescales in the photon triggers. The observed shape of the discriminant seems compatible with expectations, within the statistical precision granted by the analyzed data.

We are interested in studying a possible effect on quark efficiency, so the gluon contribution has to be subtracted. The latter is isolated in the simulation by applying a matching at MC truth level between jets and partons. Jets successfully matched to gluons are hence subtracted both from data and MC. The gluon-subtracted distributions are shown in Figures 41.

In order to evaluate the effect on signal efficiency, we have to simulate the effect of cutting on the product of two jet's likelihood, with similar kinematic properties as those expected in the case of a heavy Higgs decay. It is not possible to isolate a sample of photon+jet events with

Table 15: Efficiencies of requiring the quark-gluon discriminant to be greater than 0.2 on light quark jets, in data and MC, in three transverse momentum bins. The error on the Monte Carlo efficiency if negligible if compared to the error on data.

Jet p_T [GeV]	MC efficiency	Data efficiency
30-50	93.2%	$(95.1 \pm 3.3)\%$
50-80	91.3%	$(91.1 \pm 1.6)\%$
80-120	91.8%	$(94.0 \pm 0.8)\%$

It is not possible to isolate a sample of photon+jet events with two highly quark-enriched jets with similar kinematica properties as the ones expected from the decay of a heavy Higgs boson. We will therefore have to simulate the effect of the decay kinematics, and will proceed as follows. We choose a threshold of 0.2 on the single jet Q-G likelihood distribution, as it is expected to provide an efficiency ϵ such that $\epsilon^2 \approx 85\%$, which is the expected signal efficiency of the cut applied in the analysis. The efficiency of this cut on quark jets is measured in data and MC by applying the requirement on the gluon-subtracted distributions shown in Figures 41, and is reported in Table 15.

As no significant deviation is observed between the data and the MC prediction, the uncertainty will originate from the statistical uncertainty of the comparisons. Therefore, we expect the lowest transverse momentum bin (30-50 GeV) to play the driving role. The kinematic properties of the jets in signal will depend on the mass of the decaying Higgs boson, being on average harder as the mass increases. In order to provide a conservative estimate of this systematic uncertainty, we have considered the case of a relatively light mass Higgs boson, 250 GeV, where the relative weight of the lowest p_T bin is inflated. In the decay of a 250 GeV Higgs boson, the jets are expected to populate the three transverse momentum bins as reported in Table 16. We estimate the uncertainty U on the single jet as the average of the statistical error

in the three transverse momentum bins, weighted on the expected fractions:

$$\mathrm{U(jet)} = \sum x_i \Delta_i$$

where $i = 1, 2, 3$ are the three transverse momentum bins, x_i is the fraction of jets which fall in the given bin, and Δ_i is the statistical uncertainty of that bin. We find:

$$\mathrm{U(lead)} = 22\% \cdot \left(\frac{0.8\%}{94.0\%}\right) + 64\% \cdot \left(\frac{1.6\%}{91.1\%}\right) + 11\% \cdot \left(\frac{3.3\%}{95.1\%}\right) = 1.7\%$$

$$\mathrm{U(sublead)} = 2.3\% \cdot \left(\frac{0.8\%}{94.0\%}\right) + 32\% \cdot \left(\frac{1.6\%}{91.1\%}\right) + 68\% \cdot \left(\frac{3.3\%}{95.1\%}\right) = 3.0\%$$

We then take the product of the uncertainties of the two jets as an estimate of the uncertainty on the cut of the product of the two likelihoods, and therefore find a total systematic uncertainty of $\mathrm{U(prod)} = 4.6\%$.

8.3.6 MET uncertainty

MET affects directly only the 2 b-tag category. The dominant effects are from the knowledge of the rest of the event, such as jet energy reconstruction and pileup. Therefore, both of the above subsections cover MET uncertainty to a large extent. The uncertainty is computed following the method also used in the low mass analysis (see section 8.4.1). Requirement on the MET significance translates into about 3% inefficiency and the resulting uncertainty does not surpass this value, as shown in fugure 42.

Figure 42: Efficiency in data and Monte Carlo (after $t\bar{t}$ subtraction) for various MET significance cuts in the 2 btag category after the preselection cuts. The difference between data and Monte Carlo efficiencies (magnified by 10) is also shown.

8.3.7 Production mechanism

The expected kinematics of the Higgs production is subject to uncertainties due to limited knowledge of the underlying parton distribution functions (PDFs) as well as the shortcomings in the theoretical prediction (missing higher orders in the perturbation series). These uncertainties are propagated to an uncertainty on the selection acceptance and efficiency. Their additional effect on the Higgs production cross section is discussed in a separate section below.

The PDF uncertainties is evaluated according to the PDF4LHC recommendations, by evaluating the selection efficiency for the PDF sets CT10 [23], MSTW2008NLO [24] and NNPDF2.1 [25] and their error sets. Table 8.3.7 summarizes the resulting acceptance uncertainties. The envelope of the various PDF sets is used as the total uncertainty, as recommended and amounts to 2-4% without strong dependence on Higgs mass or b-tag category.

Table 16: Efficiencies of requiring the quark-gluon discriminant to be greater than 0.2 on light quark jets, in data and MC, in three transverse momentum bins. The error on the Monte Carlo efficiency if negligible if compared to the error on data.

	30-50 GeV	50-80 GeV	80-120 GeV
Leading Jet	11%	64%	22%
Subleading Jet	68%	32%	2.3%

The uncertainty from the matrix element is evaluated with the help of the HQT program, which includes NLL effects and exhibits a modified Higgs p_T spectrum, especially for low Higgs masses. The nominal POWHEG sample is re-weighted as a function of Higgs p_T to match the HQT spectrum and the selection efficiency evaluated in each case. Figure 43 shows the generated Higgs p_T spectra from POWHEG and HQT for three Higgs masses. The effect of this re-weighting is strongest for the lowest Higgs mass studied (200 GeV): the efficiency drops by 3% as the softer Higgs p_T spectrum from HQT leads to slightly softer leptons and jets on average. The loss of efficiency depends only weakly on the b-tag category (0-tag: -2%, 1-tag: -5%, 2-tag: -3%). At higher Higgs masses (400 and 600 GeV) the difference in efficiency is less than one percent for all b-tag categories. The trend of decreasing uncertainties at higher Higgs masses is caused by two separate effects: Firstly, POWHEG and HQT show better agreement at higher Higgs masses. Secondly, the decay energy of the Higgs dominates over the Higgs p_T in determining the kinematics of the decay products for high Higgs masses.

Figure 43: Generated Higgs p_T distributions from POWHEG (red) and HQT (blue) several masses.

We additionally estimate the uncertainty that originates from the fact that the analysis has been tuned using gluon fusion based simulation while a real signal contains a mixture of events produced by gluon fusion and VBF. Here we compute the difference in signal acceptance between the two production mechinisms in Monte Carlo and multiply this difference with the expected fraction of VBF production, leading to a global uncertainty on the production cross section.

8.4 Low-Mass Selection

Uncertainties in the expected signal yields are mostly identical to those estimated in Table 11.

Uncertainties in the expected background yields and mass distributions shapes come from sideband extrapolation and are dominated by statistical uncertainties in the sideband samples. *Statistical uncertainties from Monte Carlo samples still to be finalized when the low-mass DY samples are available.*

Table 17: Summary of systematic uncertainties on the signal acceptance following PDF4LHC recommendations.

PDF	$M_H = 200$ GeV			$M_H = 400$ GeV			$M_H = 600$ GeV		
	0 b-tag	1 b-tag	2 b-tag	0 b-tag	1 b-tag	2 b-tag	0 b-tag	1 b-tag	2 b-tag
CT10	$^{+0.7\%}_{-3.7\%}$	$^{+0.9\%}_{-3.7\%}$	$^{+1.3\%}_{-4.6\%}$	$^{+0.8\%}_{-3.5\%}$	$^{+0.5\%}_{-3.3\%}$	$^{+0.9\%}_{-4.0\%}$	$^{+0.4\%}_{-2.6\%}$	$^{+0.4\%}_{-3.4\%}$	$^{+0.7\%}_{-4.3\%}$
all categories		$^{+1.0\%}_{-3.9\%}$			$^{+0.7\%}_{-3.4\%}$			$^{+0.6\%}_{-4.1\%}$	
MSTW2008NLO	$^{-0.6\%}_{-0.6\%}$	$^{0.5\%}_{-0.8\%}$	$^{0.9\%}_{-0.8\%}$	$^{+0.5\%}_{-0.5\%}$	$^{+0.4\%}_{-0.3\%}$	$^{+0.8\%}_{-0.3\%}$	$^{-0.30\%}_{-0.0\%}$	$^{+0.3\%}_{-0.2\%}$	$^{+0.5\%}_{-0.3\%}$
all categories		$^{0.6\%}_{-0.8\%}$			$^{+0.5\%}_{-0.4\%}$			$^{+0.4\%}_{-0.3\%}$	
NNPDF2.1	$^{+1.8\%}_{+0.3\%}$	$^{+1.7\%}_{+0.1\%}$	$^{+2.2\%}_{+0.0\%}$	$^{+1.4\%}_{+0.0\%}$	$^{+1.4\%}_{+0.3\%}$	$^{+1.5\%}_{+0.1\%}$	$^{+0.7\%}_{+0.1\%}$	$^{+1.1\%}_{+0.3\%}$	$^{+1.6\%}_{+0.3\%}$
all categories		$^{+1.8\%}_{+0.1\%}$			$^{+1.3\%}_{+0.2\%}$			$^{+1.4\%}_{\%+0.3}$	
Total	$^{+1.8\%}_{-3.7\%}$	$^{+1.7\%}_{-3.7\%}$	$^{+2.2\%}_{-4.5\%}$	$^{+2.3\%}_{-3.5\%}$	$^{+1.4\%}_{-3.3\%}$	$^{+1.5\%}_{-3.9\%}$	$^{+0.7\%}_{-2.6\%}$	$^{+1.1\%}_{-3.4\%}$	$^{+1.6\%}_{-4.3\%}$
all categories		$^{+1.8\%}_{-3.9\%}$			$^{+1.3\%}_{-3.4\%}$			$^{+1.4\%}_{-4.1\%}$	

8.4.1 MET resolution

The uncertainty on the MET cut signal efficiency is estimated by comparing Monte Carlo and data. We expect no real MET in the signal and Z+jets events reproduce quite well the features of signal events. The $t\bar{t}$ contribution, as predicted by Monte Carlo, is therefore subtracted to the MET and MET significance distributions (also in data), as shown in Fig. 44. The difference in efficiency of the chosen MET cut in data and Monte Carlo is then estimated as systematics. This systematics is shown in Fig. 45 and that stays below 4% for all the MET cut above 40 GeV.

Figure 44: MET and MET significance distributions in the 2 btag category after the preselection cutslisted in Sec.**??**. $t\bar{t}$ contribution, as predicted by Monte Carlo, has been subtracted, even in data.

Figure 45: Efficiency in data and Monte Carlo (after $t\bar{t}$ subtraction) for various MET (left) and MET significance cuts (right) in the 2 btag category after the preselection cutslisted in Sec.**??**. The difference between data and Monte Carlo efficiencies (magnified by 10) is also shown.

To estimate the systematics uncertainty on signal efficiency due to Pile-Up simulation, we shift the distribution of the number of observed Pile-Up events by ± 1. This choice, clearly arbitrary, is still expected to cover conservatively possible difference in Pile-Up between data and MC.

8.4.2 Jet energy scale

Similarly to what has been done for the high mass analysis (see Sec. 8.3.2), we compare the efficiency for signal events with jet energy scale modified by $\pm 1\sigma$. The results for several Higgs masses are reported in Tab. 20.

8.4.3 Production mechanism

Similarly to what has been done for the high mass analysis (see Sec. 8.3.2), we compare the efficiency for signal events in gg and VBF production. The results for several Higgs masses are reported in Tab. 21.

9 Statistical Analysis and Results

TO BE REVISED... COPY FROM 11-125

We determine the expected upper limits to the Standard Model Higgs production cross section as a function of the Higgs boson mass assuming as reference an integrated luminosity of 1 fb^{-1}.

Table 18: Summary of systematic uncertainties due the VBF.

M_H	$M_H = 200$ GeV			$M_H = 400$ GeV			$M_H = 600$ GeV		
	0 b-tag	1 b-tag	2 b-tag	0 b-tag	1 b-tag	2 b-tag	0 b-tag	1 b-tag	2 b-tag
Δ_{eff}	20%	15%	6%	-2%	4%	4%	-8%	6%	-0.3%
total	7%			11%			13%		
uncertainty	2.4%	1.8%	0.7%	-0.16 %	0.3 %	0.3 %	-1.4 %	1 %	0%
total	2%			0.1%			-1.4%		

Table 19: Systematics on Pile-Up simulation

Higgs mass	expected yields	yields for PU-1	yields for PU+1	systematics
150	13.36	13.58	13.09	-2.0%, + 1.6%
130	2.49	2.56	2.44	-2.0%, +2.8%

Table 20: Systematics on jet energy scale: yields are quoted for 4.6 fb^{-1}

Higgs mass	yields at nominal JES	yields with JES $+1\sigma$	yields with JES -1σ	systematics
170	2.14	2.04	2.06	-4.7%, -3.7%
160	7.91	8.23	7.24	+4.0%, -8.5%
150	13.36	13.85	12.62	+3.7%,-5.5%
140	10.3	10.8	9.8	+4.4%, - 5.3%
130	2.49	2.7	2.3	+8.4 %, -7.6%

We use the official tool developed by the CMS Higgs combination group [26] that supports different methods, including Bayesian approach, frequentist profile likelihood, and Feldman-Cousins [27] methods, and the modified frequentist CLs method [28] with Cousins-Highland integration of nuisance parameters for the treatment of systematic uncertainties [29]. The tool uses the `RooStats` [30] engine from ROOT as internal implementation. We determine the expected upper limit to the Higgs boson production cross section times branching fraction to $\ell\ell b\bar{b}$. The limit is expressed as ratio r of the determined upper limit to the cross section times branching ratio divided by its standard model expectation. A value of the Higgs boson mass M_{H} is excluded if, for that mass hypothesis, r is less than one.

There are two main possible general approaches to determine the limit to the Higgs cross sections.

The simples approach, referred to as "cut and count" analysis, uses only the number of events selected within a given window in the reconstructed Higgs mass, , around an assumed value M_{H} of the Higgs boson mass. The Higgs cross section limit is determined from the expected number of signal and background events passing the selections s and b respectively. We combine using Poissonian statistics the counting information from the two channels with electrons and muons. The sources of systematic uncertainties on s and b are taken into account in the determination of the limit assuming log-normal distributions of the nuisance parameters.

Another possible approach, referred to as "shape analysis", takes into account the measured distribution of variables that can discriminate Higgs signal events against background events. In the case of the analysis, the main discriminant variable is the reconstructed Higgs boson mass, , which is peaked around the true Higgs boson mass, M_{H}, for the signal and has a broader distribution for background processes. We can develop an analysis that does not select signal events cutting on , and we can use the distribution of for selected events, comparing to the expected distribution from signal and background. The possible sources of systematic

Table 21: Systematics on Higgs production

Higgs mass	efficiency in gg	efficiency in VBF	systematics on VBF component	systematics on total yi
150	2.49%	2.25%	-11%	-0.9%
130	0.713%	0.837%	-15%	-1%

uncertainties on both the expected signal and background yields and the shape of signal and background distributions are considered in the limit extraction procedure. The shape analysis can be implemented in the Higgs combination tool using an unbinned approach with an extended likelihood function defined by:

$$-\ln\mathcal{L} = \sum_{j=1}^{k}\left[s_j + b_j - \sum_{i=1}^{n_j}\left(s_j\mathcal{P}_s^{(j)}(m_i^{(j)}) + b_j\mathcal{P}_b^{(j)}(m_i^{(j)})\right)\right], \tag{3}$$

where k is the number of channels (in our case $k = 2$, and $j = 1, 2$ correspond to the electon and muon channels), n_j is the number of selected candidate events in the channel j, s_j and b_j are the expected signal and background events, $\mathcal{P}_s^{(j)}$ and $\mathcal{P}_b^{(j)}$ are the probability density functions for signal and background, and $m_i^{(j)}$ are the n_j values of of the selected candidates in the channel j. As alternative, the distribution of can be sub-divided into bins, and in that cases the number of selected events in each bin is considered with the corresponding expected number of events from signal and background in that bin, and the information from all bins is combined using a Poissonian likelihood, similarly to the "cut and count" case, but with more channels.

Results combining all the H→ZZ→ 2ℓ2q categories (with and without tagging) are described in Reference [31].

10 Conclusion

This analysis is great. It's full of new and exciting features, each increasing the power of the analysis by orders of magnitude.

References

[1] G. Abbiendi et al., The ALEPH, DELPHI, L3 and OPAL Collaborations, The LEP Working Group for Higgs Boson Searches, "Search for the Standard Model Higgs Boson at LEP", *Phys. Lett* **B565** (2003) 61–75. doi:10.1016/S0370-2693(03)00614-2.

[2] T. Aaltonen et al, The CDF and D0 Collaborations, "Combination of Tevatron searches for the standard model Higgs boson in the W+W- decay mode", *Phys. Rev. Lett.* **104** (2010) 061802. doi:10.1103/PhysRevLett.104.061802.

[3] T. Aaltonen et al, The CDF, D0 Collaborations, the TEVNPHWG Working Group, "Combined CDF and D0 Upper Limits on Standard Model Higgs Boson Production with up to 8.2 fb-1 of Data", *FERMILAB-CONF-11-044-E* (2011) arXiv:1103.3233.

[4] The ALEPH, CDF, D0, DELPHI, L3, OPAL, SLD Collaborations, the LEP Electroweak Working Group, the Tevatron Electroweak Working Group, and the SLD electroweak and heavy flavour groups, "Precision Electroweak Measurements and Constraints on the Standard Model", *LEPEWWG* **2010-01** (2010) arXiv:1012.2367.

[5] W. Adam, V. Adler, B. Hegner, L. Lista, S. Lowette, P. Maksimovic, G. Petrucciani, S. Rappoccio, F. Ronga, R. Tenchini and R. Wolf, "PAT: The CMS Physics Analysis Toolkit", *Jour. of Phys., Conf. Series* **219** (2010) 032017. doi:10.1088/1742-6596/219/3/032017.

[6] CMS Collaboration, ""Electron reconstruction and identification at $\sqrt{s}$=7 TeV"", CMS Physics Analysis Summary CMS-PAS-EGM-10-004, (2010).

[7] CMS Collaboration, ""Performance of muon identification in pp collisions at $\sqrt{s}$=7 TeV"", CMS Physics Analysis Summary CMS-PAS-EGM-10-004, (2010).

[8] CMS Collaboration, "Jet Performance in pp Collisions at $\sqrt{s}$=7 TeV", CMS Physics Analysis Summary CMS-PAS-JME-10-003, (2010).

[9] CMS Collaboration, ""Commissioning of the Particle-Flow Reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV"", CMS Physics Analysis Summary CMS-PAS-PFT-10-002, (2010).

[10] CMS Collaboration, ""Trigger strategies for Higgs searches in 2011"", CMS Analysis Note CMS-AN-2011-065, (2011).

[11] CMS Collaboration, ""Search for the Standard Model Higgs Boson in the Decay Channel $H \to ZZ \to 2l2b$"", CMS Analysis Note CMS-AN-2011-399, (2011).

[12] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm ", *JHEP* **04** (2008) 063. doi:10.1088/1126-6708/2008/04/063.

[13] CMS Collaboration, ""Jet Energy Corrections determination at 7 TeV"", CMS Physics Analysis Summary CMS-PAS-JME-10-010, (2010).

[14] Particle Data Group Collaboration, "Review of particle physics", *J. Phys.* **G37** (2010) 075021. doi:10.1088/0954-3899/37/7A/075021.

[15] Y. Gao, A. V. Gritsan, Z. Guo et al., "Spin determination of single-produced resonances at hadron colliders", *Phys.Rev.* **D81** (2010) 075022, arXiv:1001.3396. doi:10.1103/PhysRevD.81.075022.

[16] A. De Rujula, J. Lykken, M. Pierini et al., "Higgs look-alikes at the LHC", *Phys.Rev.* **D82** (2010) 013003, arXiv:1001.5300. doi:10.1103/PhysRevD.82.013003.

[17] CMS Collaboration, "Angular Analysis of Resonances $pp \to X \to ZZ$", CMS Analysis Note CMS-AN-2010-351, (2010).

[18] CMS Collaboration, "Quark-Gluon Jet Discrimination through Particle Flow Jet Structure", CMS Analysis Note CMS-AN-2011-215, (2011).

[19] A. Hocker et al. arXiv:0703039.

[20] "Mistag note reference should go here",.

[21] CMS Collaboration, "Absolute jet energy correction uncertainty", CMS Analysis Note CMS-AN-2010-304, (2010).

[22] CMS Collaboration, ""Jet Energy Resolution in CMS at $\sqrt{s} = 7$ TeV"", CMS Physics Analysis Summary CMS-PAS-JME-10-014, (2010).

[23] H.-L. Lai, M. Guzzi, J. Huston et al., "New parton distributions for collider physics", *Phys.Rev.* **D82** (2010) 074024, arXiv:1007.2241. doi:10.1103/PhysRevD.82.074024.

[24] A. Martin, W. Stirling, R. Thorne et al., "Parton distributions for the LHC", *Eur.Phys.J.* **C63** (2009) 189–285, arXiv:0901.0002. doi:10.1140/epjc/s10052-009-1072-5.

[25] R. D. Ball, V. Bertone, F. Cerutti et al., "Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology", *Nucl.Phys.* **B849** (2011) 296–363, arXiv:1101.1300. doi:10.1016/j.nuclphysb.2011.03.021.

[26] C. H. C. Group, "Documentation of the RooStats-based statistics tools for Higgs PAG", *CMS TWiki* SWGuideHiggsAnalysisCombinedLimit (2011).

[27] G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals", *Phys. Rev* **D 57** (1998) 3873–3889. doi:10.1103/PhysRevD.57.3873.

[28] A. L. Read, "Modified Frequentist Analysis of Search Results (The CLs Method)", *CERN OPEN* **2000-205** (2000).

[29] R. D. Cousins and V. L. Highland, "Incorporating systematic uncertainties into an upper limit", *Nucl. Instr. Meth* **A 320** (1992) 331–335. doi:doi:10.1016/0168-9002(92)90794-5.

[30] L. Moneta, K. Belasco, K. Cranmer et al., "The RooStats Project", in *13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2010)*. SISSA, 2010. arXiv:1009.1003. PoS(ACAT2010)057.

[31] S. Bolognesi, A. Bonato, A. D. Del Re, A.and Gritsan et al., "Search for a Semileptonic Decay of a SM Higgs or BSM Boson $H \to ZZ \to 2l2j$", *CMS Note* **CMS-AN-2011/100** (2011).

A Neural Network

A.1 Signal Optimization Based on Helicity Neural Network

In order to gauge the effectiveness of the currently used helicity likelihood discriminant a parallel multivariate analysis has been peformed. A Neural Network is applied to separate Higgs signal from background using the five helicity angles of the final objects in the analysis as inputs. A training and test evaluation has been performed with the framework of the TMVA package [19] using the real mixture of MC processes as background and Higgs MC. Both the training and testing samples for both background and signal are constructed and events randomly mixed outside of the Neural Network. There are 37,874 signal events and 38,378 background events evenly split between testing and training. The Neural Network was trained on the Monte Carlo generated for a hypothetical Higgs mass of 400 GeV/c^2. This training was on events that passed the previously explained preselection, the Z boson mass window cuts previously explained, but before the cut on MET significance.

A.1.1 Neural Network Architecture

The architecture of the Neural Network consists of two hidden layers with N and N neurons respectively, where N is the number of variables, and one output node as shown in Figure 46.

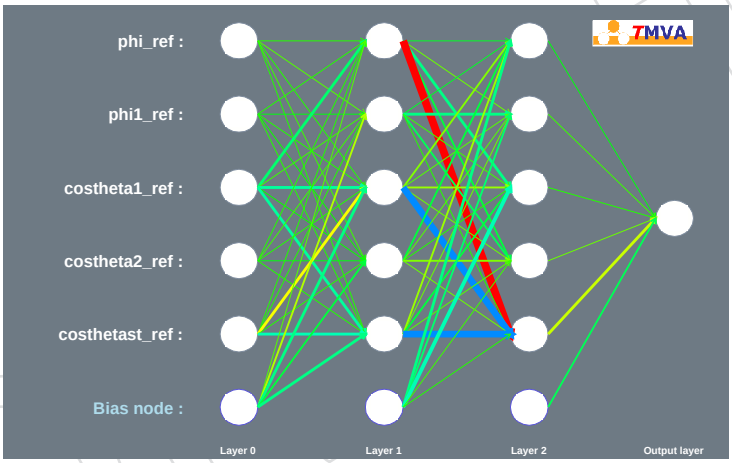


Figure 46: Neural Network architecture used for the training.

A.1.2 Input Variables and Training

The input variables for the Neural Network are the same variables that are used for the Helicity Likelihood Discriminant, phi, phi1, costheta1, costheta2, and costhetastar. The input variable distributions for both signal and background are shown in Figure 47. The default values are used when booking the Multi Layer Perceptron Neural Network except for the Hidden Layers as described in the previous subsection, and the number of cycles which is set to 1000. The background rejection versus signal efficiency for the input signal and background is shown in Figure 48.

A.1.3 Results Based on Helicity Neural Network

The training and testing of the Neural Network can bee seen in Figure 49. While there is separation between the signal and background, and good agreement between testing and training,

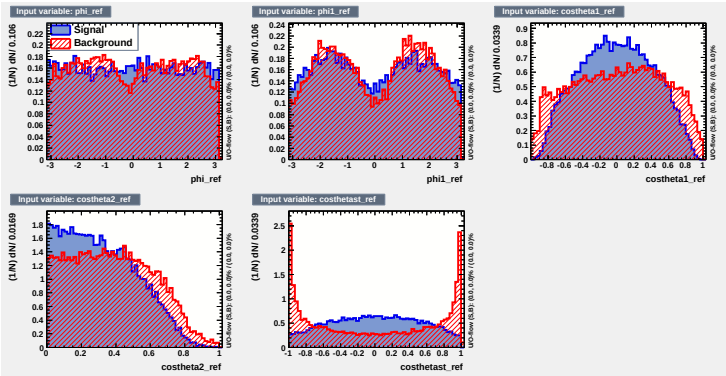


Figure 47: Signal and Background for the the Input Variables for the Neural Network training.

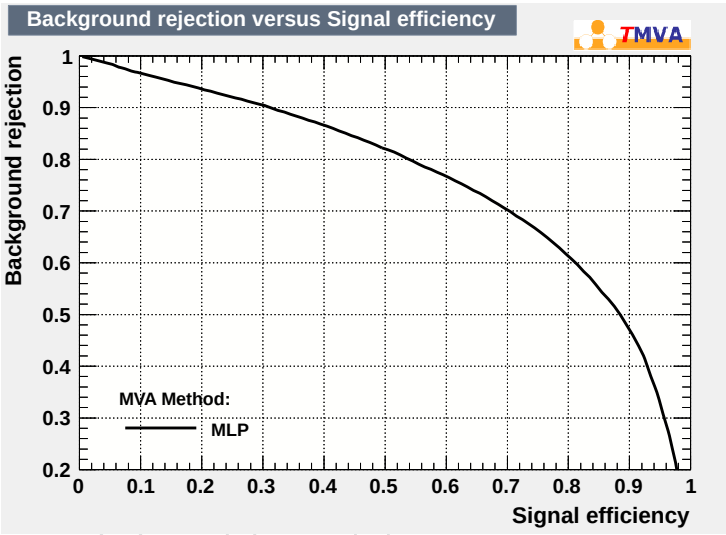


Figure 48: Background Rejection Versus Signal Efficiency for the input and output of the Neural Network.

there is a definate spike in background at the same place as the signal spikes in the MLP training, with a similar but less notacable effect for the Likelihood trinaing. The training was done with Monte Carlo generated for a hypothetical Higgs mass of 400 GeV but then applied to the Monte Carlo for hypothetical Higgs masses of 200, 300, 400, and 500 GeV. This is because the angular components should not depend on Higgs mass. In addition to the MLP neural network done each training and testing was additionaly performed with a Likelihood. The performance of the two analysis is almost identical in all cases, especialy when looking in tight hypotetical Higss mass regions.

In the 1tag region after applying the additional MET significance cut the separation between signal and background looks similar to the training for Higgs 300, 400, and 500 GeV, but does not have much discrimination power for a Higgs of 200 GeV. The discrimination power is virtually the same after applying an additional cut of of -6%/+10% of the Higgs mass for all four cases as seen in Figure 50. When comparing background rejection versus signal efficiency between the Neural Network and Likelihood Discriminate the performance is for all practical purposes the same. See Figure 52.

In the 2tag region after applying the additional MET significance cut and a cut of of -6%/+10% of the Higgs mass, the discriminating power of the Neural Network is similar to the 1tag case with poor ability for a Higgs of 200 GeV, but good seperation for Higgs of 300, 400, and 500 GeV. This is shown in Figure 51. When comparing background rejection versus signal efficiency between the Neural Network and Likelihood Discriminate the performance is roughly the same, except for the Higgs of 200 GeV case where the Neural Network is almost consistantly better than the Likelihood Discriminate. This is shown in Figure 53.

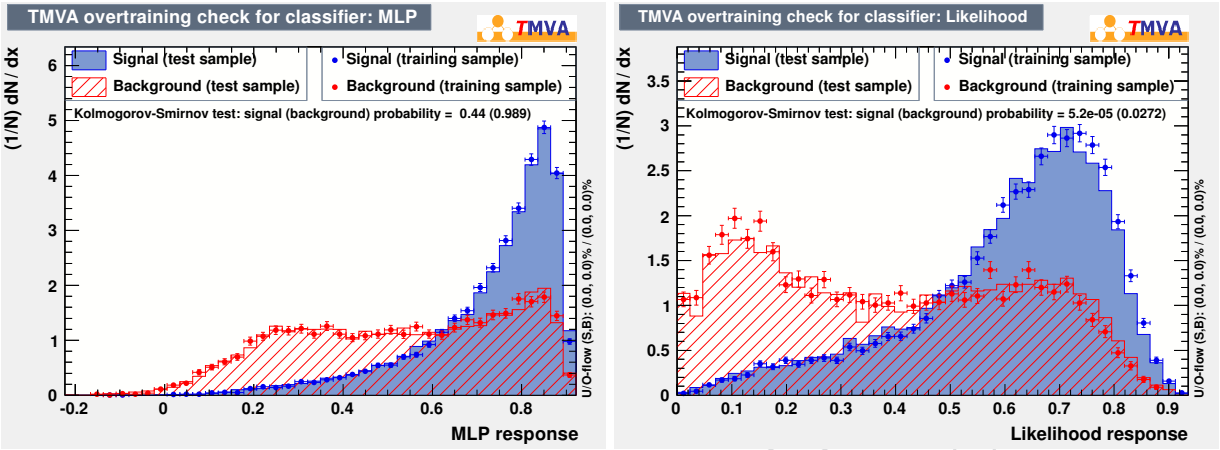


Figure 49: The trainings are done after preselection and addisionaly require at least one TCHEM jet (Track Counting High Efficiency > 1.9) Left: Training 400 GeV Higgs boson with a MLP neural network. Right: Training 400 GeV Higgs boson with a Likelihood.

A.1.4 Potential Improvement to to MVA

A neural network or likelihood should be able to take advantage of extra information then simply uisng the 5 decay angles as previously shown. One such variable that offeres good discrimination power in the preselection region is $Z_{ll}pt$ / $\sum pt$, where $\sum pt$ = l0pt + l1pt + j0pt + j1pt + met, and $Z_{ll}pt$ and $\sum pt$ are scalar quantities. This variable distribution is shown in Figure 55. While adding this variable to a MLP training improves the discrimination power in the inclusive pretag region once we look in a mass window of -6%/+10% around Higgs 400 adding $Z_{ll}pt$ / $\sum pt$ to the training actually lowers the descrimination power of the neural network. This same performance drop is seen when training is done with the addition of the reconstructed higgs mass as well.

While training on each Higgs mass can give addisional descrimination power over trainng on one Higgs mass and applying it to multiple hypothetical Higgs masses there are some difficultis. It is more conveniant to be able to train on just one higgs mass and then apply this training to a range of hypothetical higgs masses. An example of training on a Higgs 400 GeV sample and then applying this training to various hypothetical Higgs masses is show in Figure 56. This training is for a MLP neural network tranined on the 5 decay angles. This training is applied after preselection and requiring at least one TCHEM jet. This shows that good performance can be acheived using the MLP by only training on one Higgs mass sample, without the need for aditional trainings, or parameratising a descriminate as a function of the reconstructed Higgs mass.

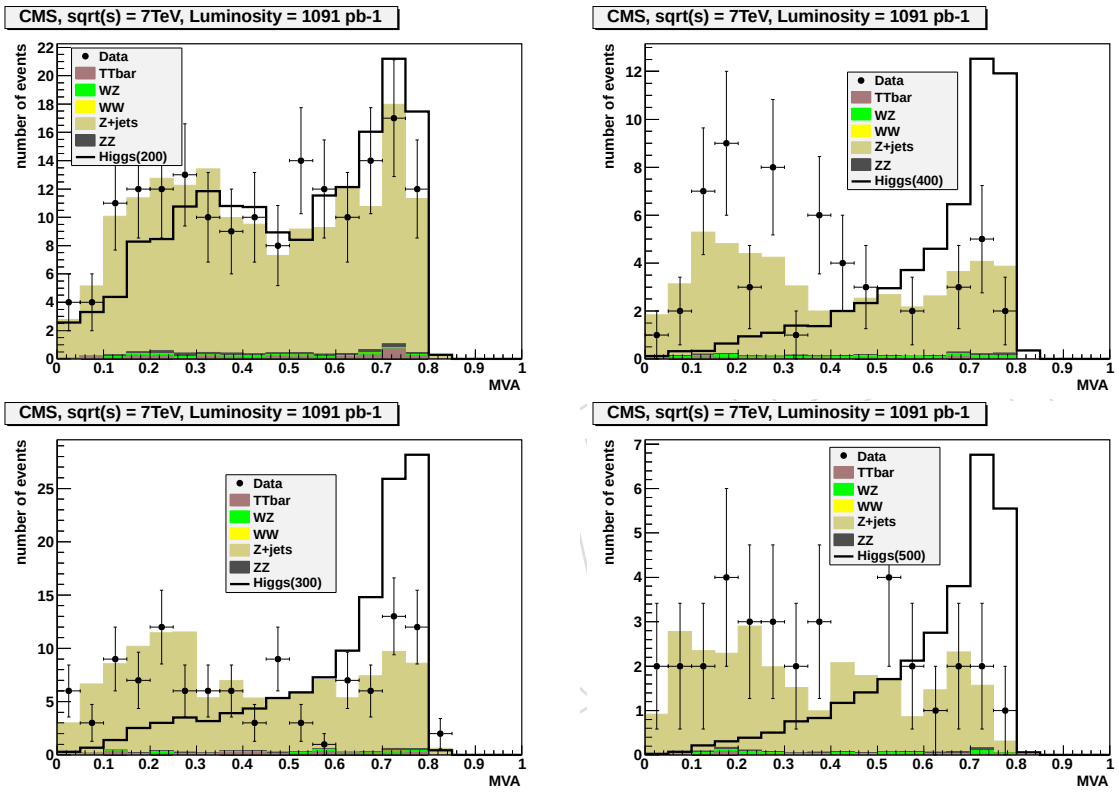


Figure 50: Signal and background Neural Network output in the 1tag region after preselection, the Z boson mass window cuts, cut on MET significance, and a -6%/+10% Higgs mass window. The signal is scaled to the sum of the Monte Carlo background.

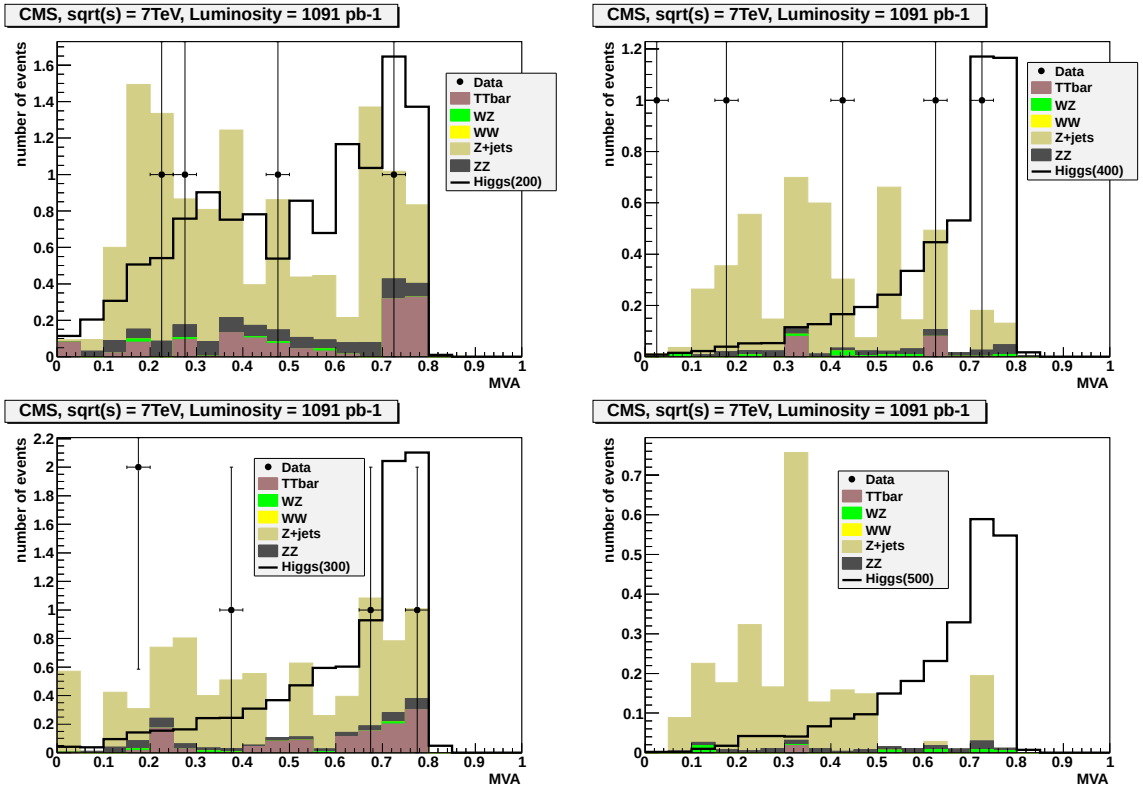


Figure 51: Signal and background Neural Network output in the 2 tag region after preselection, the Z boson mass window cuts, cut on MET significance, and a -6%/+10% Higgs mass window. The signal is scaled to the sum of the Monte Carlo background.

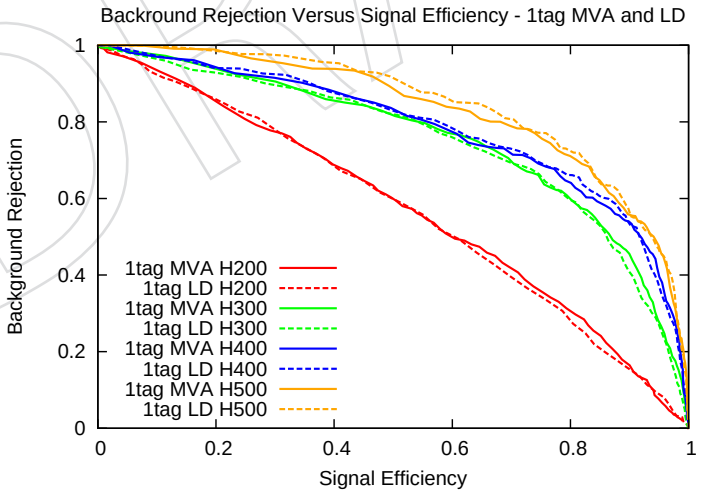


Figure 52: Background Rejection Versus Signal Efficiency in the 1tag region comparing the Multi Variant Analysis output to the the Helicity Likelihood Discriminant for a Higgs mass of 200, 300, 400, and 500 GeV. This is calculated after preselection cuts, Z boson mass cuts, cut on MET significance, in a -6%/+10% Higgs mass window.

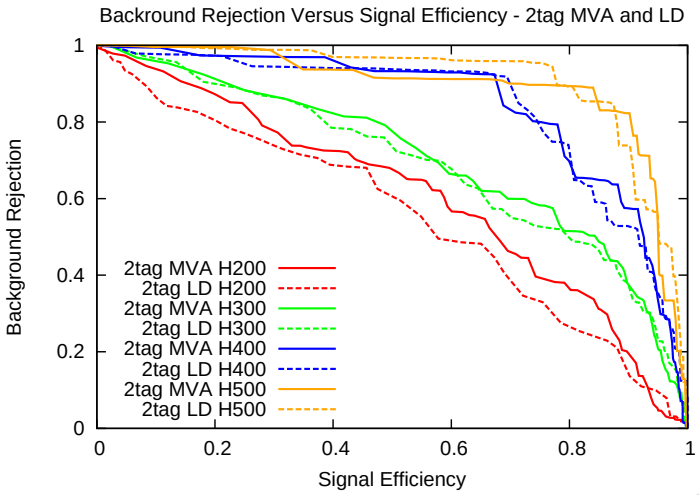


Figure 53: Background Rejection Versus Signal Efficiency in the 2tag region comparing the Multi Variant Analysis output to the the Helicity Likelihood Discriminant for a Higgs mass of 200, 300, 400, and 500 GeV. This is calculated after preselection cuts, Z boson mass cuts, cut on MET significance, in a -6%/+10% Higgs mass window.

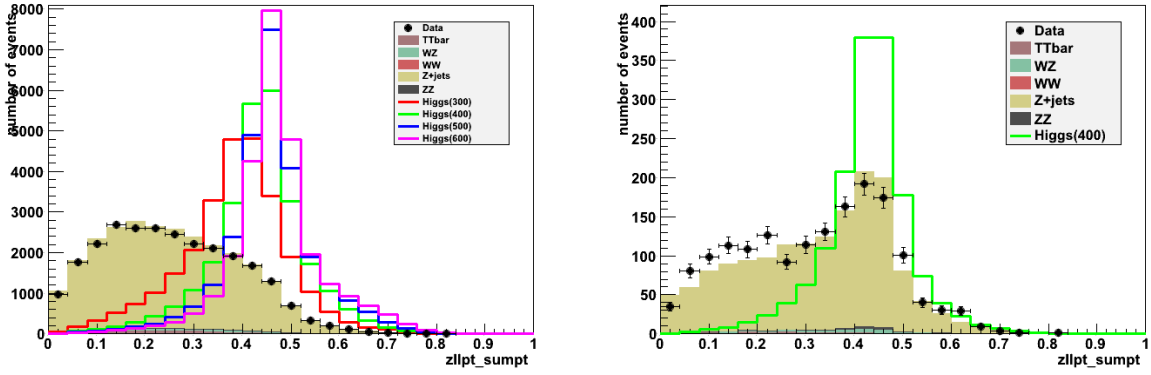


Figure 54: Signal samples are normalized to background. Left: $\frac{Z_{ll}pt}{\sum pt}$ after preselection. Right: $\frac{Z_{ll}pt}{\sum pt}$ after preselection and $376 < m_{ZZ} < 440$ GeV.

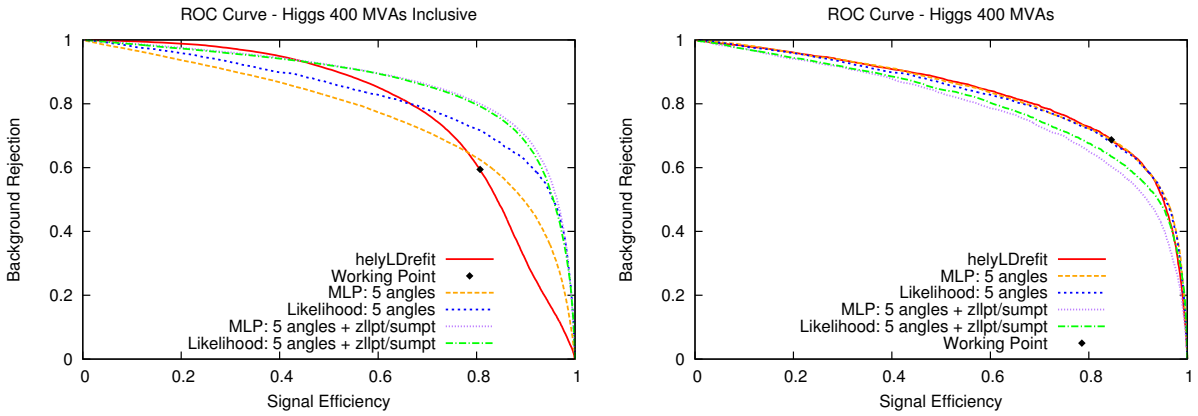


Figure 55: Applying a MLP training to preselection and at least one TCHEM jet. The working point is the equivelent performance of current analysis that we apply in the two tag region. For comparison the helyLDrefit varialbe is also shown. Left: ROC curves after preselection. Right: ROC curves after preselection and $376 < m_{ZZ} < 440$ GeV.

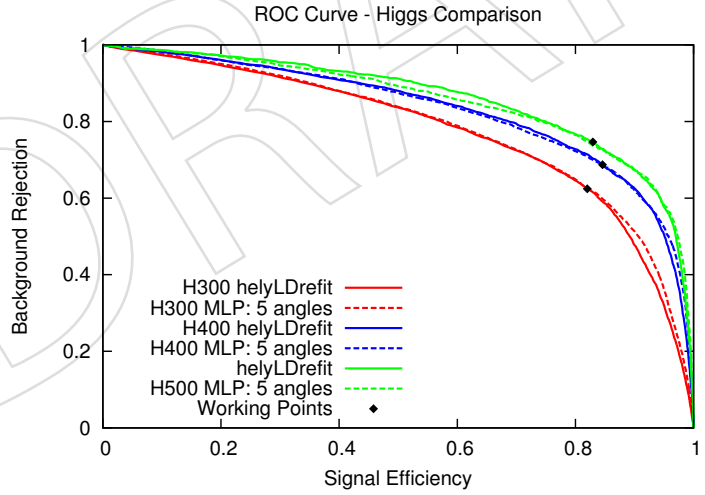


Figure 56: This training is for a MLP neural network tranined on the 5 decay angles of a Higgs 400 GeV sample. This training is applied after preselection and requiring at least one TCHEM jet to samples with a Higgs mass of 300,400, and 500 GeV. The working point is the background rejection point that we currently acheive in the two tag region in our analysis. The helyLDrefit variable is shown for comparison.