This Dissertation

entitled

SEARCH FOR LEPTON FLAVOR VIOLATING DECAYS

OF HIGGS BOSONS

WITH THE CMS EXPERIMENT

typeset with NDdiss2$_\varepsilon$ v3.2017.2 (2017/05/09) on May 8, 2019 for

Nabarun Dev

This LaTeX 2$_\varepsilon$ classfile conforms to the University of Notre Dame style guidelines as of Fall 2012. However it is still possible to generate a non-conformant document if the instructions in the class file documentation are not followed!

> Be sure to refer to the published Graduate School guidelines at `http://graduateschool.nd.edu` as well. Those guidelines override everything mentioned about formatting in the documentation for this NDdiss2$_\varepsilon$ class file.

This page can be disabled by specifying the "`noinfo`*" option to the class invocation.* (i.e.,`\documentclass[...,noinfo]{nddiss2e}`)

This page is *NOT* part of the dissertation/thesis. It should be disabled before making final, formal submission, but should be included in the version submitted for format check.

NDdiss2$_\varepsilon$ documentation can be found at these locations:

`http://graduateschool.nd.edu`
`https://ctan.org/pkg/nddiss`

SEARCH FOR LEPTON FLAVOR VIOLATING DECAYS OF HIGGS BOSONS WITH THE CMS EXPERIMENT

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in
Physics

by

Nabarun Dev

Colin Philip Jessop, Director

Graduate Program in Physics
Notre Dame, Indiana

May 2019

SEARCH FOR LEPTON FLAVOR VIOLATING DECAYS OF HIGGS BOSONS WITH THE CMS EXPERIMENT

Abstract

by

Nabarun Dev

DEDICATED TO

To my family

CONTENTS

FIGURES

TABLES

PREFACE

Long time ago in a galaxy far far away....(preface is optional)

ACKNOWLEDGMENTS

I would like to acknowledge the light side of the force, Master Kenobi and Grand Master Yoda.

SYMBOLS

c	speed of light
m	mass
e	elementary charge
E	energy

CHAPTER 1

INTRODUCTION

The standard model of particle physics is the most complete description of nature available today. The discovery of the Higgs Boson added another feather to the hat of the standard model...

...expand...

CHAPTER 2

THEORETICAL BASES

2.1 The Standard Model

2.2 Physics beyond the standard model

CHAPTER 3

EXPERIMENTAL SETUP

..intoduce...

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [1] is a powerful proton-proton synchrotron. It was built and is operated at the European Center for Nuclear Research (CERN) and is situated about 100 m underground close to Geneva, Switzerland. It has a circumference of 26.7 km and uses a tunnel previously built for LEP (Large Electron Positron Collider). Being a particle-particle collider, it consists of two rings with counterrotating beams which are steered using magnets and accelerated using radiofrequency resonating cavities. These beams are made to intersect at four collision points around the LHC ring, at one of which rests the CMS detector. Besides proton-proton collisions the LHC can also collide heavy ions (lead-lead collisions) or heavy ions with protons (lead-proton collisions). Since starting operation in September 2008 the LHC has been the world's most powerful apparatus and will probably remain so in the forseeable future. The following section describes proton-proton collisions at the LHC as the data used in the subsequent physics analysis corresponds to events from these collisions.

The injector chain that supplies protons to the LHC consists of four CERN accelerators that actually predate the LHC: Linac 2, PSB (Proton Synchroton Booster), PS (Proton Synchotron) and SPS (Super Proton Synchotron). This is illustrated in figure **??**. The proton source is simply a tank of hydrogen gas. The hydrogen atoms

are ionized to yield protons which are then fed in to the Linac 2, a linear accelerator. This accelerates the protons to an energy of about 50 MeV which are then fed into a series of circular accelerators starting with the PSB which accelerates the protons to 1.4 GeV. The PS then accelerates them to 25 GeV, and they are then sent to the SPS which accelerates them to 450 GeV before being finally fed into the LHC beampipe. Inside the LHC the protons are accelerated by sixteen radiofrequency cavities which are made to oscillate at 400 MHz and the proton beam is sorted into discrete packet called 'bunches'. The beam is steered by 1232 Niobium-Titanium superconducting dipole magnets and collimated using quadrupole magnets. This magnet system is kept at a temperature below 2 K, using a pressurised bath of superfluid helium at about 0.13 MPa, and operates at fields above 8T. The LHC has three sophisticated vacuum systems: the insulation vacuum for cryomagnets, the insulation vacuum for helium distribution, and the beam vacuum.

It takes about 4 minutes and 20 seconds to fill up each of the LHC rings wih protons, and about 20 minutes for the proton beam to reach its current peak energy of 6.5 TeV. At this point, each LHC beam contains 2808 bunches with 1.5×10^{11} protons per bunch, colliding at a center of mass energy (COM) of 13 TeV. It is anticipated for the COM energy to increase to 14 TeV in 2018. Looking for physics beyond the standard model by colliding protons at such high energies is one of the primary aims of the LHC.

Another important parameter for a collider like the LHC is the instantaneous luminosity (referred to as just luminosity in the following), $\mathscr{L}$. The number of events (N) generated per second for some processes is given by:

$$\frac{dN}{dt} = \sigma \mathscr{L} \tag{3.1}$$

where σ is the cross-section of the processes. The luminosity of the LHC can be

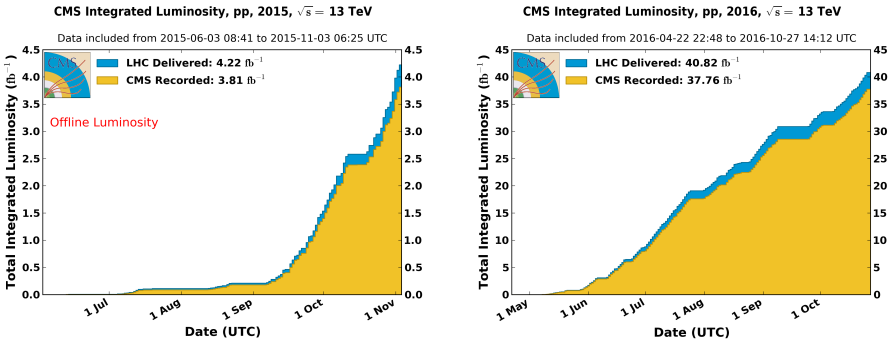


Figure 3.1: Evolution of integrated luminosity in 2015 and 2016 delivered by LHC (blue), and collected by CMS detector (orange) [2].

also expressed in terms of only beam parameters as:

$$\mathscr{L} = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta^*} F \tag{3.2}$$

where N_b is number of protons in a bunch, n_b is number of bunches per beam, f_{rev} is the revolution frequency, γ_r the relativistic gamma factor, ϵ_n the transverse beam emittance, β^* the beta function at the collision point, and F is a reduction factor coming from the fact that the beams cross at an angle.

This luminosity intergrated over time represents the total number of events collected per unit cross section and is called the integrated luminosity (L).The LHC has already reached its nominal design luminosity of $10^{34} cm^{-2} s^{-1}$, and it has delivered data amounting to a more than $36 fb^{-1}$, only in 2016. Figure 3.1 shows the amount of data delivered by the LHC overlaid with the subset collected by the CMS detector in 2015 and 2016.

In the longer term, it is planned to keep the LHC running, punctuated with several scheduled stops for upgrades and maintenance, at least until late 2030s. During this period it is anticipated to operate at increasingly higher luminosities helping collect unprecedented amounts of data. Figure 3.2 shows an overview of the long term LHC schedule.

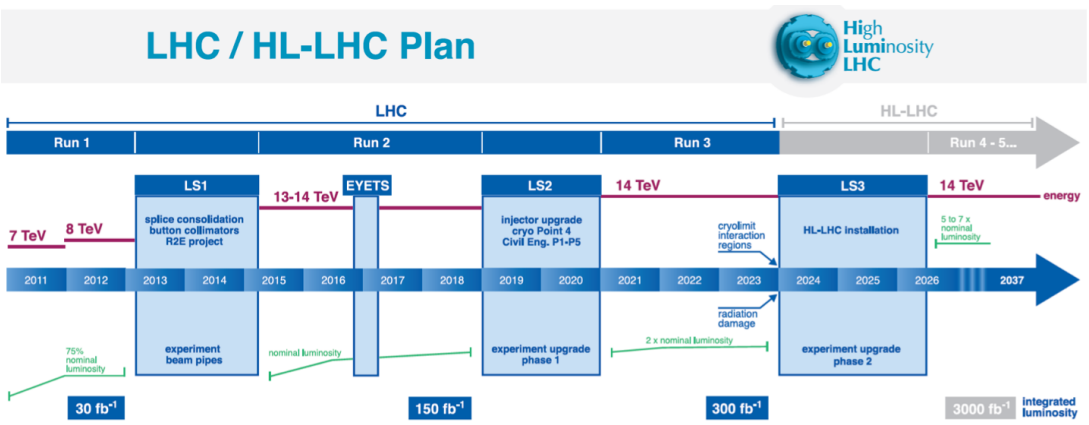


Figure 3.2: Overview of the long term LHC schecdule [3].

3.2 The CMS Detector

The Compact Muon Solenoid [4] is a general multipurpose particle physics detector that is placed in one of the four collision points of the LHC. It is 28.7m long with a diameter of 15.0m, weighs 14000 tons and is composed of several subdetectors. Its aim is to study a broad array of physics, from making precise measurements of known processess to searches for exotic processes predicted by a multitude of BSM theories. In order to be able to pursue its physics aims at the challenging LHC conditions, the CMS experiment needs to meet several requirements which primarily include good muon identification and momentum resolution over a wide range of momenta and angles, good dimuon mass resolution, good charged-particle momentum resolution and reconstruction efficiency, good electromagnetic energy resolution, good diphoton and dielectron mass resolution, good missing-transverse-energy and dijet-mass resolution. The backbone of the CMS is a superconducting solenoid that houses its tracking and calorimetry systems and provides an axial magnetic field of 3.8T. The inner-most layer is the silicon pixel and strip tracker that measures the trajectories of charged particles. Surrounding the tracker are the lead tungstate crystal electromagnetic calorimeter (ECAL) which measures the energy of electrons and photons,

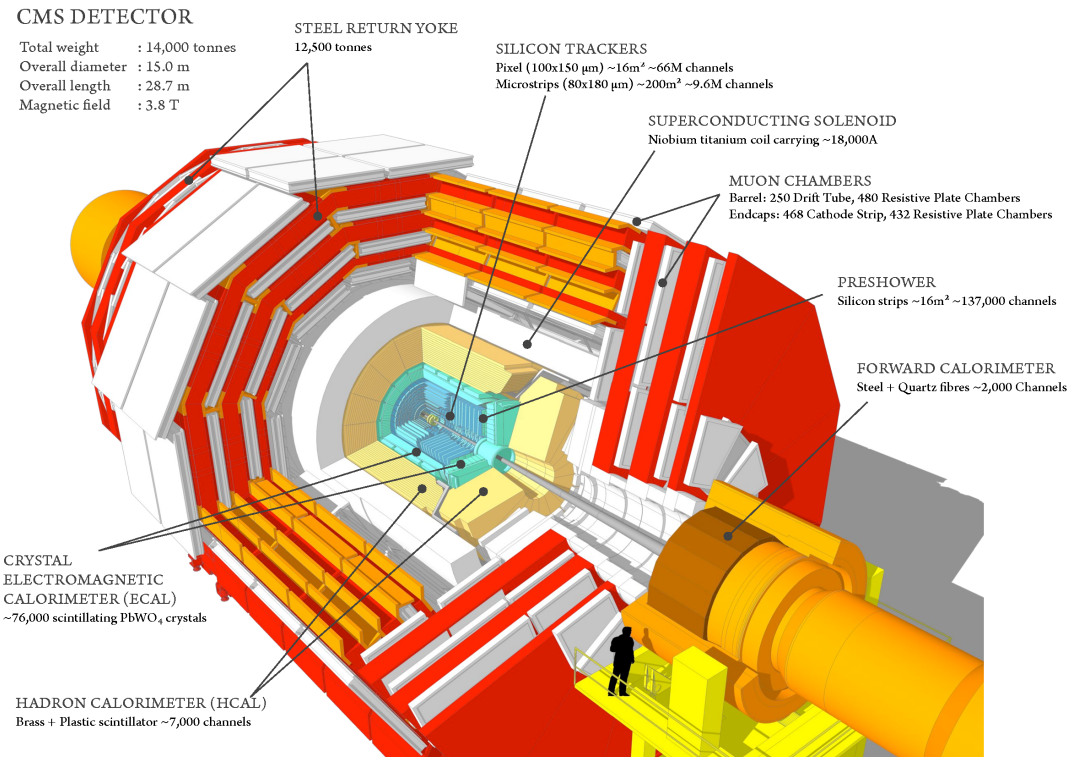


Figure 3.3: Layered View of the CMS detector

and the hadronic calorimeter (HCAL) which measures the energy of heavier particles that pass through the ECAL. The ECAL also contains a preshower detector for extra spatial precision. Outside the solenoid is the muon system which has gas-ionization detectors placed in the steel yoke of the magnet. This is the outermost component of CMS and measures the momenta of muons that traverse through it. A sophisticated two-level trigger system that helps filter out a small fraction of most interesting events among millions produced at the LHC also forms a vital part of the CMS.The powerful solenoid, sophisticated muon system and its compact design (given its complexity) give CMS its name. Figure 3.3 shows a layered view of the detector. The following sections describe it in further detail.

3.2.1 Coordinate Conventions

The CMS detector has adopted a right-handed coordinate system, the origin of which lies at the nominal collision point inside the experiment.The x-axis points radially inward towards the center of the LHC while the y-axis points vertically upwards. This makes the z-axis point along the beam direction. At point 5 of LHC (a village named Cessy in France) where the CMS is, the z axis points toward the Jura Mountains. In cylindrical co-ordinates, the polar angle θ is measured from the z-axis while the azimuthal angle ϕ is measured from the x-axis in the x-y plane. The polar angle is used to define the pseudo-rapidity $\eta = -ln(tan(\frac{\theta}{2}))$ which is a close approximation for rapidity if $E \gg m$. The rapidity is a Lorentz invariant quantity under boosts in the z-direction. Since it is typical of particles that CMS sees to have $E \gg m$, the Lorentz invariance approximately holds for pseudo-rapidity as well.

3.2.2 CMS Trigger

3.2.3 Charged Particle Tracking System

3.2.4 Electromagnetic Calorimeter

3.2.4.1 ECAL trigger project

3.2.4.2 Anomaly detection project for ECAL DQM

3.2.5 Hadronic Calorimeter

3.2.6 Muon System

CHAPTER 4

OBJECT RECONSTRUCTION AND EVENT GENERATION

4.1 Introduction

This chapter is divided into two parts. In the first part, the procedure for the generation of simulated events is described. This is done in several distinct stages with the output of one stage serving as an input for the next. A suite of software packages, developed mostly by the particle and nuclear physics communities, is used to achieve this. This part concludes by detailing the simulated datasets used in the analyses described in this thesis. In the second part of this chapter, the reconstruction of physics objects is described in detail. It starts with a description of the particle flow algorithm which is kind of a global event reconstruction scheme for the entire event. This is followed by descriptions of track , muon and electron reconstructions. Reconstruction of jets is described next followed by description of composite objects used in the analysis such as collinear mass and transverse mass. Brief desciptions of tau lepton reconstruction and b-tagging of jets is also included.

4.2 Event Simulation

A pp collision at the LHC, like any hadronic collision, is more complex than the hard interaction of two participating partons. The proton being a composite object, the colliding partons from the hard interaction are accompanied by other quarks and gluons that interact and rearrange themseleves into colorless objects. A pp collision thus consists of: the Hard Scattering which represents the part of the collision where

two partons in the initial state interact by exchanging a high transverse momentum, and the Underlying Event that represent the interaction of the everything else in the collision except the partons in hard scattering. In addition to the implementing the above, i.e. physics of a pp collision that produces a bunch of final state particles, the event simulation also has to include interactions of these particles with the CMS detector. Monte Carlo methods, that use generation of random numbers to simulate sampling from a given probability distribution , are used to model the above event simulations [5].

4.2.1 Monte Carlo method

Monte Carlo (MC) methods (named after a famous casino in the city state of Monaco) are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results [6]. In particle physics, these methods play a key role in generation of events and are used primary for : generation of samples from specified probabilty distributions, and the calculation of integrals. Programs which implement the above method, called MC event generators, use generation of random numbers to make decisions about physics processes. These can range from selection of processes are generated in the collision, to which decay channel a particle decays in, to making decisions on how the particle interacts with detector material. Usually, each such decision is the result of a draw from a distribution which depends only on the current state the process is in, and not on previous states. The MC generator is provided as input the distributions that represent the physics of the generated particles, their production, their decay modes and their couplings. A MC generator starts by using a pseudo-random number generator that usually outputs a random number between 0 and 1 with. Although, true random number generation can only be done by physical processess, modern pseudo-random number generators are known to generate numbers with a high degree of randommness. Starting from

this distribution, the MC event generator uses one of the various methods such as the inverse-transform method, or the rejection sampling method to convert this uniform distribution into a desired probablity distribution, $p(x)$. It is then possible to generate random numbers according to this distribution to simulate physical processes.

4.2.2 CMS simulation pipeline

The MC simulation of events in CMS consists of the following sequential steps. The first step is simulation of the Hard Scattering.As mentioned earlier, this represents the primary hard interaction in a collision where two partons in the initial state interact by exchanging high transverse momentum resulting in a final state with two or more partons. The parton density function (pdf) which parametrizes the distributions of the partons inside each hadron are used to model the momenta of incoming partons. It represents the probability of finding a parton of a certain flavour at a certain longitudinal momentum fraction, when the hadron, that contains it, is probed at a certain scale. The PDF are extracted from fits to the data, mainly from ep collisions, and various PDF sets are available for each parton flavour. Commonly used pdf sets include ones provided by the CTEQ, HERA (H1 and ZEUS) and NNPDF collaborations. The LHAPDF library provides a unified C++ interface to all major PDF sets. The matrix element formulation is used to model the hard scattering process to leading order in perturbative QCD, or to higher orders depending on the generator. The next step is simulation of the parton shower. The hadronization and radiation of quarks and gluons in the initial and final states cannot be feasibly encapsulated in the matrix element computation. Parton shower describes these missing parts. The matrix element calculations are combined with the parton shower by one of the different matching schemes which ensure that there is no double counting of terms present in both the matrix element and the partion shower expansion. The matching schemes that are most often used are MLM [7], CKKW [8] and FxFx [9]. The simu-

lation of the Underlying Event comes next. Underlying event includes everything in the collision that is not associated with the primary hard scattering process. They consist mostly of soft QCD interactions, and implemented using the MC event generators and interfaced with the matrix element simulation. The hadronization of the quarks and gluons is simulated next and it conisists of recombination of individual partons into colorless hadrons. Lastly the decay of short-lived particles is simulated.

An important part of the event generation chain is the simulation of pileup. The protons circulate inside the LHC not as a continuous beam but in discrete closely packed bunches. This leads to more than one proton-proton collision per bunch crossing, i.e. pileup both in-time and out-of-time (see chapter 3). Event generators add pile-up events to the hard scattering samples by randomly simulating soft inelastic collisions and overlapping them. The distribution of the number of pileup interactions in data is hard to predict. MC event generators usually produce events for a scenario with a higher number of pileup vertices, and with a flat disctribution of number of vertices . This is afterwards reweighted to match the observed distribution of pileup interactions in data.

Several MC generators have been developed. Some of these can produce all components of the above simulation pipeline while some calculate only the matrix element and need to be interfaced with other generators for the simulation of remaining parts. Pythia [10] and Herwig [11] can produce the entire chain while Powheg [12–17], aMC@NLO [18] and Madgraph [19] produce up to matrix element stage. Powheg and aMC@NLO can perform next-to-leading order calculations.

Finally, the Geant4 (GEometry ANd Tracking) [20] package is used to simulate the interaction of physical particles after the collision, produced by pipeline described above, with a sophisticated and complex simulation of the detector itself. This simulated detector response is used as input for the same physics reconstruction algorithms (desribed in the next section), that are used to reconstruct the data, thus enabling

a direct comparison of the two. If differences are observed in the behavior of these reconstruction algorithms for MC events in comparsion to observed data, the MC events are tuned to the behavior observed in data.

4.3 MC samples used for the analyses

The ggH and VBF Higgs boson samples are generated with POWHEG 2.0 while an extension of POWHEG 2.0 [21] is used for the WH and ZH simulated samples. For the $H \rightarrow \mu\tau_e$ analysis, only the gluon fusion (ggH) production mode has been considered. Samples are generated for a range of H masses from 200 to 900 GeV.

The Z + jets and W + jets processes are simulated using the MG5_aMC@NLO generator at leading order (LO) with the MLM jet matching and merging scheme. The same generator is also used for diboson production which is simulated at next-to-LO (NLO) with the FxFx jet matching and merging scheme. POWHEG 2.0 and 1.0 are used for top quark-antiquark ($t\bar{t}$) and single top quark production, respectively. The POWHEG and MADGRAPH generators are interfaced with PYTHIA 8 for parton showering, fragmentation, and decays.

As mentioned earlier in this chapter, additional pileup interactions are also a part of the MC generation pipeline. All simulated samples are reweighted to the pileup distribution observed in data. An event weight is applied based on the number of simulated pileup events and the instantaneous luminosity per bunch-crossing, averaged over the run period. Several other scale factors are used to reweight the events in order to get the MC simulation to match the data closely. These include scale factors based on trigger, lepton identification, lepton isolaton and b-jet tagging efficiencies.

4.4 Physics Object Reconstruction

4.4.1 Particle Flow

4.4.2 Track Reconstruction

4.4.3 Electron Reconstruction

4.4.4 Muon Reconstruction

4.4.5 Jet Reconstruction

4.4.6 MET, MT and Collinear Mass

4.4.7 Tau Lepton and others

4.5 Datasets

The data analysed in this search was gathered by the CMS detector in 2016 during proton-proton collisions at the LHC, corresponding to an integrated luminosity of $35.9 fb^{-1}$. This data corresponds to a center-of-mass energy of 13 TeV and a spacing of 25ns between bunch crossings in the LHC with an average of about 30 collisions per bunch crossing. The subset of samples used among all collected by CMS are the ones having at least one isolated muon having transverse energy over 24 GeV, as triggered by the CMS high level isolated muon trigger (HLT_IsoMu24 in CMS parlance).

CHAPTER 5

EVENT SELECTION

5.1 Introduction

This chapter describes in detail the event selection criteria for the analyses, and how they were chosen. It starts by introducing the backgrounds that each of selection criterion is trying to reduce in order to get a higher ratio of number of signal events to background events, leading to a better sensitivity for the search. This is followed by the procedure for arriving at the best possible set of selection criterion. For the h $\rightarrow \mu\tau_e$, two methods of selection were developed. The first method developed involves placing requirements on several kinematic variables, and then using the resulting ditribution of M_{col} as discriminant for a binned likelihood fit (see section **??** for description of statistical procedures). We call this method M_{col} fit method. The second method developed involves using a Boosted Decision Trees (BDT) discriminator for classification of signal and background events. The output distribution of the BDT discriminator is then used to perform the fit. We call this method BDT method. The BDT method is found to have greater sensitivity, as discussed later in the chapter. However, the M_{col} fit method is also presented as a complementary method and acts like a cross-check for the BDT method. For H $\rightarrow \mu\tau_e$ analysis, only the M_{col} fit method is developed. This is in part due to the dificulties foreseen in training a BDT with much fewer events available in H $\rightarrow \mu\tau_e$ analysis, and in part since this is the very first time the H $\rightarrow \mu\tau_e$ search is being performed, a simpler analysis was felt to be adequate.

Both analyses were performed blinded [22] in the signal region. All selection criterion and methods described below were developed without the knowledge of the observed data in the range of variable spectra where the signal is expected to be present. This is considered an optimal way of eliminating the unintended biasing of a result in a particular direction and is a standard methodology in particle physics analyses.

5.2 h125: $h \to \mu\tau_e$ analysis

5.2.1 $h \to \mu\tau_e$: Final state signature and backgrounds

The signature of the $h \to \mu\tau_e$ analysis final state consists of a muon that comes promptly from the Higgs and has a hard p_T spectrum, along with a softer electron of opposite sign charge that comes from the tau lepton, and missing transverse momentum from the tau decay. It is interesting to note that the signature is similar to the $h \to \tau_\mu\tau_e$ decay that is allowed by the SM and since been observed [23], but with significant kinematic differences. In $h \to \mu\tau_e$ decay the μ comes directly from the Higgs resulting in its p_T spectrum peaking and spreading out to much higher values. Also there are fewer neutrinos in $h \to \mu\tau_e$, coming from the decay of the single τ. The decay products of this highly boosted tau are closely aligned, leading to a narrow separation between the e and the $\vec{p}_T^{\,miss}$ in the azimuthal plane. The same is not true in the $h \to \tau_\mu\tau_e$ decays. These differences are illustrated pictorially in Fig. 5.1.

The most dominant backgrounds consists of $Z \to \tau\tau$ events coming from Drell-Yan production and $t\bar{t}$ production. In $Z \to \tau\tau$ events, one τ can decay to an e and the other to a μ. This background peaks at lower values of M_{col} than the signal events but there is significant overlap with the signal spectrum. In $t\bar{t}$ production, each of the top quarks can decay into a bottom and a W with the W bosons then decaying to a e and μ. The other backgrounds are smaller and include (in no particular order)

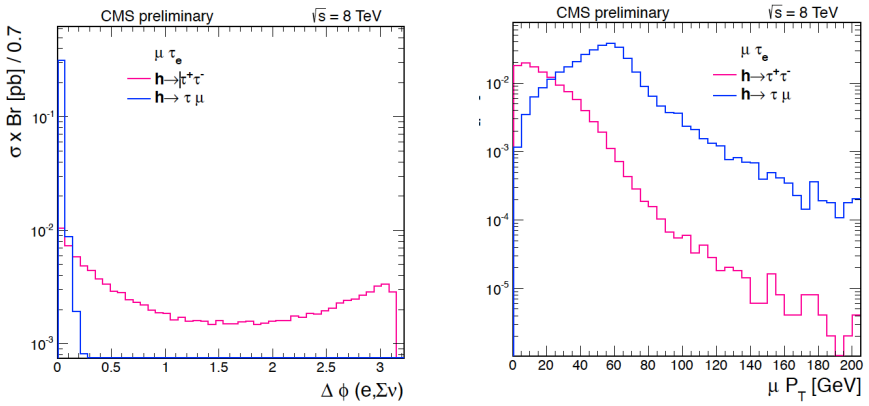


Figure 5.1: Illustration of the differences in p_T^{μ} and $\Delta\phi(e, \vec{p}_T^{miss})$ spectrums in $h \to \mu\tau_e$ and $h \to \tau_\mu\tau_e$ processes.

electroweak diboson production (WW, WZ and ZZ), h boson decays allowed by the SM ($H \to \tau\tau, WW$), $W\gamma^{(*)}$ + jets ,single top production, W + jets events, $Z \to \ell\ell$ ($\ell = e, \mu$) + jets and QCD multijet backgrounds. These backgrounds are described in more detail, along with there estimation and validation techniques in section 6.

5.2.2 $h \to \mu\tau_e$:Baseline selection and categorization

A baseline selection is defined first in order to ensure that we have clean and well-defined events faithful to the final state signature of the signal process. An isolated and well-identified μ is thus required to be present along with a well-identified and isolated e of opposite sign charge. They are required to be separated by $\Delta R > 0.3$. The identification criterion applied for μ and e have been described in sections 4.4.4 and 4.4.3. Isolation criterion, as measured by I_{rel} (described in 4.4.7), are required to have values $I_{rel}^{e} < 0.15$ and $I_{rel}^{\mu} < 0.1$. The p_T of these candidates are required to be above minimal thresholds required by trigger, identification and isolation requirement. Both candidates are also required to be within the fiducial region of the detector. The μ is required to have $p_T^{\mu} > 26\,GeV$ and $|\eta^{\mu}| < 2.4$.The e is required to have $p_T^{e} > 10\,GeV$ and $|\eta^{e}| < 2.3$. Only events with two or fewer jets are considered.

All jets considered must have $p_T > 30\,\text{GeV}$, $|\eta| < 2.4$ and satisfy the loose identification criterion described in section 4.4.5. Events with one or more jets arising from a b-quark (b-tagged jets) are vetoed. Cleaning events with b-tagged jets reduce some contribution from backgrounds which give rise to b-quarks such as $t\bar{t}$ and single top. Also, as described in 4.4.5, any event with one or more jets within $\Delta R < 0.4$ of either lepton candidates is also rejected. Further, an event is rejected if it has additional μ or e, or any τ_{had} candidates. All the above baseline selection requirements have been summarized in Table 5.1. All the events were required to pass isolated muon triggers with a p_T threshold of 24 GeV. The trigger selection has been described in detail in section 3.2.2. The distributions of the M_{col} and several other kinematic variables after the baseline selection just described, are shown in Figs. 5.2 and 5.3. These distributions act as the starting point for development of stricter kinematic selections looking at the different shapes of signal and backgrounds distributions for different variables.

Table 5.1: Baseline selection criteria for $h \to \mu\tau_e$ analysis.

Variable	μ	e
p_T	$> 30\,\text{GeV}$	$> 10\,\text{GeV}$
$\|\eta\|$	< 2.4	< 2.3
I_{rel}	< 0.15	< 0.1
Cleaning requirements		
$\Delta R(\mu, \text{e}) > 0.3$		
No additional μ, e or τ_{had}		
No b-tagged jets with $p_T > 30\,\text{GeV}$		
No jets with $\Delta R(\mu, jet) < 0.4$ and $p_T > 30\,\text{GeV}$		
No jets with $\Delta R(\text{e}, jet) < 0.4$ and $p_T > 30\,\text{GeV}$		

At this point the events are divided into several buckets, called categories. This

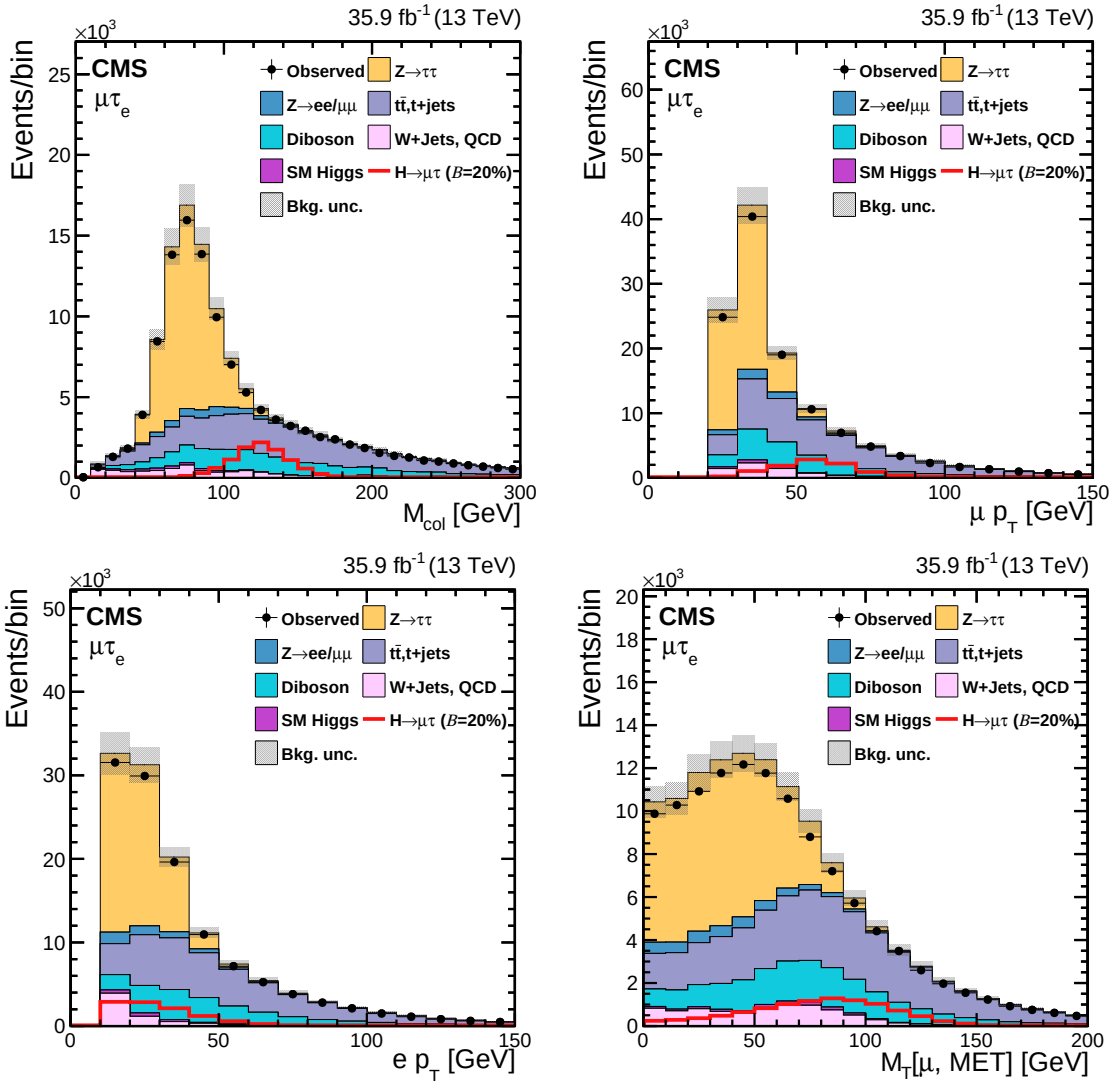


Figure 5.2: Distributions of kinematic variables after baseline selction for h $\rightarrow \mu\tau_e$ analysis (1).

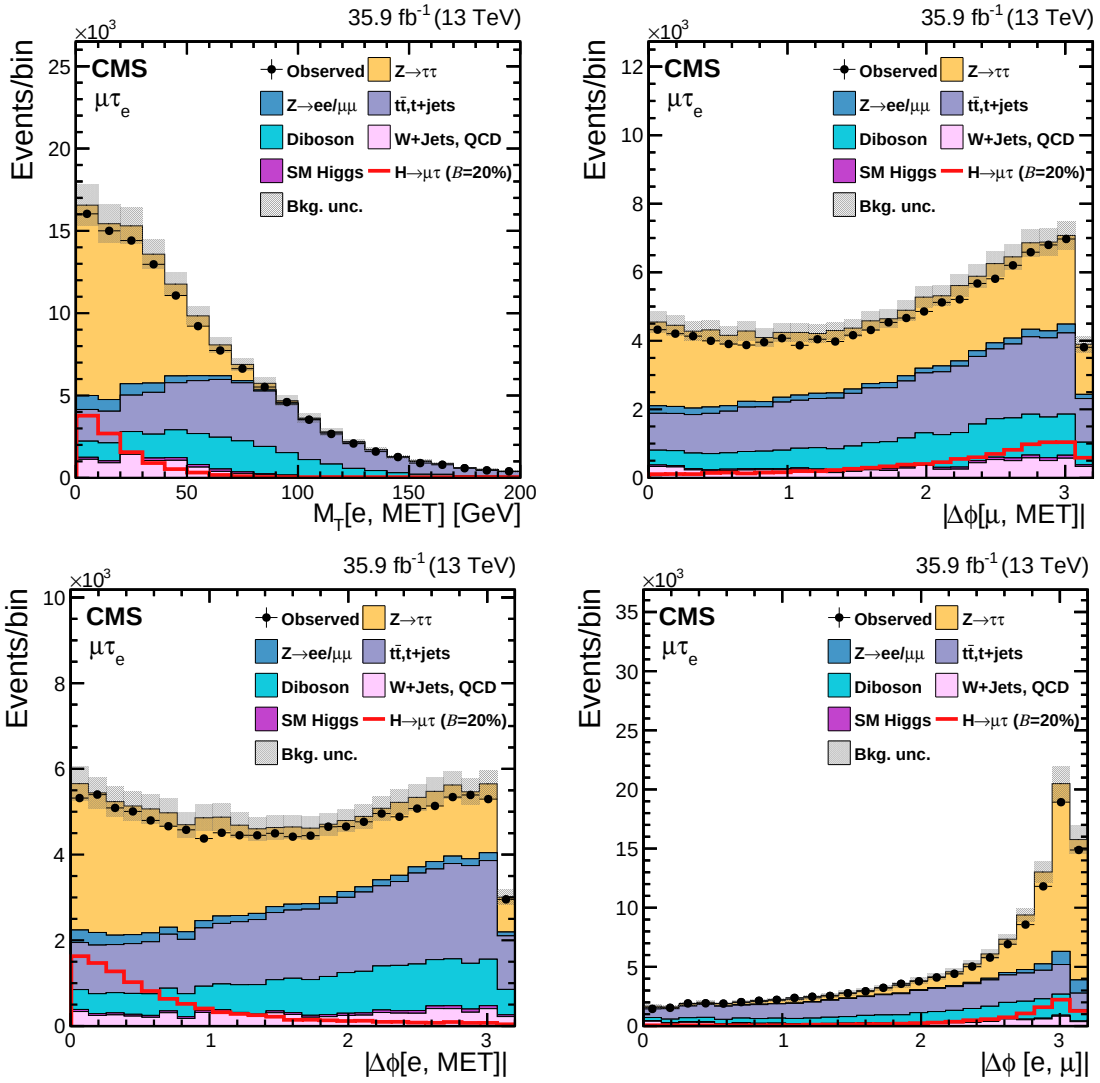


Figure 5.3: Distributions of kinematic variables after baseline selction for h $\rightarrow \mu\tau_e$ analysis (2).

is done on the basis of number of jets present in the event. In events with 2 jets the invariant mass of the di-jet system (M_{jj}) is also used for categorization. The topology of events containing different number of such jets can be different. For example, in events with one energetic jet the h produced can be boosted resulting in the azimuthal separation of the μ and e (that come from its decay) to be narrower than events with no jets. Each of this categories enchance the contribution of different h boson production mechanisms, and requiring different optimal selection criteria in each category helps increase the sensitivity of the search. The categories in order of decreasing number of signal events are:

- **0-jet category**: These are events that do not have any jet. This category enhances the gluon-gluon fusion (GGF) contribution.
- **1-jet category**: Events that have 1 jet are put in this category. This category enhances the GGF production with initial state radiation (ISR). Some VBF events where one jet has escaped detection can also enter this category.
- **2-jet GGF category**: This category contains events that have 2 jets with the additional requirement that $M_{jj} < 550\,\text{GeV}$. The dominant contribution comes from GGF production in association with two jets.
- **2-jet VBF category**: This category contains events that have 2 jets with the additional requirement that $M_{jj} \geq 550\,\text{GeV}$. The dominant contribution comes from VBF production which is characterized by presence of two jets with high dijet mass.

5.2.3 $h \to \mu\tau_e$: M_{col} fit selection

In the M_{col} fit method, the selection is performed by placing kinematic cuts on several variables to enhance the signal-to-background ratio. There are several variables considered for this and they include: the azimuthal separation ($\Delta\phi$) between μ

and e, between e and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, between μ and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, denoted respectively by $\Delta\phi(\mathrm{e},\mu)$, $\Delta\phi(\mathrm{e},\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$, $\Delta\phi(\mu,\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$, and the transverse mass between μ and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, between e and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, devnoted respectively by $M_T(\mu)$ and $M_T(\mathrm{e})$. The $\mathrm{h} \to \mu\tau_e$ decay being a 2-body decay, the μ and e are expected to be well separated in the azimuthal plane. Therefore, selecting events with a $\Delta\phi(\mathrm{e},\mu)$ larger than a threshold can help reject background events while keeping the signal that is peaked at high $\Delta\phi(\mathrm{e},\mu)$ values. This can be seen from Fig 5.3 (bottom right). Both neutrinos in the signal process come from the decay of the same τ. These neutrinos form the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$. As mentioned earlier, the τ being much lighter than the h, it is highly boosted and its decay products i.e. e and the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ are expected to be close to each other in the azimuthal direction. Thus $\Delta\phi(\mathrm{e},\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$ is expected to peak at values close to zero for signal events, as seen in Fig 5.3 (bottom left). Given that all backgrounds have relatively flat shape for this variable throughout the $\Delta\phi$ range, requiring $\Delta\phi(\mathrm{e},\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$ to be lower than a threshold works as a strong rejection criterion against the backgrounds. Following a similar line of reasoning, the μ is expected to be well separated from the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ resulting in $\Delta\phi(\mu,\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$ for signal events to peak at high values, as seen in Fig 5.3 (top right). Further, as the $M_T(\ell)$ (defined in section 4.4.6) contains negative of the cosine of $\Delta\phi(\ell,\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$ term, it is expected to be peak at values similar to $\Delta\phi(\ell,\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$. This can be seen from Fig 5.3 (top left) and Fig 5.3 (bottom right) which show signal events for $M_T(\mu)$ and $M_T(\mathrm{e})$ peak at relatively higher and lower values than most backgrounds respectively. In particular, requiring $M_T(\mu)$ to be larger than a threshold can help reject a lot of $\mathrm{Z} \to \tau\tau$ events which is the most dominant background in the 0-jet category. All the above variables have some amount of correlation with one another (see the correlation matrix shown in Fig. 5.6. The optimization procedure used to arrive at the most optimal set of kinematic thresholds for these variables is described in detail in the next paragrhaph. The thresholds on the $\mathrm{p_T}$ of the μ and e have not been made stricter to avoid biasing the selection

toward energetic leptons that sculpt the background M_{col} distribution to mimic the signal peak. This effect could potentially reduce the shape discrimination power of the signal extraction procedure. Only in the 0-jet category category the requirement on $\mathrm{p_T}$ of the μ is made marginally stricter by requiring $\mathrm{p_T^{\mu}} > 30\,\mathrm{GeV}$. All other lepton $\mathrm{p_T}$ requirements are allowed to remain the same as baseline selection and are not included in the optimization procedure.

The aim of the optimization procedure is to maximize the sensitivity of the analysis. In other words, we want to select a set of thresholds which increases a quantity such as the $\frac{S}{\sqrt{S+B}}$ ratio where S and B are the number of estimated signal and background events respectively. It is also necessary to ensure alongwith, that the entire spectrum of distribution of the discriminant variable (that is used int the final max-likelihood fit to extract results) is well-populated, escpecially in the region where the signal is expected to appear. A bad fit can potentially degrade the sensitivity of the analysis. Taking both of the above points into consideration, the thresholds have been optimized to obtain the most stringent (lowest) possible expected limits. The definition and procedure of extacting the expected limit is given in section **??**). To do the optimization of the kinematic thresholds, we start by requiring the baseline selection. Then for a variable in consideration,e.g.- $\Delta\phi(\mathrm{e}, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$, we look at the expected limit while making the threshold progressively stricter until we reach a point where making the threshold any stricter degrades (increases) the expected limit. We repeat this procedure for all variables and note the stringent expected limit for each (by tightening thresholds of only that variable). This concludes one round of the optimization. For the next round we start by requiring the baseline selection. In addition we require that the variable that achieved the best possible expected limit among all variables in the last round satisfy its corresponding threshold. Lets call this variable variable1. We now repeat the same procedure as the last round for all but variable1. Say the variable that gave us the best possible expected limit this

round is variable2. For the start of the following round variable2 is is required to satisfy its corresponding threshold. Then all the other variables (including variables that were had chosen thresholds in earlier rounds such as variable1 here) are made to go through the same procedure. This is done because the optimum value of threshold for variables chosen earlier might shift as new variables are chosen. This process is continued until the expected limit becomes no further stringent in sucessive rounds. This optimization was done separately for each of the four categories. The final set of thresholds arrived at in this way for the h $\rightarrow \mu\tau_e$ M_{col} fit analysis are listed in Table. 5.2. This method of choosing the optimal set of thresholds is sometimes called the n-1 procedure, and the idea is conceptually similar to forward/backward selection methods used in statistical learning to build optimal models.

TABLE 5.2

FINAL SELECTION CRITERIA FOR h $\rightarrow \mu\tau_e$ M_{col} FIT ANALYSIS.

Category	0-jet	1-jet	2-jet GGF	2-jet
$\text{p}_{\text{T}}^{\mu}$	$> 30\,\text{GeV}$	–	–	–
$M_T(\mu)$	$> 60\,\text{GeV}$	$> 40\,\text{GeV}$	$> 15\,\text{GeV}$	$> 15\,\text{GeV}$
$\Delta\phi(\text{e}, \vec{p}_{\text{T}}^{\text{miss}})$	< 0.7	< 0.5	< 0.3	< 0.3
$\Delta\phi(\text{e}, \mu)$	> 2.5	> 2.0	–	–

5.2.4 h $\rightarrow \mu\tau_e$: BDT method selction

In the BDT method, a boosted decision trees (BDT) classifier is used to discriminate signal events from background events. A decision tree is a classifier which works by building a tree structure based on binary splits (as shown in Fig. 5.4). Starting from the root node of the tree (which contains all the events which we want to classify), a sequence of binary splits is made using input variables provided to the classifier. At each split, the variable which provides best purity of split or equivalently, in our case the best separation of signal and background events, is used. The same variable can thus be used for splitting several nodes and the splitting is continued until a desired some stopping criterion such as depth of the tree, purity of leaf nodes , minimum number of events in a leaf node etc. is reached. All events end up in one of the leaf nodes. If an event ends up in a leaf node in which signal events form the majority fraction, it is classified as a signal event. Otherwise, it is classified as a background event. Boosting is a class of ensemble machine learning techniques which help in enhancing performance of weak classifiers by sequentially building classifiers using reweighted (boosted) versions of the training data and then taking a weighted majority vote of the sequence of classifiers thus produced. Boosting also stabilizes the response of the classifiers with respect to fluctuations in the training data. In other words it helps avoid overfitting to the training data. When the boosting technique is a applied to produce an ensemble of decision trees, the resulting ensemble of classifiers is called a Boosted Decision Trees classifier. A detailed overview of how decision trees and boosting works, and the chosen value of parameters used in training the BDTs for this analysis is given in appendix A.

The BDT is trained using events that satisfy the baseline selection criteria. Simulated GGF and VBF events weighted by their cross-section are used as signal events for training. For background, a mixture of $t\bar{t}$ and Drell-Yan events are used, also weighted by their respective cross-sections. The $t\bar{t}$ and Drell-Yan backgrounds are

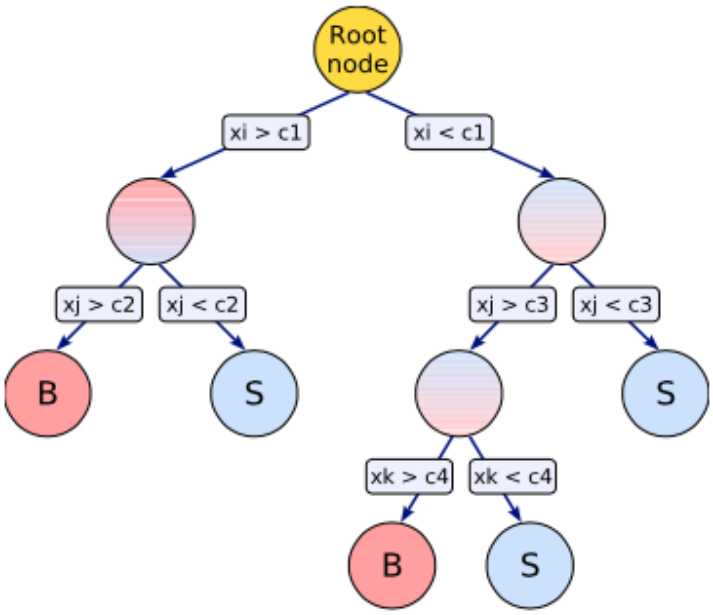


Figure 5.4: Illustration of decision tree. [24]

the most dominant backgrounds. The Drell-Yan background is the most dominant background in 0-jet and 1-jet category, while the $t\bar{t}$ background is the most dominant in both 2-jet categories. It also has many kinematic characteristics in common with diboson and single-top bankgrounds. A suite of input variables is used in training of the BDT. They are as follows:

- Transverse mass between the μ and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$: $M_T(\mu)$.
- Transverse mass between the e and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$: $M_T(\mathrm{e})$.
- Azimuthal angle between the e and μ: $\Delta\phi(\mathrm{e},\mu)$.
- Azimuthal angle between the e and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$: $\Delta\phi(\mathrm{e},\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$.
- Azimuthal angle between the μ and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$: $\Delta\phi(\mu,\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$.
- Collinear mass: M_{col}.
- Muon $\mathrm{p_T}$: $\mathrm{p}_{\mathrm{T}}^{\mu}$.

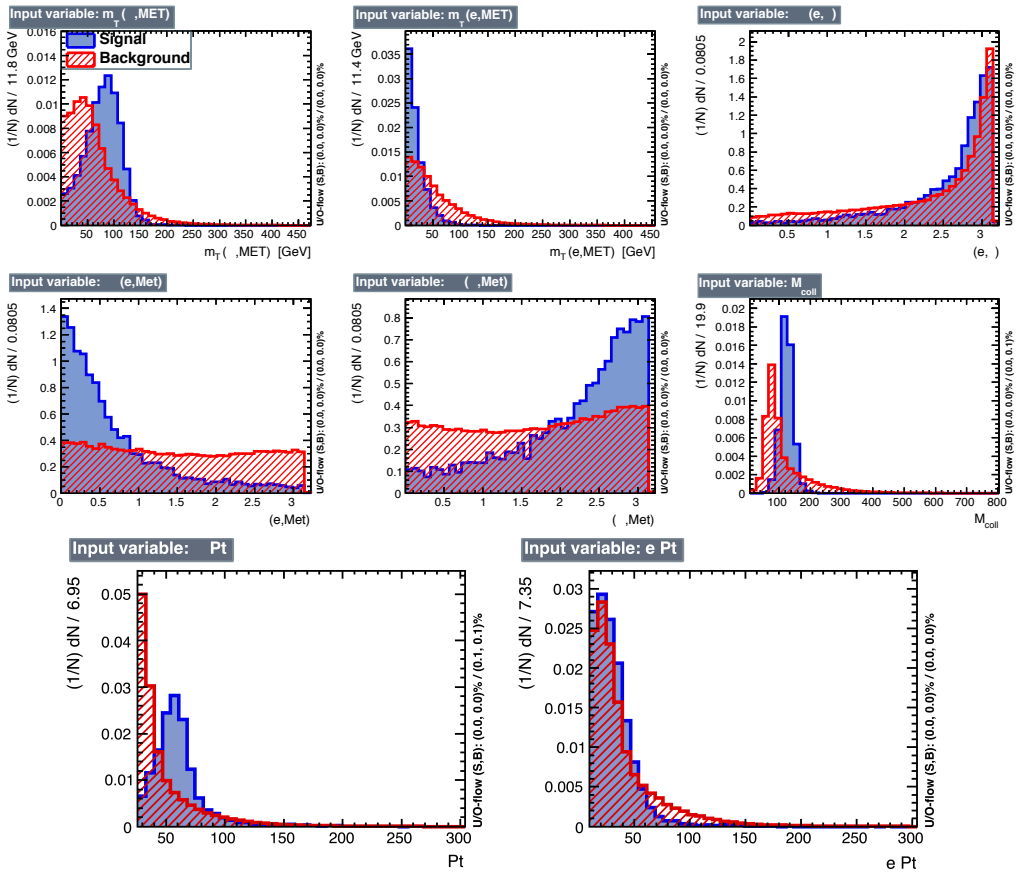


Figure 5.5: Normalized distributions of the input variables for BDT method. The signal (blue) is composed of a weighted mixture of GGF and VBF events, whereas the background (red) is made of $t\bar{t}$ and Drell-Yan events. All events were required to satisfy the baseline selection criteria.

- Electron p_T: p_T^e.

The distributions of these variables normalized to the total number of events in the input sample to the BDT is shown in Fig. 5.5. The correlations between these variables in signal and background events are shown in Fig. 5.6.

The training was done with a 800 decision tree ensemble, each tree having a maximum depth of 4. The gini-index criterion was used for splitting the data at each node. Further, AdaBoost (adaptive bossting) method was used for boosting (see appendix A for details of these techniques). A training to testing split of 70:30 split

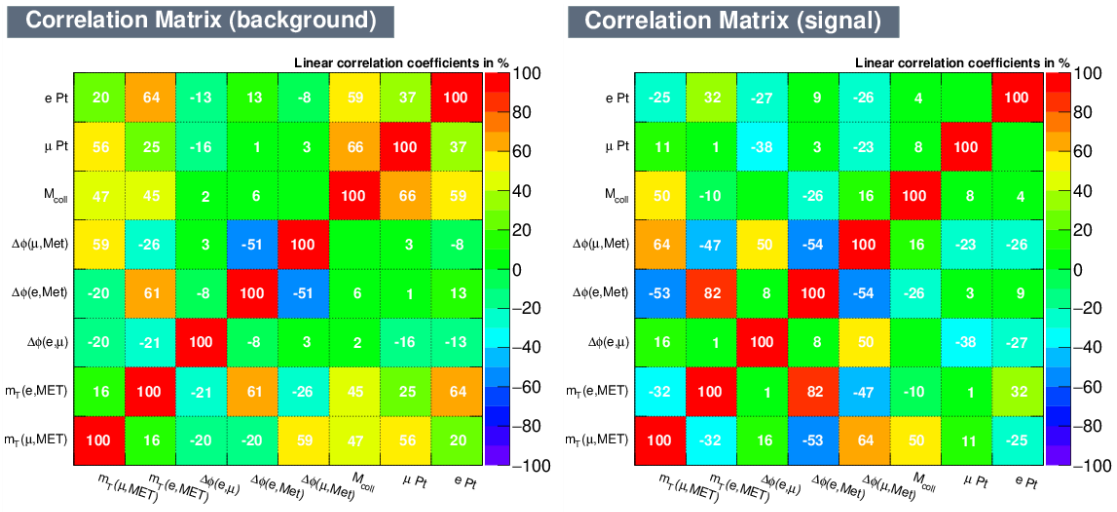


Figure 5.6: Correlations between input variables for signal events (right) and background events (left).

was used. Fig. 5.7 shows the distribution of the BDT response for training and testing samples. The training and testing distributions for both signal and background events match well, suggesting that there is no overtraining. The distribution of BDT response is used in max-likelihood fit to extract results, as discussed in section 7.3.3.

5.3 Heavy higgs:H $\to \mu\tau_e$ analysis

5.3.1 H $\to \mu\tau_e$: Final state signature and backgrounds

The signature of the H $\to \mu\tau_e$ analysis final state is very similar to that of h $\to \mu\tau_e$. It also consists of a muon that comes promptly from the Higgs and has a hard p_T spectrum, along with a softer electron that comes from the tau lepton, and missing transverse momentum from the tau decay. The p_T^{μ} spectrum is expected to be harder for higher H boson masses. The topologies being similar, the kinematic properties discussed in section 5.2.1 for h $\to \mu\tau_e$ analysis also apply to the H $\to \mu\tau_e$ analysis. The H boson mass peaks for all the simulated samples illustrated in Fig 5.8.

The most dominant backgrounds for H $\to \mu\tau_e$ consists of events from $t\bar{t}$ and

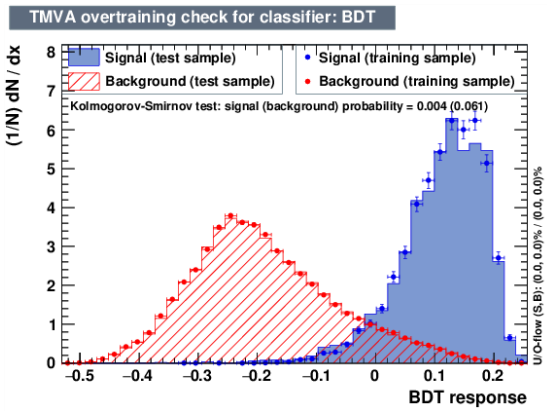


Figure 5.7: Distribution of BDT response for training (dots) and test(fill) distributions for both signal(blue) and background(red) events.

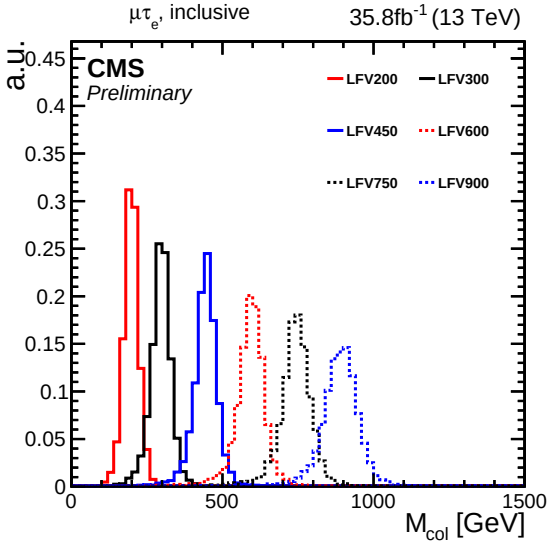


Figure 5.8: Illustration of simulated signal mass peaks for H $\rightarrow \mu\tau_e$ analysis for different H boson masses.

electroweak diboson production. Unlike $h \to \mu\tau_e$ analysis, $Z \to \tau\tau$ events from Drell-Yan production form a very small background as the $Z \to \tau\tau$ spectrum peaks at much lower values (around Z boson mass) of collinear mass that the signal events coming from heavy H boson decays. The other backgrounds come from h boson decays ($H \to \tau\tau, WW$), $W\gamma^{(*)}$ + jets ,single top production, W + jets events, $Z \to \ell\ell$ ($\ell = e, \mu$) + jets and QCD multijet backgrounds. These backgrounds are described in more detail, along with there estimation and validation techniques in section 6.

5.3.2 $H \to \mu\tau_e$: Baseline selection and categorization

The baseline selection for $H \to \mu\tau_e$ is similar to that of $h \to \mu\tau_e$ with the exception of higher p_T thresholds. Just like $h \to \mu\tau_e$, an isolated and well-identified μ is thus required to be present along with an well-identified and isolated e of opposite sign charge. They are required to be separated by $\Delta R > 0.3$. The identification and isolation criteria have been described in sections 4.4.4, 4.4.3 and 4.4.7. All events are required to pass a single muon trigger with the threshold of 50 GeV. The trigger selection has been described in detail in section 3.2.2. The μ is required to have $p_T^\mu > 53\,\text{GeV}$ and $|\eta^\mu| < 2.4$. The e is required to have $p_T^e > 10\,\text{GeV}$ and $|\eta^e| < 2.3$. Only events with zero or one jet are considered. Jets must have $p_T > 30\,\text{GeV}$, $|\eta| < 2.4$ and satisfy the loose identification criterion described in section 4.4.5 to be considered. As only GGF production mode is considered for the $H \to \mu\tau_e$ analysis, events with more than one jet make negligent contribution and are rejected. All other other criteria are same as the $h \to \mu\tau_e$ analysis. The entire set of baseline selection criteria for $H \to \mu\tau_e$ has been summarized in table 5.3.

The events are then divided into categories, with motivations similar to the $h \to \mu\tau_e$ analysis (see section 5.2.2), on the basis of number of jets present in the event. The two categories for $H \to \mu\tau_e$ are:

- **0-jet category**: These are events that do not have any jet. This category

Table 5.3: Baseline selection criteria for H $\to \mu\tau_e$ analysis.

Variable	μ	e
p_T	$> 53\,\text{GeV}$	$> 10\,\text{GeV}$
$\lvert\eta\rvert$	< 2.4	< 2.3
I_{rel}	< 0.15	< 0.1
Cleaning requirements		
$\Delta R(\mu, \text{e}) > 0.3$		
No additional μ, e or τ_{had}		
No b-tagged jets with $p_T > 30\,\text{GeV}$		
No jets with $\Delta R(\mu, jet) < 0.4$ and $p_T > 30\,\text{GeV}$		
No jets with $\Delta R(\text{e}, jet) < 0.4$ and $p_T > 30\,\text{GeV}$		

enhances the gluon-gluon fusion (GGF) contribution.

- **1-jet category**: Events that have 1 jet are put in this category. This category enhances the GGF production with initial state radiation (ISR).

The distributions of several kinematic variables after the baseline selection and categorization are shown in Figs. 5.9 and 5.10.

5.3.3 H $\to \mu\tau_e$: mcol fit selection

Just like the M_{col} fit method in h $\to \mu\tau_e$, the selection is performed by placing kinematic cuts on several variables to enhance the signal-to-background ratio. The variables considered are: $\Delta\phi(\text{e}, \mu)$, $\Delta\phi(\text{e}, \vec{p}_T^{\text{miss}})$, $\Delta\phi(\mu, \vec{p}_T^{\text{miss}})$, $M_T(\mu)$ and $M_T(\text{e})$. In addition, the p_T of the μ and e are also considered. Since we are looking for a decay in an extended mass range (200-900 GeV) in H $\to \mu\tau_e$, and not in a particular region like the h $\to \mu\tau_e$ analysis, the potential effect of background mimicking the signal, in particular due to higher p_T thresholds of the leptons, is not apparent. The motivations for using these variables remain much the same like the h $\to \mu\tau_e$ analysis owing to similarities in topology. They are motivated by the facts that the only source of MET is the τ, and the τ being lighter than the H, its visible products are closely aligned, and the p_T spectrum of the prompt lepton (μ) is hard.

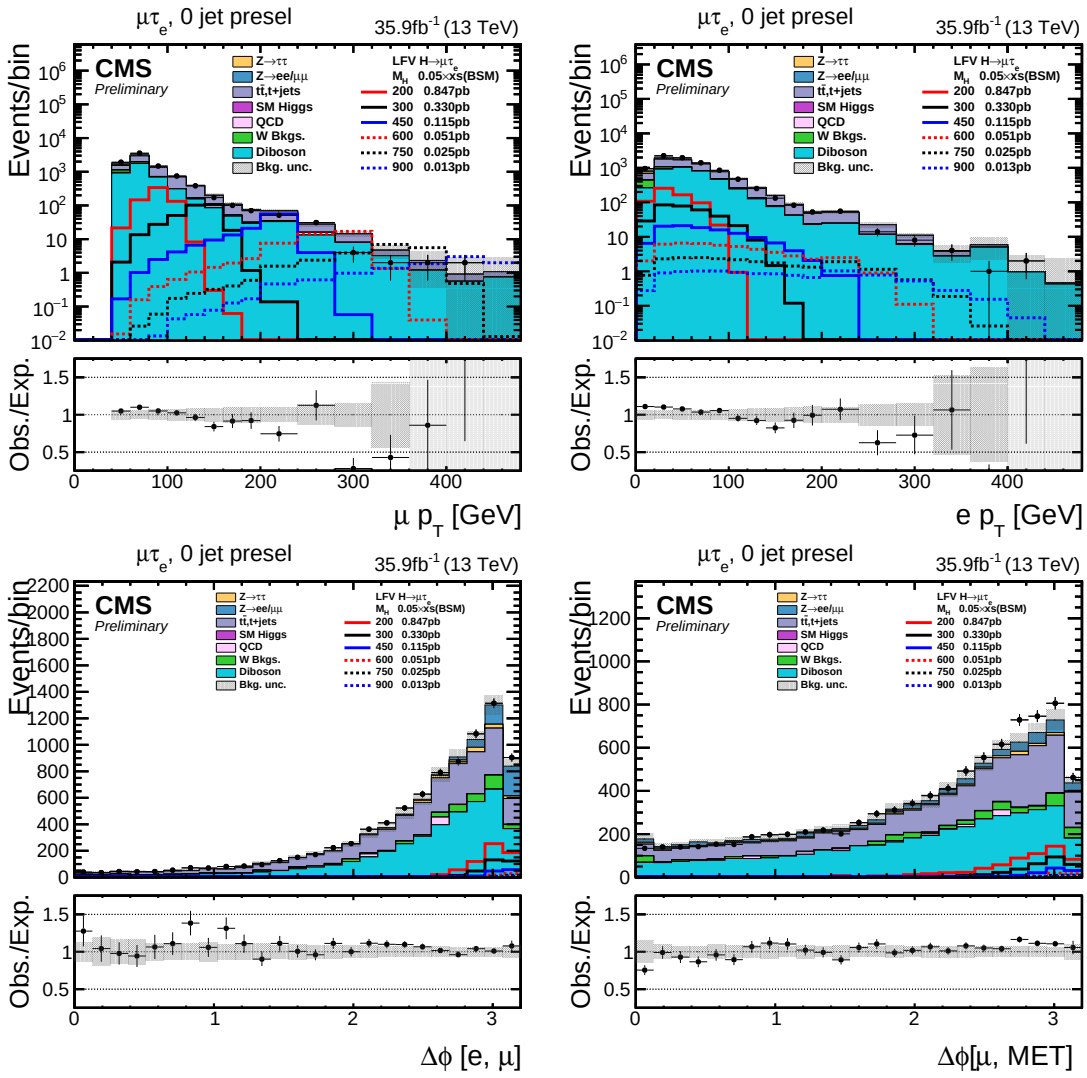


Figure 5.9: Distributions of kinematic variables after baseline selction for 0-jet category of H $\rightarrow \mu\tau_e$ analysis.

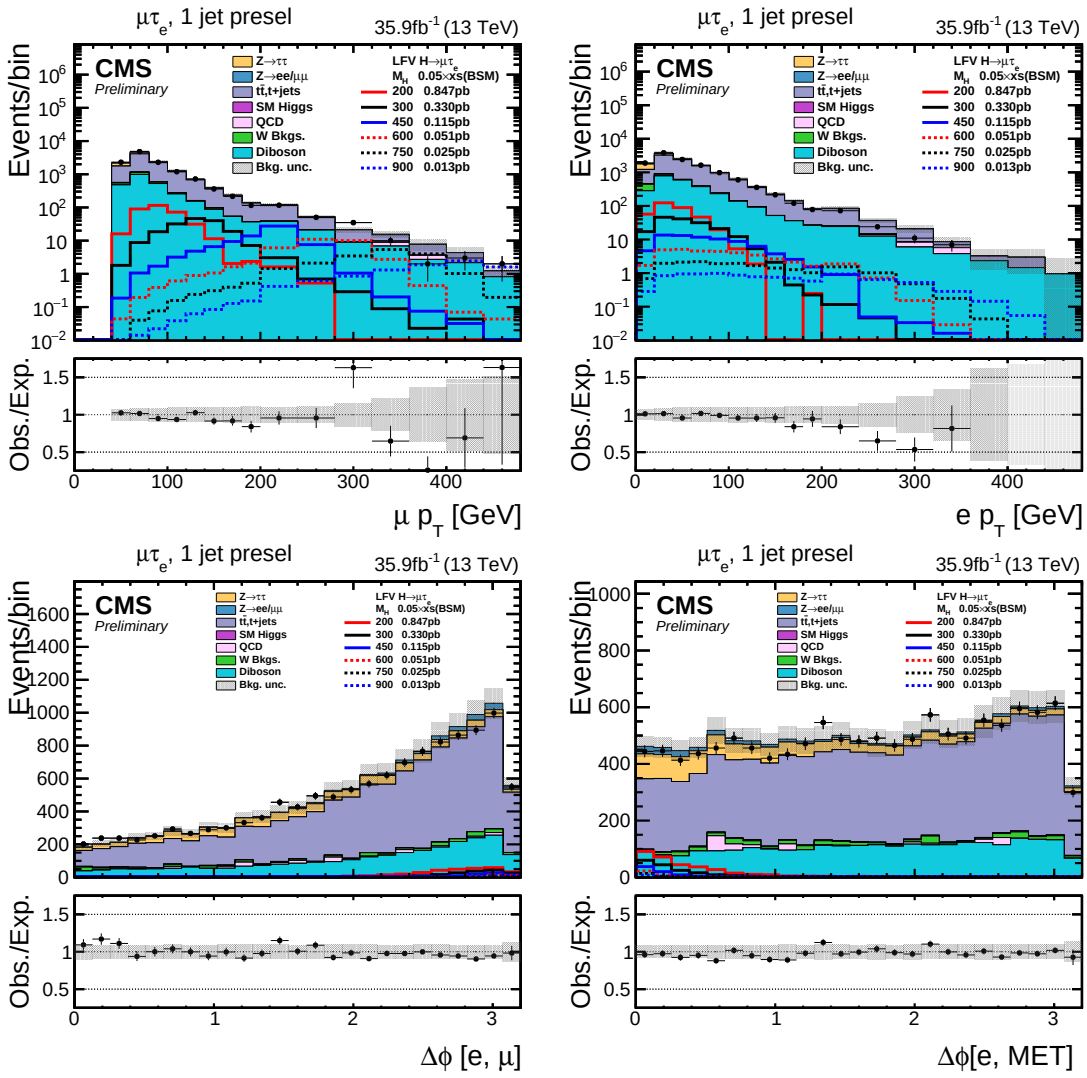


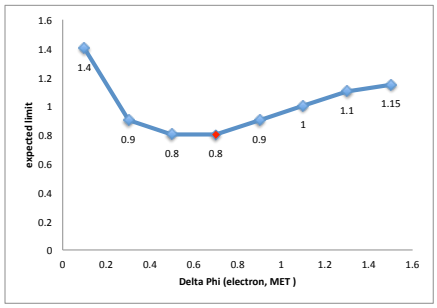
Figure 5.10: Distributions of kinematic variables after baseline selction for 1-jet category of H $\rightarrow \mu\tau_e$ analysis.

The procedure for optimization of the thresholds of for these variables is exactly the same as described in section 5.2.3. Further to get better sensitivity in the entire mass range from 200 to 900 GeV, two separate sets of thresholds are optimized, for each category. One set is optimized to provide better sensitivity in the 200-450 GeV mass range. The simulated signal for the H mass of 200 GeV is used when calculating expected limits during the optimization procedure for this mass range. The other set is optimized to provide better sensitivity in 450-900 GeV mass range. The simulated signal for H mass of 450 GeV is used when calculating expected limits during the optimization procedure for this mass range. A few illustrations of the optimization procedure are shown in Fig. 5.11. The final set of thresholds arrived at in this manner, for both mass ranges and both categories of the $H \to \mu\tau_e$ $M_{\rm col}$ fit analysis, are listed in Table. 5.4. The $M_{\rm col}$ distributions after requiring these selections is used in a max-likelihood fit to extract results, as discussed in section 7.3.3.

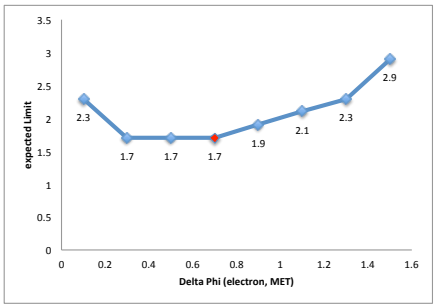
TABLE 5.4

FINAL SELECTION CRITERIA IN EACH CATEGORY OF THE $H \to \mu\tau_e$ ANALYSIS.

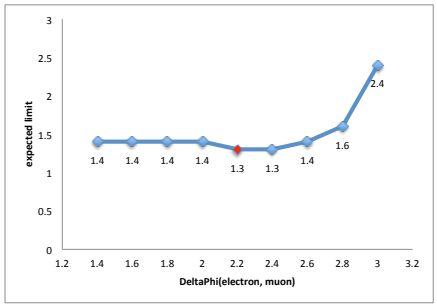
	Low mass range	High mass range
0-jet	$p_T^\mu > 60\,\text{GeV}$, $p_T^e > 10\,\text{GeV}$ $\Delta\phi(e, \vec{p}_T^{\,\text{miss}}) < 0.7$ $\Delta\phi(e, \mu) > 2.2$	$p_T^\mu > 150\,\text{GeV}$, $p_T^e > 10\,\text{GeV}$ $\Delta\phi(e, \vec{p}_T^{\,\text{miss}}) < 0.3$ $\Delta\phi(e, \mu) > 2.2$
1-jet	$p_T^\mu > 60\,\text{GeV}$, $p_T^e > 10\,\text{GeV}$ $\Delta\phi(e, \vec{p}_T^{\,\text{miss}}) < 0.7$ $\Delta\phi(e, \mu) > 2.2$	$p_T^\mu > 150\,\text{GeV}$, $p_T^e > 10\,\text{GeV}$ $\Delta\phi(e, \vec{p}_T^{\,\text{miss}}) < 0.3$ $\Delta\phi(e, \mu) > 2.2$



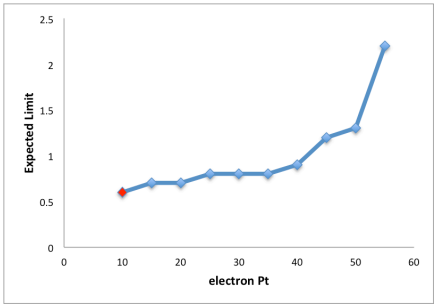
(a) Low mass range 0 jet $\Delta\phi(\mathrm{e}, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$



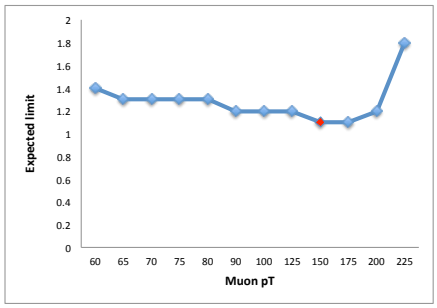
(b) Low mass range 1 jet $\Delta\phi(\mathrm{e}, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$



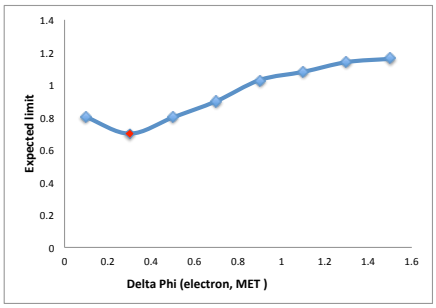
(c) Low mass range 0 jet $\Delta\phi(\mathrm{e}, \mu)$



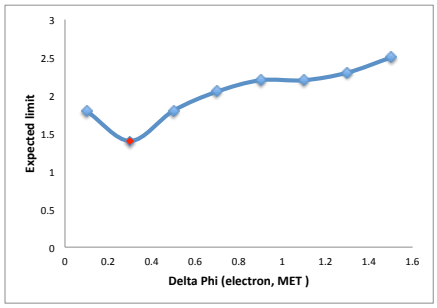
(d) Low mass range 0 jet $\mathrm{p}_{\mathrm{T}}^{\mathrm{e}}$



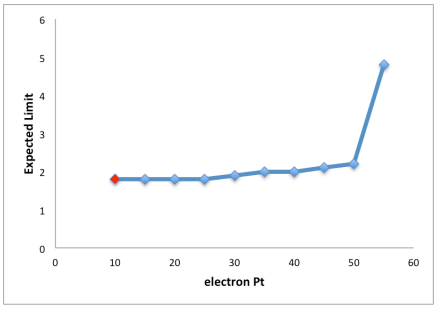
(e) High mass range 1 jet $\mathrm{p}_{\mathrm{T}}^{\mu}$



(f) High mass range 0 jet $\Delta\phi(\mathrm{e}, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$



(g) High mass range 1 jet $\Delta\phi(\mathrm{e}, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$



(h) High mass range 1 jet $\mathrm{p}_{\mathrm{T}}^{\mathrm{e}}$

Figure 5.11. Examples of cut optimisation for the $\mathrm{H} \to \mu\tau_e$ analysis

CHAPTER 6

BACKGROUND ESTIMATION AND VALIDATION

6.1 Introduction

This chapter describes the techniques used for estimation of the backgrounds in the analyses. Each background is estimated individually. For large backgrounds, the estimation is validated using regions enriched in those backgrounds.

6.2 h125: $h \to \mu\tau_e$ backgrounds

6.2.1 $Z \to \tau\tau$

The $Z \to \tau\tau$ background is the dominant background in 0-jet and 1-jet categories of the analysis. It is an irreducible background and arises when one τ coming from the Z boson decay further decays into a μ and the other decays into a e. This background is estimated from simulated monte-carlo events. In a $Z \to \ell\ell$ events from Drell-Yan production including $Z \to \tau\tau$, the $m_{\ell\ell}$ and Z_{p_T} distributions are found to be different in data and simulation. In order to correct for this, a set of reweighting factors is calculated using a dedicated control region enriched in $Z \to \mu\mu$ events. The set of reweighting factors are applied as a function of generator-level $m_{\ell\ell}$ and Z_{p_T} in the signal region of the analysis. A more detailed study of this effect and calculation of the reweighting factors can be found in the following references [23].

To validate this estimation, we look at agreement between observed data and simulation in a region enriched in $Z \to \tau\tau$ events. This region is constructed by requiring, in addition to the baseline selection, the p_T of the $\mu < 40$ GeV. The p_T in

$Z \to \tau\tau$ events is on softer side of the spectrum compared to other backgrounds which are more spread out, as seen in Fig. 5.2 (top right). The $M_T(\mu)$, as seen in Fig. 5.2 (bottom right), is required to be less than 60 GeV following similar reasoning. Further the invariant mass of the e and μ is required to be in between 30 GeV and 70 GeV in order to isolate the Z peak. The distributions of BDT response and M_{col} in this $Z \to \tau\tau$ enriched region are shown in Fig. 6.1, for the categories where this background is dominant. The plots show good agreement between data and background.

6.2.2 $t\bar{t}$

Tops decay into W bosons and a b-quark more than 90% of the time. The W boson can decay leptonically into a μ and e making it a background for the analysis. The b-tagging veto applied at the baseline selection level is able to somewhat supress this background. However it still forms a large fraction of the background for the analysis. In fact, it is the largest background in both 2-jet categories. It is also large in the 1-jet category. We estimate the $t\bar{t}$ background using simulation. The background estimation is validated in two separate control regions enriched in $t\bar{t}$. The first control region is formed requiring the baseline selection but with a inverted b-tagging veto. In other words, at least 1 b-tagged jet is required to be present in the event. The distributions of BDT response (top) and M_{col} (bottom) in this region are shown in Fig. 6.6 for categories where the $t\bar{t}$ background is large. The second control region is constructed using kinematic selection criteria. In particular, in addition to the baseline selection criteria with the b-tag veto removed, we require M_T(e) (see Fig. 5.3 top left) to be greater than 50 GeV. The distributions of BDT response (top) and M_{col} (bottom) in this second control region are shown in Fig. 6.3. Given that the uncertainty bands in these control region plots only contain uncertainties on normalization (and not shape-based uncertainties, as discussed in section 7.3.3, and included in the max likelihood fit used to extract results), the data over background

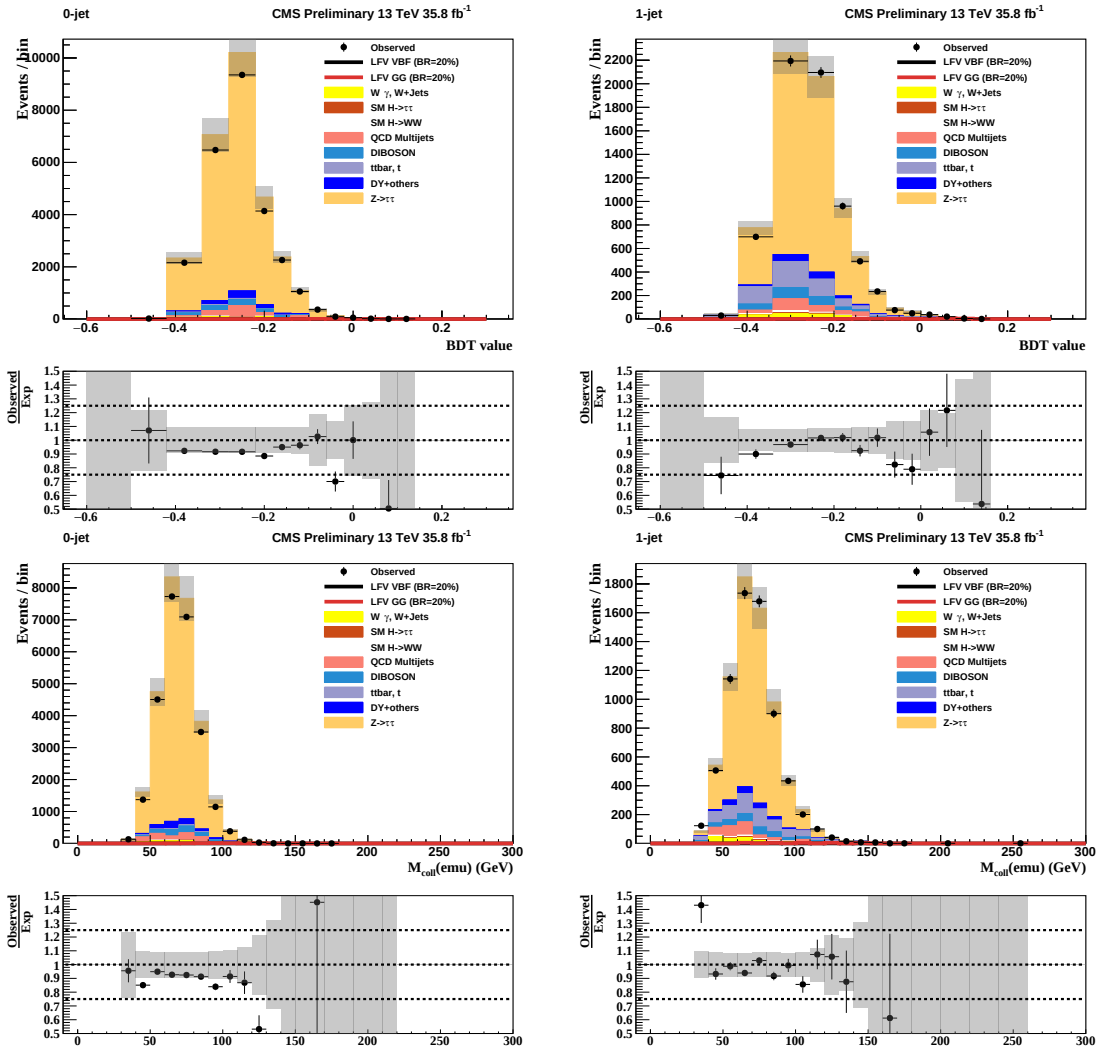


Figure 6.1: Distributions of BDT response (top) an M_{col} (bottom) in Z $\rightarrow \tau\tau$ enriched region for 0-jet (left) and 1-jet (right) categories.

estimation ratio is reasonable in these regions. Further, a normalization uncertainty of 10% is applied on the $t\bar{t}$ estimation in the signal region based on these control regions.

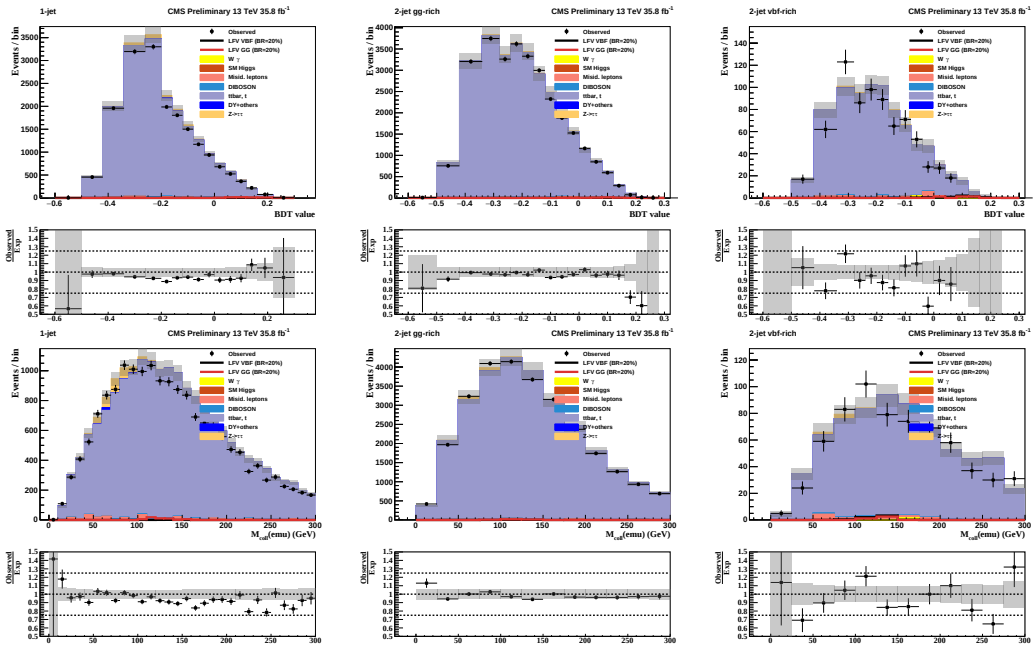


Figure 6.2. Distributions of BDT response (top) an M_{col} (bottom) in the first $t\bar{t}$ enriched region, as described in the text.

6.2.3 Misidentified lepton background

Another source of background which is relatively much smaller than $t\bar{t}$ or $\text{Z} \to \tau\tau$ arises from jets misidentified as leptons in W+jets or SM events comprised uniquely of jets produced through the strong interaction, referred to as quantum chromodynamics (QCD) multijet events. In W+jets events, one lepton candidate is a real lepton from the W boson decay while the other lepton is a misidentified jet. In QCD events,

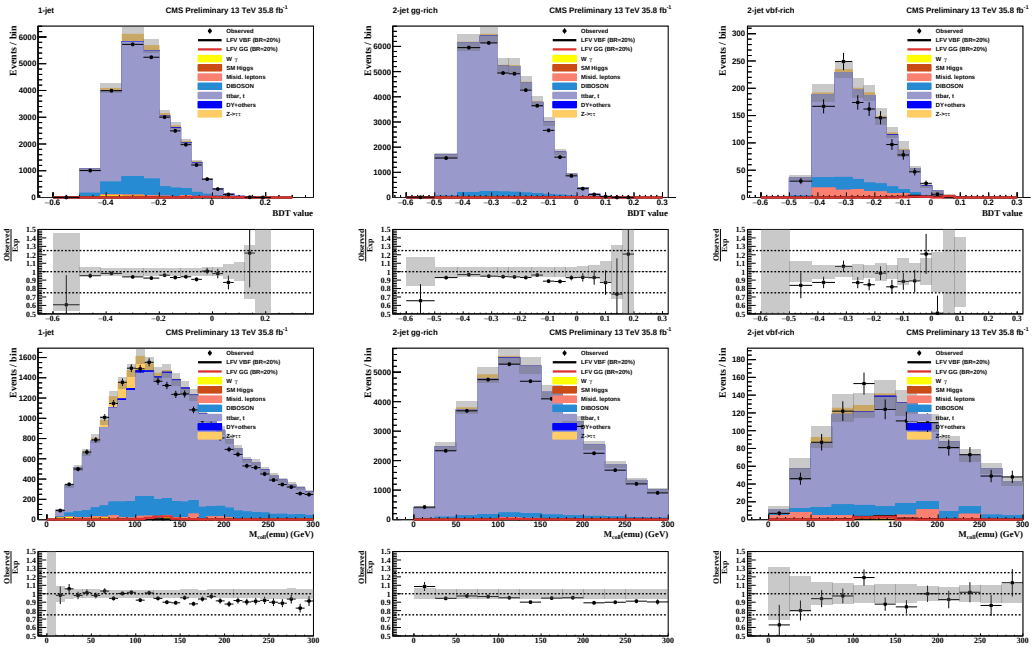


Figure 6.3. Distributions of BDT response (top) an M_{col} (bottom) in the second $t\bar{t}$ enriched region, as described in the text.

both leptons in the final state are misidentified jets. The baseline selection criteria requires the leptons to be well identified and isolated. This makes it difficult for a jet to masquerade as a lepton. In case of the μ, this is even more so since it is required to satisfy high p_T thresholds as well. Consequently, these events form a small part of the background. This is in contrast to a final state where the non-prompt lepton is a hadronically decaying τ instead of an electronically decaying one. This background would be much larger in such a case.

The W + jets background contribution to the misidentified-lepton background is estimated using simulation. The QCD multijet contribution is estimated from collision data events where the leptons have like-sign charge. The expected yield from non-QCD processes in this region is subtracted using simulation. The resulting sample is then rescaled to account for the differences between the composition in the like- and opposite-sign charge regions. The scaling factors are extracted from

samples enriched QCD multijet events, and the procedure is illustrated in Ref. []. This background is validated in a control region that is obtained by requiring the baseline selection but inverting the isolation criteria. In other words events with well-isolated μ and e are rejected. The particular isolation thresholds required for this region are:$0.1 < I^{\mathrm{e}}_{\mathrm{rel}} < 1$ or $0.15 < I^{\mu}_{\mathrm{rel}} < 0.25$. The distributions of BDT response and M_{col} in this qcd enriched region are shown in Fig. 6.4. The plots show good agreement between data and background.

6.2.4 Other backgrounds

The other backgrounds in the analysis make relatively much smaller contrbutions. Electroweak diboson production (WW, WZ and ZZ) contributes a similar number of events as the misidentified lepton background, and is estimated from simulation. WW events make the largest contribution, followed by WZ and ZZ events. This is because WZ and ZZ events have additional leptons in their final state which have to miss detection in order for the event to be a background. SM decays of the h boson also forms a small but non-negligible background. These come particularly from $\mathrm{h} \to \tau\tau$ and $\mathrm{h} \to \mathrm{WW}$ decays. Other backgrounds include $Z \to \ell\ell$ $(\ell = \mathrm{e}, \mu)$ + jets, single-top quark production and $\mathrm{W}\gamma^{(*)}$ + jets. All of these are estimated using simulation.

6.3 Heavy Higgs: $\mathrm{H} \to \mu\tau_e$ backgrounds

The background processes in the $\mathrm{H} \to \mu\tau_e$ analysis are similar to $\mathrm{h} \to \mu\tau_e$ but differ in relative contribution, and are overall much smaller. This is due to the fact that the $\mathrm{H} \to \mu\tau_e$ analyses searches for LFV decay in a higher mass, higher $\mathrm{p_T}$region. In particular, $\mathrm{Z} \to \tau\tau$ background which is the most dominant in $\mathrm{h} \to \mu\tau_e$ is now very small. The $\mathrm{Z} \to \tau\tau$ background peaks around the Z boson mass, and the high $\mathrm{p_T}$ cuts in this anlysis reject most of these events. The dominant backgrounds in $\mathrm{H} \to \mu\tau_e$ are $t\bar{t}$ production, followed by electroweak diboson production which have

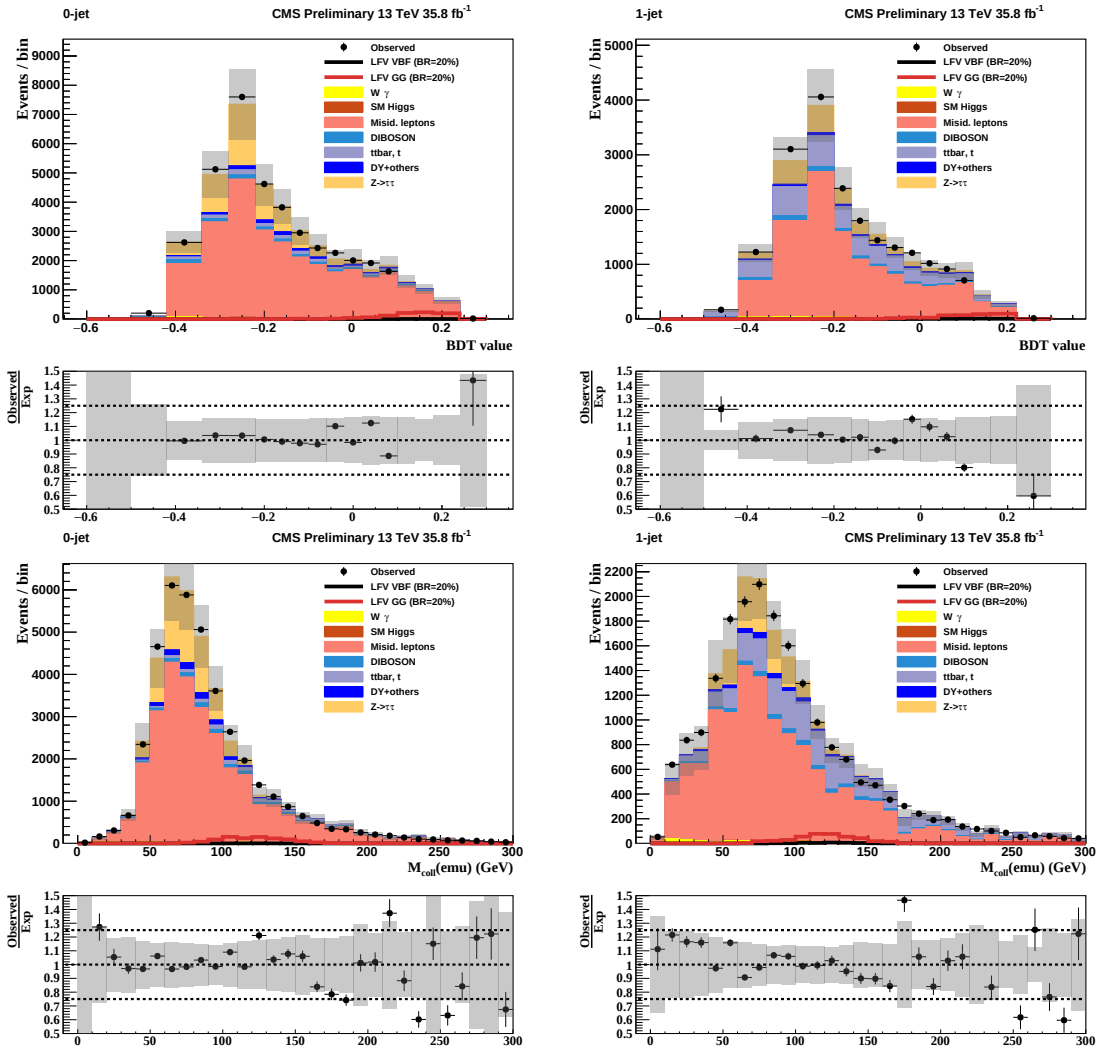


Figure 6.4: Distributions of BDT response (top) an M_{col} (bottom) in QCD enriched region for 0-jet (left) and 1-jet (right) categories.

a relatively flatter p_T distribution and survive the strict p_T requirements.

$t\bar{t}$ production is the largest background in the H $\to \mu\tau_e$ analysis. We estimate this background using simulation. A control region enriched in $t\bar{t}$ events is constructed by requiring the baseline selection with the b-tag veto removed, and with the additional requirement that at least 1 b-tagged jet be present. Fig. 6.5 (left) shows the M_{col} distribution of this sample. To take into account the residual data to background estimation difference, an overall normalization scale factor of 0.886 is extracted from this region, and is applied to the background estimation in the signal region. The same control region above is shown in Fig. 6.5 (right), after the background has been scaled by the above factor for illustration. Distributions of several other kinematic variables (after the above rescaling) in the $t\bar{t}$ control region are shown in Fig. 6.6. They show reasonable agreement between data and estimated background.

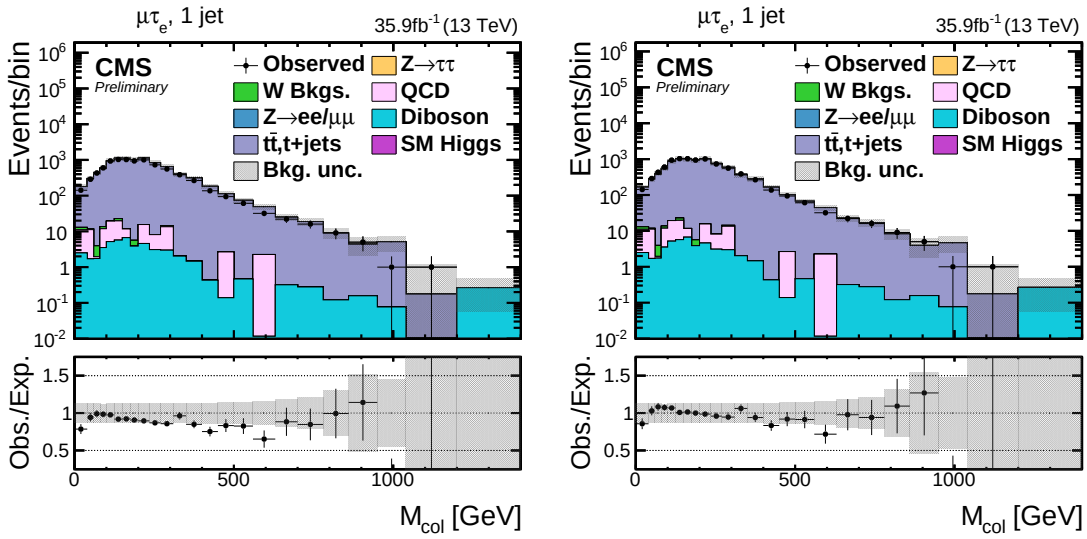


Figure 6.5: M_{col} distribution in $t\bar{t}$ enriched control region as defined in the text before the application of the scale factor (left) and after (right),for the H $\to \mu\tau_e$ analysis.

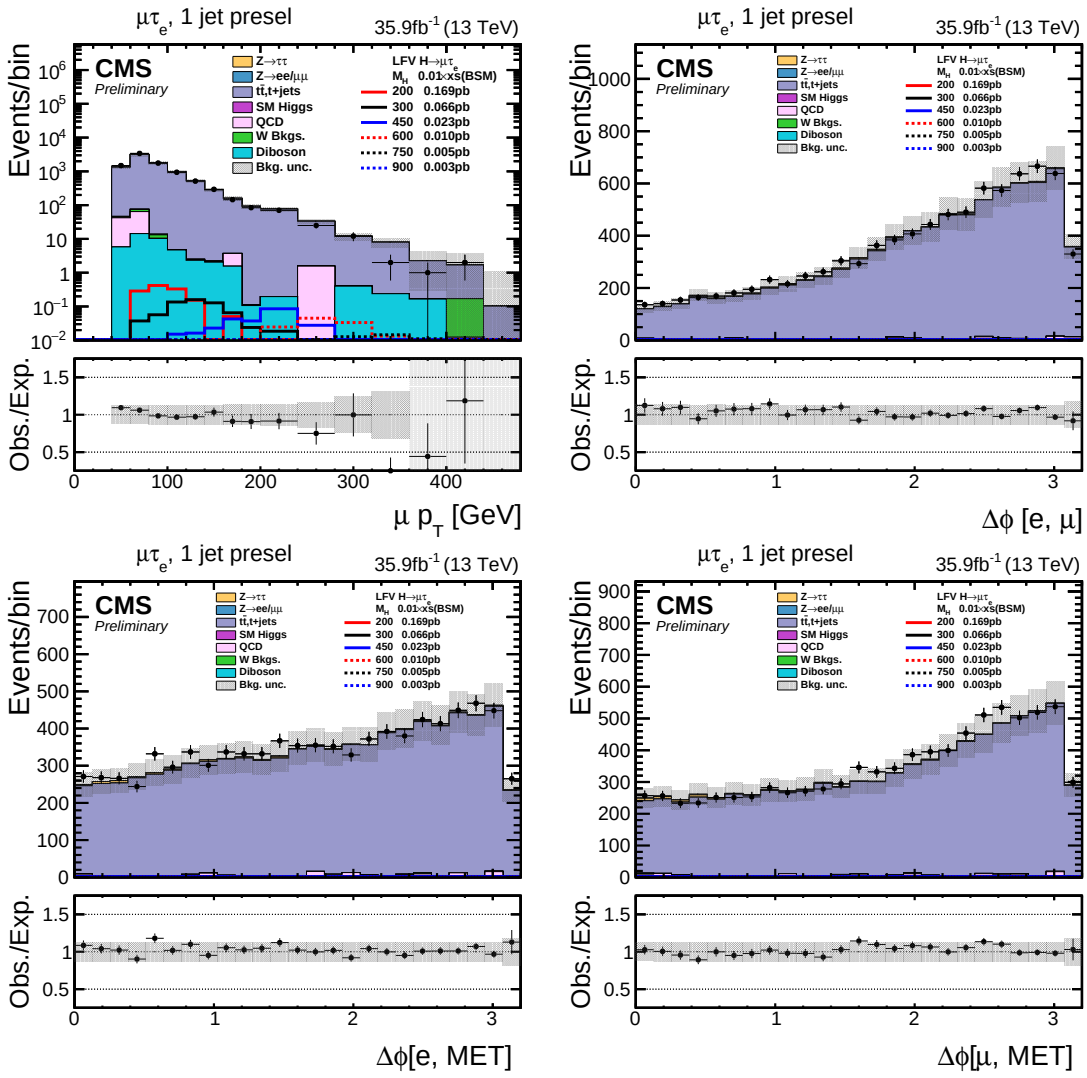


Figure 6.6: Distributions of several kinematic variables in the $t\bar{t}$ enriched control region for H $\rightarrow \mu\tau_e$ analysis.

Electroweak diboson production (WW, WZ and ZZ) forms the next largest background in $H \to \mu\tau_e$ analysis. It is estimated using simulation. All other backgrounds are much smaller. This can be seen from the distributions of kinematic variables after baseline selection, as can be seen from Figs. 5.9 and 5.10. The misidentified lepton background is even smaller here than $h \to \mu\tau_e$. The higher p_T requirement makes it even less likely for jets to be able to be misidentified as leptons. This background is estimated using the same technique as $h \to \mu\tau_e$, as described in section 6.2.3. The $Z \to \ell\ell$ ($\ell = e, \mu$) + jets and $Z \to \tau\tau$ backgrounds are estimated from simulation. Other backgrounds include SM h boson decays, $h \to WW$, $h \to \tau\tau$, single-top quark production and $W\gamma^{(*)}$ + jets, and are also estimated using simulation.

CHAPTER 7

SIGNAL EXTRACTION AND SYSTEMATIC UNCERTAINTIES

7.1 Introduction

The analysis is in essence a sophisticated counting experiment. The presence of a signal is indicated by an excess of events over the predicted background, in the distribution of a signal variable. For our analyses the signal variables are collinear mass or BDT output, as described in Sections 5 and 4. Given that there are several uncertainties, both experimental and theoretical and also due to the innate randomness in the process, it is possible that an excess is observed when there is no signal. So, when an excess is observed, a p-value which represents the probability that the excess is due to statistical fluctuations is computed. A very low p-value is taken to indicate that the excess corresponds to an observed signal and not merely a statistical fluctuation. Conversely, if no excess is observed (upper exclusion) limits are set on the product of branching fraction and production cross-section. A 95% CL (confidence level) is taken as a requirement for ruling out a signal at or above a certain value known i.e. upper exclusion limit. The first part of this chapter describes the statistical methods used, that very closely follow the procedure used for LHC Higgs boson search and described in [25].

Several sources of systematic uncertainties need to be considered when making the above measurement. The sources of these uncertainties can be theoretical, experimental or purely statistical in nature. Further, they can effect only the overall scale of the distributions (used to make the measurement), or effect there shape i.e.

change the scale differently in each bin of the distribution. All the uncertainties used in the analyses and their sources are described in the secon part of this chapter.

7.2 Statistical methods for signal extraction

In the following section, the expected signal event yields are denoted by s, and backgrounds by b. The parameter μ that appears is the signal strength modifier, which changes the signal production cross-sections of all the production mechansims by exactly the same scale μ.

7.2.1 Likelihood function

The Poisson distribution is an appropriate model for n, the number of times an event occurs in an interval if the following assumptions are true [26].

- The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
- The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals. This rate is the average number of events in the interval. λ.
- Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur.

The poisson probablity of distribution is then given by:

$$P(n_e vents) = \frac{e^{-\lambda}\lambda^n}{n!} \tag{7.1}$$

For a counting experiments such as ours, the above conditions approximately hold. The expected number of events is $\mu \cdot s + b$. The likelihood function $\mathcal{L}(data|\mu)$ is then

given by:

$$\mathcal{L}(data|\mu) = \prod_{i=1}^{bins} \frac{(\mu \cdot s_i + b_i)^{n_i}}{n_i!} e^{-\mu \cdot s_i - b_i} \tag{7.2}$$

,where n_i is the number of events observed in the bin i of the distribution, and s_i and b_i are expected number of signal and background events in that bin respectively.

7.2.2 Treatment of systematic uncertainties

All systematic uncertainties are handled by introducing them as nuisance parameters. Nuisance parameters are parameters that influence the model but are not of interest in our measurement, e.g., if we are interested in knowing only the mean of a population that is expected to be distributed as a gaussian, the standard deviation becomes a nuisance parameter for the model that we fit. In our experiment, the nuisance parameters are embedded into the likelihood function. In order for the likelihood function to have a clean factorised form [25], all sources of uncertainties considered are considered 100%-corrrelated or uncorrelated. If an uncertainty is partially correlated, it is either separated into 100%-corrrelated or uncorrelated components, or considered 100%-corrrelated or uncorrelated, depending on whichever is a more conservative estimate. The full suite of nuisance parameters is represented as θ. These effect the expected signal and backgeound yields which are now represented as $s(\theta)$ and $b(\theta)$. Each component of θ is associated with a default value $\tilde{\theta}$, reflecting our degree of belief on the real value of θ. The pdf (probablity distribution function) $\rho(\theta|\tilde{\theta})$ can then be interpreted as a posterior distribution from measurements of $\tilde{\theta}$. Using Bayes' theorem:

$$\rho(\theta|\tilde{\theta}) = \rho(\tilde{\theta}|\theta) \cdot \pi_\theta(\theta), \tag{7.3}$$

where the priors $\pi_\theta(\theta)$ are taken as flat distributions representing no prior knowledge of θ. This reformulation allows us to use the pdf of $\tilde{\theta}$ instead, i.e. $\rho(\tilde{\theta}|\theta)$ to directly constrain the likelihood of the measurement. The likelihood function after

the introduction of systematic uncertainties now becomes:

$$\mathcal{L}(data|\mu,\theta) = Poisson(data|\mu \cdot s(\theta) + b(\theta)) \cdot \rho(\tilde{\theta}|\theta) \tag{7.4}$$

Systematic unceraintites that effect only the overall scale of the distributions, correspond to a mutliplicative factor in the signal and/or background yields, and are described by log-normal pdfs. Log-normal pdfs are characterised by the width κ, and are well-suited for positively valued observables. The log-normal distribution looks like:

$$\rho(\theta|\tilde{\theta}) = \frac{1}{\sqrt{2\pi}\ \ln(\kappa)}\exp\ (\frac{\ln(\theta/\tilde{\theta})^2}{2(\ln\ \kappa)^2})\frac{1}{\theta} \tag{7.5}$$

Systematic uncertainties that effect the scale of the distribution differently in each been have the effect of altering its shape along with its scale. Such uncertainties are called shape uncertainties [27], and are modeled using a linear extrapolation method [28]. In practice, two alternate distributions obtained by varying the nuisance by ± 1 standard deviation are used, and a parameter is added to the likelihood that smoothly interpolates between these shapes.

7.2.3 Calculation of exclusion limits

The CL_s method [29–31] is used to set upper exclusion limits when no excess of data over background is observed. The test statistic used generally for hypothesis testing in searches at the LHC, uses profiling of nuisances as described above, and is based on the likelihood ratio [32], which by the Neyman-Pearson lemma is known as the most powerful discriminator. This is denoted by $\tilde{q_\mu}$, and is given by:

$$\tilde{q_\mu} = -2\ \ln\frac{\mathcal{L}(\text{data}|\mu,\hat{\theta_\mu})}{\mathcal{L}(\text{data}|\hat{\mu},\hat{\theta})},\ \text{with}\ 0 \leq \mu \leq \hat{\mu} \tag{7.6}$$

,where $\hat{\theta_\mu}$ refers to the conditional maximum likelihood estimators of θ, i.e. the

set of nuisances parameters that maximize the likelihood for a given signal strength μ, while $\hat{\mu}$ and $\hat{\theta}$ refer to the global maximum likelihood estimators for μ and θ. The lower constraint on $\hat{\mu}$ i.e., $\hat{\mu} \geq 0$ ensures that the signal rate cannot be negative, while the upper constraint that $\hat{\mu}$, which is the global maximum value, cannot be less than the value of μ under consideration is imposed to guarantee that upward fluctuations of data such that $\hat{\mu} \geq \mu$ are not considered as evidence against the signal hypothesis,i.e., a signal of strength μ.

Now, using equation 7.6, the observed value of the test statistic,$\tilde{q_\mu}^{obs}$, is calculated for the signal strength μ. Also, maximum likelihood estimators for the nuisance parameters, for the background-only($\mu = 0$) and signal-plus-background(current $\mu >$ 0 under consideration) hypotheses are calculated. They are denoted by $\hat{\theta_0}^{obs}$ and $\hat{\theta_\mu}^{obs}$ respectively, and are used to generate toy Monte carlo pseudo-datasets. These pseudo datasets are used to construct pdfs, using equation 7.6, of test statistics $f(\tilde{q_\mu}|0, \hat{\theta_0}^{obs})$ and $f(\tilde{q_\mu}|\mu, \hat{\theta_\mu}^{obs})$ by treating them as they were real data. Example of these distributions are shown in Fig. 7.1.

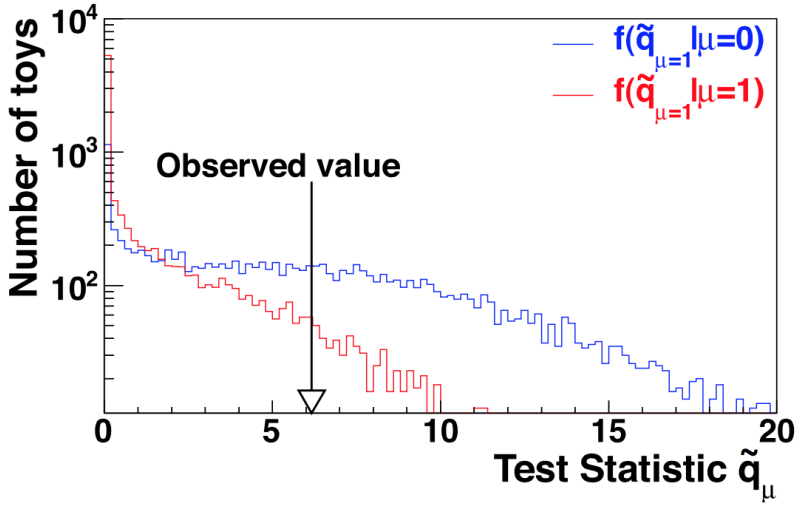


Figure 7.1: Test statistic distributions for ensembles of pseudo-data generated for signal-plus-background (red) and background-only (blue) hypotheses. [25]

Having constructed the above pdfs, it is now possible to calculate the probabilities of the observations under both hypotheses. The first quantity that we calculate is:

$$p_\mu = P(\tilde{q_\mu} \geq \tilde{q_\mu}^{obs} | \text{signal-plus-background}) = \int_{\tilde{q_\mu}^{obs}}^{\text{inf}} f(\tilde{q_\mu} | \mu, \hat{\theta_\mu}^{obs}) d\tilde{q_\mu} \tag{7.7}$$

The above quantity corresponds to $\text{CL}_{\text{s+b}}$ and measures the incompatibilty of data with signal-plus-background hypothesis. This quantity alone is not adequate for hypothesis testing in situations when the signal is so small that both hypotheses are compatible with the observation and a downward fluctuation of the background can lead to an inference of signal.

The second quantity we calculate is:

$$1 - p_b = P(\tilde{q_\mu} \geq \tilde{q_\mu}^{obs} | \text{background-only}) = \int_{\tilde{q_\mu}^{obs}}^{\text{inf}} f(\tilde{q_\mu} | 0, \hat{\theta_0}^{obs}) d\tilde{q_\mu} \tag{7.8}$$

This quantity corresponds to CL_{b} and measures the incompatibilty of data with the background. The incompatibilty of the data with background-only hypothesis alone doesn't tell us that it is indeed compatible with the signal, and so is not considered a good test of the signal hypothesis.

The ratio of the two quatities referred to as CL_{s} [29–31] helps deal with both situations above well, and is given by:

$$\text{CL}_{\text{s}} = \frac{p_\mu}{1 - p_b} \tag{7.9}$$

The 95% CL is then arrived at by iterating over μ until we have $\text{CL}_{\text{s}} = 0.05$. And the amount of signal or above, given by that μ, denoted as $\mu^{95\% CL}$, is said to be excluded at 95% CL.

7.2.4 Median expected Limits

Upper exclusion limits calculated using toy datasets of background-only expectation, are called expected limits. A large set of background-only pseudo-data is generated, and CL_s and $\mu^{95\%CL}$ is calculated for each of them. The median expected limit is calculated by integrating over this distribution until the 50% quantile is reached. The $\pm 1\sigma$ bands are calculated similary by integrating the distribution to the appropiate quantiles are reached. The calculation of median expected limits does not involve using the observed data and hence can be calculated when the analyses is blinded to prevent experimenter's bias (as mentioned in Section 5.1). This can be use to maximize the sensitivity of the search, as described in Sections 5.2.3 and 5.3.3. A more stringent(lower) median limit corresponds to a more sensitive search.

7.2.5 Quantifying an excess of events

In case an excess of data over background is observed, it is necessary to make sure beyond a reasonable doubt that the excess is not merely a fluctuation. This is quantified using the background-only p-value, which is the probability for the background to fluctuate and give an excess of events as large or larger than that observed. The same test statistic as equation 7.6 is used with the signal stength set to 0 to correspond to the background-only hypothesis:

$$\tilde{q_0} = -2 \ln \frac{\mathcal{L}(\text{data}|0, \hat{\theta}_0)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})}, \text{ with } 0 \leq \hat{\mu} \tag{7.10}$$

The constraint on $\hat{\mu}$ being greater than 0 is required so that a deficit of events in observed data is not interpreted in the same manner as we would an excess. In other words a departure from the background hypothesis in the form of deficit of events is not considered in favour of the signal hypothesis. Following the same procedure as calculation of observed limits 7.2.3 and generating pseudo-data, the distribution

$f(\tilde{q_0}|0,\hat{\theta_0}^{obs})$ is constructed. The p-value is then given by:

$$p_0 = P(\tilde{q_0} \geq \tilde{q_0}^{obs}) = \int_{\tilde{q_0}^{obs}}^{\inf} f(\tilde{q_0}|0,\hat{\theta_0}^{obs}) d\tilde{q_0} \tag{7.11}$$

The p-value can be converted to significance $\mathcal{Z}_0$, which is an equivalent way of quantifying an excess and is related to the p-value by the following:

$$p_0 = \int_{\mathcal{Z}_0}^{\inf} \frac{1}{\sqrt{2\pi}} exp(-x^2/2) dx \tag{7.12}$$

Broadly, the signficance corresponds to how far into the tail of the distribution (i.e., away from the most probable value), assuming background hypothesis, the test statistic value corresponding to the observed data lies. The farther it is, the less likely it is to have been a fluctuation. The conventional standard in high energy physics to be able to claim observation of a process is a significane of 5σ, which corresponds to a p-value of 2.8×10^{-7}.

7.2.6 Experminetal uncertainties

7.2.7 Signal extraction

7.3 Heavy Higgs Analysis

7.3.1 Theoretical uncertainties

7.3.2 Experminetal uncertainties

7.3.3 Signal extraction

CHAPTER 8

INTERPRETATION OF RESULTS

CHAPTER 9

CONCLUSION

APPENDIX A

BOOSTED DECISION TREES

A.1 Introduction

BIBLIOGRAPHY

1. L. Evans and P. Bryant. LHC machine. In *Journal of Instrumentation*, volume 3, August 2008.

2. CMS Collaboration. CMS integrated luminosity - public results. Website. `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`.

3. Joel Butler et al. Phase II Upgrade Scope Documentt. *CERN-LHCC-2015-019*, 2015.

4. CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3, August 2008.

5. Stefan Gieseke .et al Andy Buckley, Jonathan Butterworth. General-purpose event generators for lhc physics. *Physics Reports*, 504, July' pages = 2011.

6. Wikipedia. Monte carlo method. Website, . `https://en.wikipedia.org/wiki/Monte_Carlo_method`.

7. F. Krauss et al. J. Alwall, S. Hche. Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions. *EPJC*, 53, 2008.

8. Leif Lonnblad. Correcting the color dipole cascade model with fixed order matrix elements. *JHEP*, 0205, 2002.

9. Rikkert Frederix and Stefano Frixione. Merging meets matching in MC@NLO. *JHEP*, 12:061, 2012. doi: 10.1007/JHEP12(2012)061.

10. Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. A brief introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852, 2007. doi: 10.1016/j.cpc.2008.01.036.

11. David Grellscheid et al. Johannes Bellm, Stefan Gieseke.

12. Paolo Nason. A new method for combining NLO QCD with shower Monte Carlo algorithms. *JHEP*, 11:040, 2004. doi: 10.1088/1126-6708/2004/11/040.

13. Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with parton shower simulations: the POWHEG method. *JHEP*, 11:070, 2007. doi: 10.1088/1126-6708/2007/11/070.

14. Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 06:043, 2010. doi: 10.1007/JHEP06(2010)043.

15. Simone Alioli, Keith Hamilton, Paolo Nason, Carlo Oleari, and Emanuele Re. Jet pair production in POWHEG. *JHEP*, 04:081, 2011. doi: 10.1007/JHEP04(2011)081.

16. Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. NLO Higgs boson production via gluon fusion matched with shower in POWHEG. *JHEP*, 04:002, 2009. doi: 10.1088/1126-6708/2009/04/002.

17. E. Bagnaschi, G. Degrassi, P. Slavich, and A. Vicini. Higgs production via gluon fusion in the POWHEG approach in the SM and in the MSSM. *JHEP*, 02:088, 2012. doi: 10.1007/JHEP02(2012)088.

18. J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014. doi: 10.1007/JHEP07(2014)079.

19. Johan Alwall, Michel Herquet, Fabio Maltoni, Olivier Mattelaer, and Tim Stelzer. MadGraph 5: going beyond. *JHEP*, 06:128, 2011. doi: 10.1007/JHEP06(2011)128.

20. S. Agostinelli et al. GEANT4 — a simulation toolkit. *Nucl. Instrum. Meth. A*, 506:250, 2003. doi: 10.1016/S0168-9002(03)01368-8.

21. Gionata Luisoni, Paolo Nason, Carlo Oleari, and Francesco Tramontano. $HW^{\pm}/HZ$ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO. *JHEP*, 10:083, 2013. doi: 10.1007/JHEP10(2013)083.

22. Aaron Roodman. Blind Analysis in Particle Physics . 2003. doi: arXiv:physics/0312102v1.

23. CMS Collaboration. Observation of the Higgs boson decay to a pair of τ leptons. *Phys. Lett. B*, 779:283, 2018. doi: 10.1016/j.physletb.2018.02.004.

24. A. Hoecker et al. *TMVA - Toolkit for Multivariate Data Analysis*, 2017. `http://tmva.sourceforge.net/`.

25. CMS and ATLAS Collaborations. Procedure for the lhc higgs boson search combination in summer 2011. Technical report, August 2011.

26. Wikipedia. Poisson distribution. Website, . `https://en.wikipedia.org/wiki/Poisson_distribution`.

27. J. S. Conway. Incorporating nuisance parameters in likelihoods for multisource spectra. In *PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva*, January 2011.

28. A. L. Read. Linear interpolation of histograms. *Nucl. Instrum. Meth*, 425, April 1999.

29. A. L. Read. Presentation of search results: The CL_s technique. *Journal of Physics G*, 28, September 2002.

30. A. L. Read. Modified frequentist analysis of search results (The CL_s method). In *Workshop on confidence limits, CERN, Geneva, Switzerland, 17-18 Jan 2000*, August 2000.

31. Thomas Junk. Confidence level computation for combining searches with small statistics. *Nuclear Instruments and Methods A*, 434, September 1999.

32. Eilam Gross Glen Cowan, Kyle Cranmer and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *European Physics Journal C*, 71, February 2011.

This document was prepared & typeset with pdfLaTeX*, and formatted with* NDdiss2$_\varepsilon$ *classfile (v3.2017.2[2017/05/09])*