UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Studying the Higgs to tau-pair process with CMS Open Data

Asdrubal Eduardo Cruz Basante

Física

Trabajo de fin de carrera presentado como requisito

para la obtención del título de

Licenciado en Física

Quito, 19 de mayo de 2021

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN DE TRABAJO DE FIN DE CARRERA

Studying the Higgs to tau-pair process with CMS Open Data

Asdrubal Eduardo Cruz Basante

Nombre del profesor, Titulo académico	Edgar Carrera, Ph.D.

Quito, 19 de mayo de 2021

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Asdrubal Eduardo Cruz Basante

Código: 00127808

Cédula de identidad: 1723246599

Lugar y fecha: Quito, 19 de mayo de 2021

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

Dedicado a mi padre y madre.

Agradecimientos

Me gustaría expresar un profundo agradecimiento a mi familia, gracias por todo su apoyo. Y a mi supervisor, Edgar Carrera, por su guía y estímulo durante mi carrera universitaria.

Resumen

En este trabajo se estudia el decaimiento del bosón de Higgs en dos leptones τ, utilizando como estado final un leptón muón y un leptón τ que decae hadrónicamente (τ_h). El estudio está basado en el análisis oficial de CMS "Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons" publicado en el 2014. Se utilizan datos y simulación de eventos del experimento CMS del 2012 y se los procesa en la nube bajo la plataforma Kubernetes. Estos datos están guardados en nueve set de datos de colisión protón-protón, juntos contienen 21185 archivos AOD, y corresponden a una luminosidad integrada de 11.1 fb^{-1} a una energía de centro de masa de 8 TeV. Este estudio arroja resultados muy parecidos al estudio oficial de CMS y representan un primer vistazo al proceso de evidenciar una nueva partícula.

Palabras clave: Higgs to Tau Tau, CMS, LHC, Kubernetes cluster.

Abstract

In this work the decay of the Higgs boson in two τ leptons is studied, using a muon lepton and a hadronically decayed τ lepton (τ_h) as final state. The study is based on the official CMS analysis “Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons” published in 2014. Data and simulation of events from the 2012 CMS experiment are used and processed in the cloud under the Kubernetes platform. These data are stored in nine proton-proton collision data-sets, together they contain 21185 AOD files, and correspond to an integrated luminosity of fb^{-1} at a center of mass energy of 8 TeV. This study yields results very similar to the official CMS study and represents a first look at the process of evidencing a new particle.

Keywords: Higgs to Tau Tau, CMS, LHC, Kubernetes cluster.

Contents

1 Introduction **12**

2 The CMS experiment **14**

2.1 Event reconstruction 16

2.2 Event selection 18

2.3 The CMS open data 20

3 Analysis strategy **22**

3.1 Signal process 23

3.2 Tau decay modes 23

3.3 Background processes 24

4 Computational process **27**

4.1 First part 28

4.2 Second part 29

5 Results **31**

5.1 Summary and conclusions 33

List of Figures

2.1 A cross section through a segment of the CMS detector indicating the responses of the various detection systems to different types of particles [28]. 15

3.1 Feynman diagrams for the production of the Higgs boson through gluon-gluon fusion (left) and for the production associated with vector boson fusion (right). 23

5.1 Observed primary vertice number distributions in the 8 TeV $\mu\tau_h$ channel. 32

5.2 Observed tau pseudorapidity distributions in the 8 TeV $\mu\tau_h$ channel. 32

5.3 Observed visible $m_{\tau\tau}$ distributions in the 8 TeV $\mu\tau_h$ channel. 33

5.4 Observed τ decay mode distributions in the 8 TeV µτh channel. 42

5.5 Observed muon pseudorapidity distributions in the 8 TeV µτh channel. 42

5.6 Observed τ pseudorapidity distributions in the 8 TeV µτh channel. . . 42

5.7 Observed muon isolation distributions in the 8 TeV µτh channel. . . . 42

5.8 Observed τ isolation distributions in the 8 TeV µτh channel. 43

5.9 Observed leading jet b-tag distributions in the 8 TeV µτh channel. . . 43

5.10 Observed trailing jet b-tag distributions in the 8 TeV µτh channel. . . 43

5.11 Observed di-jet $\triangle\eta$ distributions in the 8 TeV µτh channel. 43

5.12 Observed leading jet pseudorapidity distributions in the 8 TeV µτh channel. 44

5.13 Observed trailing jet pseudorapidity distributions in the 8 TeV µτh channel. 44

5.14 Observed leading jet mass distributions in the 8 TeV µτh channel. . . 44

5.15 Observed trailing jet mass distributions in the 8 TeV µτh channel. . . 44

5.16 Observed leading jet ϕ distributions in the 8 TeV µτh channel. 45

5.17 Observed trailing jet ϕ distributions in the 8 TeV µτh channel. 45

5.18 Observed leading jet p_T distributions in the 8 TeV µτh channel. 45

5.19 Observed trailing jet p_T distributions in the 8 TeV µτh channel. 45

5.20 Observed muon mass distributions in the 8 TeV µτh channel. 46

5.21 Observed τ mass distributions in the 8 TeV µτh channel. 46

5.22 Observed visible di-τ mass distributions in the 8 TeV µτh channel. . . 46

5.23 Observed di-jet mass distributions in the 8 TeV µτh channel. 46

5.24 Observed muon transverse mass distributions in the 8 TeV µτh channel. 47

5.25 Observed τ transverse mass distributions in the 8 TeV µτh channel. . 47

5.26 Observed jet number distributions in the 8 TeV µτh channel. 47

5.27 Observed primary vertice number distributions in the 8 TeV µτh channel. 47

5.28 Observed muon ϕ distributions in the 8 TeV µτh channel. 48

5.29 Observed tau ϕ distributions in the 8 TeV µτh channel. 48

5.30 Observed missing transverse momenta distributions (as a function of ϕ) in the 8 TeV µτh channel. 48

5.31 Observed muon transverse momenta distributions in the 8 TeV µτh channel. 48

5.32 Observed τ transverse momenta distributions in the 8 TeV µτh channel. 49

5.33 Observed missing transverse momenta distributions in the 8 TeV µτh channel. 49

5.34 Observed visible di-τ transverse momenta distributions in the 8 TeV µτh channel. 49

5.35 Observed di-jet transverse momenta distributions in the 8 TeV µτh channel. 49

5.36 Observed muon charge distributions in the 8 TeV µτh channel. 50

5.37 Observed τ charge distributions in the 8 TeV µτh channel. 50

Prologue

One of the goals of the CMS collaboration is to make open data sets easier to access for people outside the CMS. In other words, reduce the access threshold to this data, which are available for anyone to analyze as they see fit. In 2017, a theory group at MIT published two peer-reviewed publications using this data. A physics analysis typically spans hundreds of gigabytes of data. In CMS, this is commonly done using high-throughput batch systems, such as the HTCondor facility at CERN, but not everyone has access to this. As an option, computing resources can now be accessed through a public cloud. The objective of this work was to find a way to perform a physical analysis with the CMS open data using cloud services.

Chapter 1

Introduction

> "The apparent strengths of the forces in any field theory depend on two kinds of numerical parameter: the masses (if any) of the particles like W and Z particles that transmit the forces, and certain intrinsic strengths (coupling constants) that characterize the likelihood for particles to be emitted and reabsorbed in particle reactions. The masses arise from spontaneous symmetry breaking, but the intrinsic strengths are numbers that appear in the underlying equation of the theory."
>
> – *Steven Weinberg, Dreams of a Final Theory: The Scientist's Search for the Ultimate Laws of Nature.*

Understanding the electroweak symmetry breaking mechanism, through which the W and Z bosons become massive, is a goal of the Large Hadron Collider (LHC) physics program. In the Standard Model (SM) of particle physics [1], the electroweak symmetry is broken by the Brout-Englert-Higgs mechanism [2], which predicts the existence of a neutral scalar particle, known as the Higgs boson. On July 4, 2012, at CERN, ATLAS and CMS Collaborations announced the discovery of a new boson with a mass of around 125 GeV [3]. In addition, its properties were compatible with those of the SM Higgs boson [4]. According to SM, fermions are massive due to Yukawa couplings between the Higgs field and the fermionic fields [5]. By measuring these couplings, this boson can be identified as the SM Higgs boson. The $\tau\tau$ decay channel is the most promising, since its expected event rate in the SM is larger compared to the other lepton decay modes. In addition, the contribution of background events with respect to the $b\bar{b}$ decay mode is very small [5, 6].

This work studies the decay of the Higgs boson into two τ leptons, using a muon lepton and a hadronically decayed τ lepton as the final state. Data and simulation of events from the 2012 CMS experiment are used. The study is based on the official CMS analysis "Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons" and its simplified CMS Open Data example [6]. The goal of the original CMS analysis was to prove the existence of the Higgs boson that decays into two τ leptons. In this, all systematic uncertainties are considered, which is a very complex task. For this reason, here, the analysis is reduced to the qualitative study of the event properties without a statistical treatment. However, this reduced analysis already has a high degree of complexity, which makes it a good example of the process involved in showing the existence of a new particle. The objective is to understand and reproduce the analysis using resources outside the CMS. The thesis structure is as follows. **Chapter 2** talks about the CMS detector, event reconstruction and its selection. **Chapter 3** provides an overview of the analysis strategy. **Chapter 4** describes the computational proccess. **Chapters 5** presents the results.

Chapter 2

The CMS experiment

"For thousands of years, it had been nature--and its supposed creator--that had had a monopoly on awe. It had been the icecaps, the deserts, the volcanoes and the glaciers that had given us a sense of finitude and limitation and had elicited a feeling in which fear and respect coagulated into a strangely pleasing feeling of humility, a feeling which the philosophers of the eighteenth century had famously termed the sublime.

But then had come a transformation to which we were still the heirs.... Over the course of the nineteenth century, the dominant catalyst for that feeling of the sublime had ceased to be nature. We were now deep in the era of the technological sublime, when awe could most powerfully be invoked not by forests or icebergs but by supercomputers, rockets and particle accelerators. We were now almost exclusively amazed by ourselves."

– *Alain de Botton, The Pleasures and Sorrows of Work.*

The LHC (Large Hadron Collider) consists of 7 experiments, one of them is the CMS (Compact Muon Solenoid). This detector consists of several layers of subdetectors designed to sense different particles. The essential element of the CMS is a superconducting solenoid that has a diameter of 6 internal meters and produces a magnetic field of 3.8 T. Inside this solenoid are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass / scintillator hadron calorimeter (HCAL). All these layers are complemented by extensive forward calorimetry. In the external part of the solenoid there are gas ionization chambers destined to detect the muons [7]. A cross section of the CMS experiment is shown in Figure 2.1.

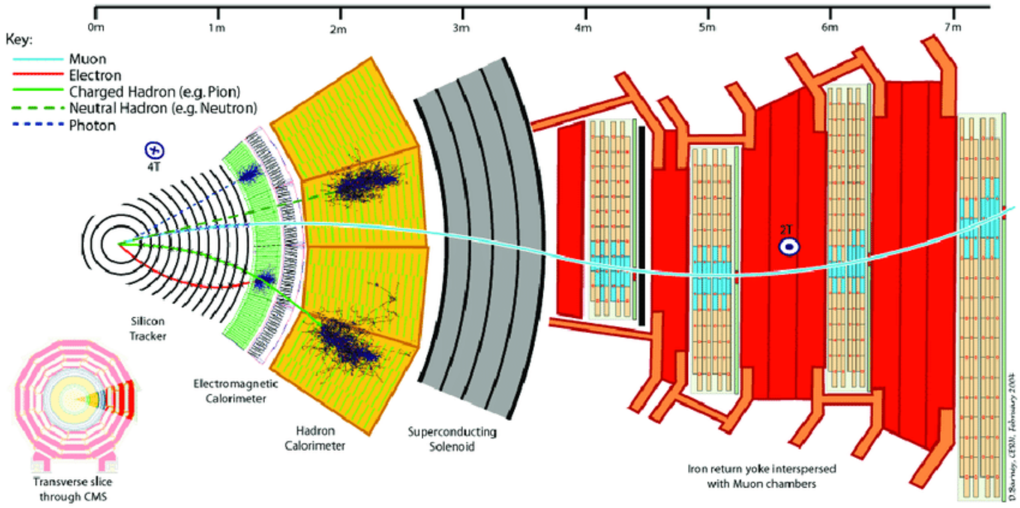
0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D. Barney, CERN, February 2004

Figure 2.1: A cross section through a segment of the CMS detector indicating the responses of the various detection systems to different types of particles [28].

The CMS has a two-level trigger system: level 1 (L1) and the high level trigger (HLT). The L1 is made up of customized processors that use the information from the calorimeters and muon detectors to select the most interesting events, all of which are done at a speed of approximately 100 kHz in a time interval of less than 4 µs. The high level trigger, or the second level, is made up of a farm of event reconstruction programs optimized for fast processing, thus further reducing the event rate to a value of around 3 kHz, for then be stored [8].

The coordinate system used in the CMS experiment is of the right-handed type, the origin is at the nominal interaction point with the x axis pointing to the center of the LHC and the y axis pointing upwards, that is, perpendicular to the LHC plane, while the z-axis points in the direction of the beam counterclockwise. From the z axis the polar angle θ is measured and in the (x, y) plane the azimuth angle φ is measured [8]. Pseudorapidity is defined as: $\eta \equiv -\ln(\tan(\theta/2))$.

2.1 Event reconstruction

In 2012, the average number of inelastic proton-proton collisions that occurred per LHC bunch crossing was 21. Collision vertices can be separated 0.5 mm in the direction of the beam using a tracking system [9]. The squared transversal momenta of all associated tracks are added, this for each vertex. The vertex with the largest sum is called the primary vertex and corresponds to the hard scattering process. The rest of the proton-proton collisions that happen in the same bunch crossing are called pileup (PU) [5].

Particles that result from proton-proton collisions, such as charged and neutral hadrons, photons, muons and electrons, can be identified and reconstructed by applying a particle-flow (PF) algorithm, which uses the information from the subdetectors of the CMS. With these particles, the missing transverse energy vector $\vec{E}_T^{miss}$, jets and candidates τ_h can be reconstructed, the isolation of the leptons can also be quantified. The FASTJET clustering algorithm reconstructs jets from all particles [10]. The combined secondary vertex (CSV) algorithm reconstructs the jets that come from the hadronization of the b quarks [11]. Furthermore, jets originating from the PU are identified and rejected using criteria based on information from the vertex and the shape of the jet [12]. From all the reconstructed particles, the $\vec{E}_T^{miss}$ and its magnitude can be calculated, with a very high resolution, using a multivariate regression process (BDT) [13].

To identify muons, a better quality track reconstruction is required as well as additional measurements from the tracker and the moun systems [14]. Electrons are identified using a multivariate discriminant and electromagnetic calorimeter measurements that must match the measurements of the tracker [15]. Using the “hadron-plus-strips” algorithm, the τ_h candidates are reconstructed, using data from charged hadrons and photons [16]. Electrons and muons that are erroneously identified as τ_h candidates are discarded, using criteria based on the consistency of measurements from the tracker, calorimeters and muon detectors [5]. To discard non-prompt and misidentified leptons, absolute lepton isolation is defined as:

$$I^L \equiv \sum_{charged} p_T + \max\left(0, \sum_{neutral} p_T + \sum_{\gamma} p_T - \frac{1}{2} \sum_{charged,\, PU} p_T\right) \tag{2.1}$$

In equation 2.1, $\sum_{charged} p_T$ is the sum of the transverse momenta of the charged hadrons, muons and electrons coming from the primary vertex and that are inside in a cone of size $\triangle R = \sqrt{(\triangle\eta)^2 + (\triangle\phi)^2} = 0.4$. $\sum_{neutral} p_T$ and $\sum_{\gamma} p_T$ are the sum of the transverse momentum of the neutral hadrons and of the photons respectively. $\sum_{charged,\,PU} p_T$ is the sum of the transverse momenta of charged hadrons from PU vertices. In addition, this sum is multiplied by a factor of 1/2, since this corresponds to the production rate of charged hadrons from neutral hadrons of proton-proton collisions [5]. The relative lepton isolation is: $R^L = I^L/p_T^L$.

Events with the SM Higgs boson signal that comes from gluon-gluon fusion or VBF are generated by POWHEG 1.0 [17], while the signatures $Z + jets$, $W + jets$, $t\bar{t} + jets$, and diboson are generated by MADGRAPH 5.1 [18]. The POWHEG and MADGRAPH generators are interconnected with PYTHIA, which continues the simulation process.

2.2 Event selection

According to the selected number of muons, electrons and candidates τ_h, the events are classified in various modes and the resulting samples are independent. To optimize the trigger and the offline selection, simulated data are used for each channel, thereby maximizing the sensitivity of the SM Higgs boson signal [15]. All the selection criteria for the LL' channel are shown in Table 1.

Table 1: Lepton selection for the LL' channel.

Channel	HLT requirement	Lepton selection		
		p_T (GeV)	$\lvert\eta\rvert$	Isolation
μμ	μ(17) & μ(8)	$p_T^{\mu_1} > 20$	$\lvert\eta^{\mu_1}\rvert < 2.1$	$R < 0.1$
		$p_T^{\mu_2} > 10$	$\lvert\eta^{\mu_2}\rvert < 2.4$	
ee	$e(17)$ & $e(8)$	$p_T^{e_1} > 20$	$\lvert\eta^{e}\rvert < 2.3$	$R < 0.1 - 0.15$
		$p_T^{e_2} > 10$		
μτ_h	μ(12–18) & τ_h(10–20)	$p_T^{\mu} >$ 17–20	$\lvert\eta^{\mu}\rvert < 2.1$	$R < 0.1$
		$p_T^{\tau_h} > 30$	$\lvert\eta^{\tau_h}\rvert < 2.4$	$I < 1.5$
$e\tau_h$	e(15–22) & τ_h(15–20)	$p_T^{e} > 20 - 24$	$\lvert\eta^{e}\rvert < 2.1$	$R < 0.1$
		$p_T^{\tau_h} > 30$	$\lvert\eta^{\tau_h}\rvert < 2.4$	$I < 1.5$
$\tau_h\tau_h$	$\tau_h(30)$ & $\tau_h(30)$ & $jet(30)$	$p_T^{\tau_h} > 45$	$\lvert\eta^{\tau_h}\rvert < 2.1$	I <1
eμ	$e(17)$ & μ(8)	$p_T^{l_2} > 20$	$\lvert\eta^{\mu}\rvert < 2.1$	$R < 0.1 - 0.15$
	$e(8)$ & μ(17)	$p_T^{l_2} > 10$	$\lvert\eta^{e}\rvert < 2.3$	

In 2012, the muon p_T threshold increased by 20 GeV for the $\mu\tau_h$ channel, as the instantaneous luminosity increased. In the decay $H \to \tau\tau$, the leptons must be oppositely charged.

In channel $l\tau_h$, the background $W + jets$ signal is reduced by the following cut [5]:

$$m_T \equiv \sqrt{2p_T^l E_T^{miss}[1 - \cos(\Delta\varphi)]} < 30GeV, \tag{2.2}$$

where p_T^l is the transverse momentum of the lepton l, and $\Delta\varphi$ is the azimuthal angle between l direction and the $\overrightarrow{E}_T^{miss}$.

The background signal $t\bar{t}$ is reduced, in the $e\mu$ channel, by the discriminant BDT. It uses several kinematic variables from the $e\mu$ system, the $\overrightarrow{E}_T^{miss}$, the shortest distance between the leptons and the primary vertex, and the b-tagging CSV discriminator of the main jet with $p_T > 20$ GeV, if there is some jet [5].

2.3 The CMS open data

The data collected by the CMS experiment is of exceptional value. It is a scientific opportunity to be able to reuse it. This opportunity requires overcoming very great challenges, since data of such magnitude has never been stored in history. One of the goals of the CMS community is to preserve its data at various levels of complexity, for later reuse by the scientific and academic community. CMS believes that free access to data will allow those who take advantage of this opportunity to reach the maximum scientific potential [19].

The information collected by the experiment is stored in files of different formats. At the most basic level is RAW data, this is the first to be generated and contain information directly related to the subdetectors, for example, which of these were activated and at what voltage. All this RAW data is stored in CERN's T0 computer center [20] and exceeds 200 petabytes of information [21]. From RAW files the AOD files (Analysis Object Data) are generated, these contain reconstructed physical objects such as electrons, muons, etc. The "legacy" data taken in the run from 2011 to 2012 (first run) is stored in this AOD format [22].

Currently, AOD files from the years 2010 to 2012 have been released, with simulations and real data. These, together with virtual machines that provide the legacy software necessary to perform an analysis, are accessible from CERN's Open Data portal [23]. Here you can find data along with instructions and examples of how to perform an analysis. The project continues to grow and develop, with the intention of facilitating the use of the data for people outside the CMS Collaboration.

The CMS classifies data into sets. In this analysis, several datasets taken during 2012 are used, these are in AOD format and correspond to simulations and real data.

The datasets corresponding to simulations used in this analysis are the following: GluGluToHToTauTau, VBF_HToTauTau, DYJetsToLL, TTbar, W1JetsToLNu, W2JetsToLNu, and W3JetsToLNu. And the real datasets used are Run2012B_TauPlusX and Run2012C_TauPlusX. In total, there are 21185 AOD files within these datasets.

Chapter 3

Analysis strategy

> "At heart, science is the quest for awesome - the literal awe that you feel when you understand something profound for the first time. It's a feeling we are all born with, although it often gets lost as we grow up and more mundane concerns take over our lives."
>
> – *Sean Carroll, The Particle at the End of the Universe: How the Hunt for the Higgs Boson Leads Us to the Edge of a New World.*

First, the notation used throughout this analysis is defined: the symbol τ_h denotes the reconstruction of a τ lepton that has decayed hadronically. The reconstruction of the τ_h candidates occurs in decay modes with one or three charged particles. The symbol l denotes an electron or a muon, and the symbol L denotes any type of reconstructed charged lepton (electron, muon, or τ_h).

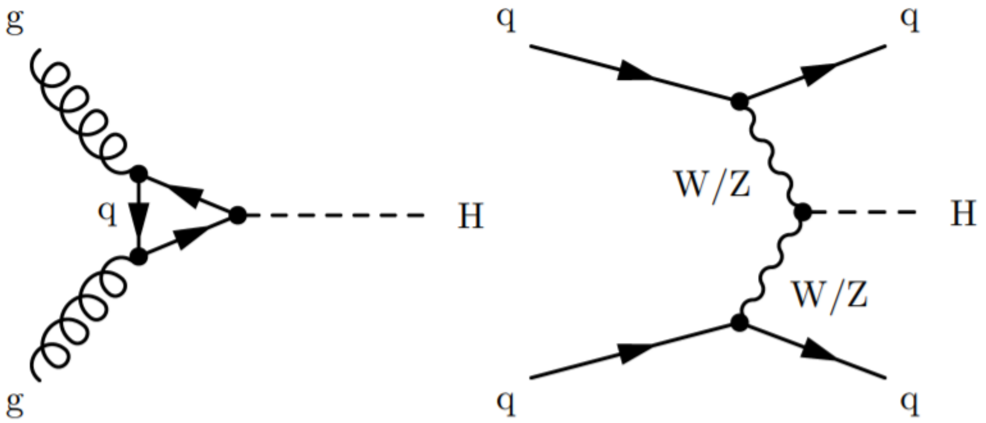

Figure 3.1: Feynman diagrams for the production of the Higgs boson through gluon-gluon fusion (left) and for the production associated with vector boson fusion (right).

3.1 Signal process

In this analysis, the signal process sought is the production of the Higgs boson that decays into two τ leptons. Gluon fusion and vector boson fusion (VBF) are the main production modes of the Higgs boson, labeled $gg \to H$ and $qq \to H$ respectively. The two Feynman diagrams that describe these processes are shown in Figure 3.1.

3.2 Tau decay modes

The life time of the τ lepton is very short, approximately 290 femtoseconds, after which it decays into other particles. The τ lepton can decay into one muon or one electron and two neutrinos, these two modes have probability around 20%. The rest of the modes consist of a combination of hadrons (such as pions and kaons) and one neutrino [6]. The final states of the process signal

$H \to \tau\tau$ contain two charged leptons, that is, there are 6 final states of τ pairs: $LL' = \mu\tau_h$, $e\tau_h$, $\tau_h\tau_h$, $e\mu$, $\mu\mu$, and ee.

It should be noted that the official CMS analysis studied the 6 mentioned decay modes, while this analysis only considers the pair $LL' = \mu\tau_h$ in which one τ lepton decays into a muon and two neutrinos and the other τ lepton decays hadronically.

3.3 Background processes

There are other processes that can produce signatures very similar to that of the Higgs boson that decays in two τ leptons, and these must be considered to make conclusions from the data [6]. The most dominant processes with a signal similar to the signal $H \to \tau\tau$ are presented below. The analysis estimates the contribution of these background processes using simulated data, except for the QCD multijet process, which uses real data from 2012.

The most dominant background process is the Z boson signal that decays into two τ leptons ($Z \to \tau\tau$). The leading production is called the Drell-Yan process [6], where a quark is annihilated with an anti-quark. Like the Higgs boson, the Z boson decays directly into two τ leptons; this process is quite difficult to distinguish from the $H \to \tau\tau$ signal.

In addition to the $Z \to \tau\tau$ signal, the Z boson has the same probability of decaying into electrons and muons ($Z \to ll$). And although this process does not have any τ lepton, during the reconstruction one can appear by chance.

Commonly, electrons or jets are the objects that are mistakenly identified as a τ_h(hadronically decayed τ lepton).

In the LHC, often, W bosons in association with jets ($W+jets$) are produced that can decay into any lepton. In the case that a muon from a W boson and a misidentified jet as a τ are selected, a signal, similar to $H \rightarrow \tau\tau$, can occur. But this process can be evaded efficiently with a cut in the selection of events in the transverse mass of the muon and in the missing energy [5].

Also, in the LHC, top anti-top pairs ($t\bar{t}$) are produced by quark anti-quark annihilation or gluon fusion. Since, most of the time, a top quark decays into a W boson and a bottom quark, identification errors can occur with the signal from the $W + jets$ process, which was explained previously. However, it is possible to discard those events where the bottom quarks decay into jets, reducing this background efficiently [6].

The multijet background, also called QCD background, covers decays with a large number of jets, this happens very often in the LHC. Such events can be wrongly selected in this analysis. Since the proper simulation of these events is computationally heavy, the contribution of this signal is taken from real data. Therefore, τ pairs are selected, with the same selection of the signal $H \rightarrow \tau\tau$, but both τ leptons are required to have the same charge. In the end, all known simulation processes are subtracted from the histogram. This resulting histogram can be considered as an estimate of the QCD multijet process, since the production of misidentified τ leptons does not depend on the charge [6].

The signal $H \rightarrow \tau\tau$ decaying in the $LL' = \mu\tau_h$ channel suffers from other contributions, these contributions with similar signatures, presented above, are treated as background in the search for $H \to \tau\tau$ decay.

To separate the events of the signal $H \to \tau\tau$ with the events of the signal $Z \to \tau\tau$, the visible mass m_{vis} of the LL' system can be used, since the signal $Z \to \tau\tau$ constitutes an important irreducible background. However, the separation of the variable m_{vis} can be limited, since the neutrinos coming from a τ lepton can take a large part of the energy. In events with $H \to \tau\tau$ and $Z \to \tau\tau$ signals coming from gluon-gluon fusion or VBF, the only source of neutrinos is the decay of the τ lepton.

Chapter 4

Computational process

> "A physics analysis usually encompasses running over hundreds of gigabytes of data. At CMS, this is usually performed using high-throughput batch systems such as the HTCondor installation at CERN and at other research institutions as well as the worldwide LHC computing grid (WLCG). Not everyone will have these resources available at their own institution, but nowadays anyone can get access to computing resources via public cloud vendors."
>
> – *CMS collaboration.*

To obtain the images shown in Chapter 5, relatively long processing is performed on the data sets specified in Chapter 2. The whole process can be divided into two parts:

The first part consists of reducing the size of the data to be analyzed, these data are in AOD format. For this, a tool is used that transforms them to a reduced format called NanoAOD, this tool is AOD2NanoAODOutreachTool [24]. In this way, the size of the files is considerably reduced.

The second part is the creation of histograms and images for most of the variables within the reduced NanoAOD files [6].

4.1 First part

This is the computationally intensive part, that is, completing the process of transforming the data from AOD to NanoAOD format. The 9 datasets mentioned contain 21285 AOD files in total, so processing them on a single computer with the AOD2NanoAODOutreachTool would take an estimated time of 2 months without any pause. But during this time things can happen that spoil the analysis and we would have to repeat it. As an alternative to this, high-throughput batch systems like HTCondor can be used at CERN, but not everyone has access to this.

Today anyone can access computing resources through a public cloud. For example, Google Cloud, and among the various services that Google Cloud offers, one of them is Google Kubenetes Engine. Kubernetes is an open source platform for managing service and workflows. This allows us to run applications in containers, and these containers are mounted on virtual machines in the GOOGLE data center [25]. Therefore, Kubernetes is an option. We are especially interested in clusters. A Kubernetes cluster is a set of nodes (or worker machines) that run containerized applications [26].

The goal is to find a way to process each AOD file in parallel, thereby reducing the overall processing time. So, a way to perform the AOD2NanoAOD step was developed using a Kubernetes cluster. Within Kubernetes you can create clusters with the number of nodes you need. A node is a working machine, and on each machine a container is created to perform each task. So, several containers are created simultaneously, and each of these processes an AOD file. In Kubernetes, each container created is named as a pod.

To complete this part of the analysis, it is necessary to configure 9 workflows to process 9 datasets with the AOD2NanoAOD tool. Detailed steps to perform this process in a Kubernetes cluster are shown in Annex A. As a result, 9 files are obtained in NanoAOD format (one for each dataset). The time to complete the AOD to NanoAOD step through the 9 workflows in the Kubernetes cluster was 24 hours.

4.2 Second part

In this part a typical workflow in CMS is followed, the full details of the steps to follow can be found in the analysis code [6]. Below is a description of each step to obtain the images presented in Chapter 5.

- First, the 9 files, in NanoAOD format, are taken and a preselection process is carried out that significantly reduces the size of the data. Here, $\mu\tau_h$ pairs that were probably originated by the Higgs boson are selected. So in this step it is produced preselected data sets from the original files.
- The files produced in the previous step retain information of quantities selected for each event. Histograms are created for these quantities for

each skimmed dataset. Since the data-driven QCD estimation exists, histograms have to be created for the selection of τ lepton pairs with the same charge.

- Then, the final graphs are created by combining the histograms produced previously, these show the data taken in the CMS together with the expectation of the background estimation.

These graphs represent a first look at the process of evidencing a new particle, since they allow us to observe the contribution of various physical processes that occurred in the CMS detector.

Chapter 5

Results

"Historically, nature has been very good at surprising us."

– *Sean Carroll, The Particle at the End of the Universe: How the Hunt for the Higgs Boson Leads Us to the Edge of a New World.*

The search for Higgs boson events that stand out from the expected background entails an adjustment based on the final variable, this is $m_{\tau\tau}$ or m_{vis} in the channel $\mu\tau_h$ [27].

The results produced in this analysis are presented below, the graphs show the data collected in the CMS in comparison to the estimation of the background processes, which are explained in Chapter 3. The analysis creates 34 plots of various observables. The figures below show 3 graphs obtained.

As mentioned above, you can separate the events of the signal $H \to \tau\tau$ from the events of the signal $Z \to \tau\tau$ (this constitutes an important irreducible

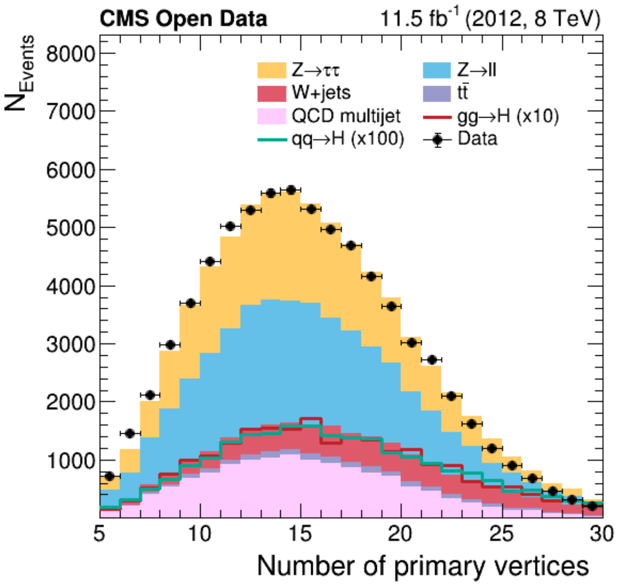

Figure 5.1: Observed primary vertice number distributions in the 8 TeV $\mu\tau_h$ channel.

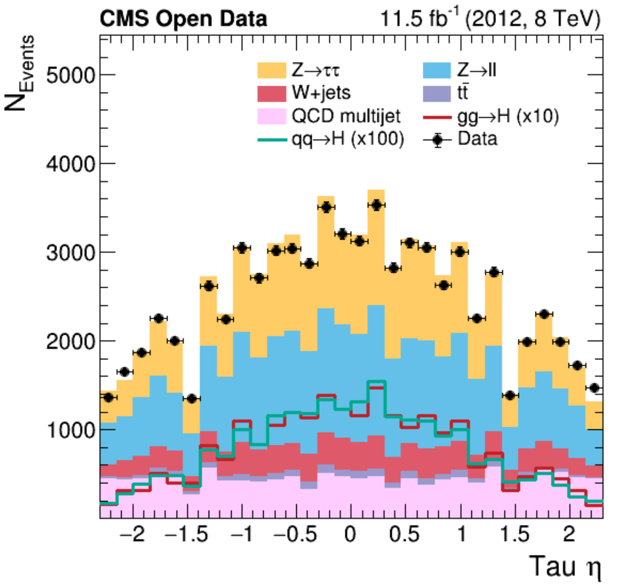

Figure 5.2: Observed tau pseudorapidity distributions in the 8 TeV $\mu\tau_h$ channel.

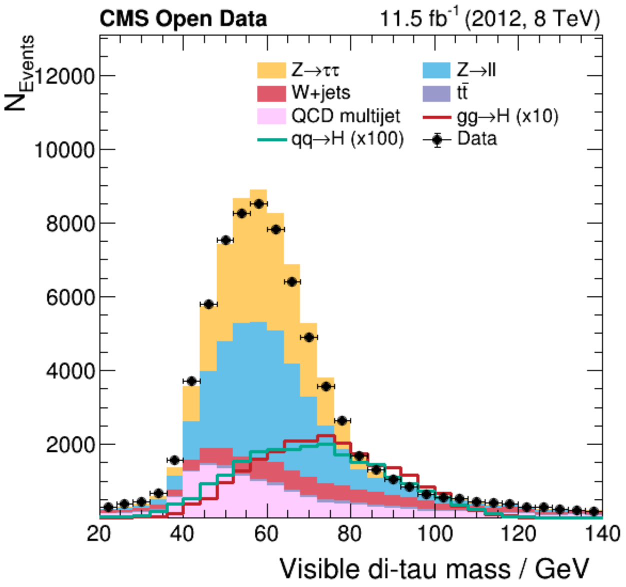

Figure 5.3: Observed visible $m_{\tau\tau}$ distributions in the 8 TeV $\mu\tau_h$ channel.

background) using the visible mass m_{vis} of the LL' system. This can be seen in Figure 5.3. However, the variable m_{vis} is limited, since the neutrinos coming from a τ lepton can take a large part of the energy and this energy is not taken into account. Neutrinos are identified as MET (missing transverse energy) objects. For this reason, the signals $gg \to H \to \tau\tau$ and $qq \to H \to \tau\tau$ are shifted to the left and are not centered at 125 GeV. Furthermore, these signals have been scaled to make these contributions visible.

Annex B contains the entire set of graphs obtained.

5.1 Summary and conclusions

A search is made for the standard model Higgs boson decaying into a pair of τ leptons, based on the official CMS analysis "Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons". The search uses nine proton-proton collision datasets recorded by CMS in 2012, together they contain 21185 AOD files, and correspond to an integrated luminosity of 11.1 fb^{-1} at a center of mass energy of 8 TeV. The analysis is performed on one of the six channels, corresponding to the final state $\mu\tau_h$. The gluon-gluon fusion and vector boson fusion are the main production modes of a Higgs boson. It was possible to reproduce the original analysis partially. This result constitutes a first look at the process of revealing the coupling between the τ lepton and the 125 GeV Higgs boson discovered in 2012 by the ATLAS and CMS collaborations.

The computational processing time during the AOD2NanoAOD step is 24 hours, which is quite good considering the number of root files that were processed, which is 21185 AOD files.

A complete physical analysis of data requires proper treatment of all uncertainties. This analysis does not study any systematic uncertainties. For this reason, the results presented are a qualitative interpretation of the observables used. In addition, the signals $gg \to H \to \tau\tau$ and $qq \to H \to \tau\tau$ were scaled for better visibility.

This analysis can be continued with an appropriate statistical study. This is often the longest part of a physical data study. First, the simulation must be improved with corrections to increase the precision of the data. Each correction must include systematic uncertainties. Second, measurements must be taken using a statistical data model, which must incorporate a physics model, such as the SM Standard Model, and statistical and systematic uncertainties.

Bibliography

[1] Weinberg, S. (1967). A model of leptons. Physical review letters, 19(21), 1264.

[2] Englert, F., & Brout, R. (1964). Broken symmetry and the mass of gauge vector mesons. Physical Review Letters, 13(9), 321.

[3] Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Khalek, S. A., Abdelalim, A. A., ... & Bansil, H. S. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Physics Letters B, 716(1), 1-29.

[4] CMS collaboration. (2012). Study of the mass and spin-parity of the Higgs boson candidate via its decays to Z boson pairs. arXiv preprint arXiv:1212.6639.

[5] Chatrchyan, S., Khachatryan, V., Sirunyan, A. M., Tumasyan, A., Adam, W., Bergauer, T., ... & Mundim, L. (2014). Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons. Journal of High Energy Physics, 2014(5), 1-72.

[6] Wunsch, S. Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012 (2019), 10.7483/OPENDATA. CMS. GV20. PR5T.

[7] Adolphi, R. (2008). The CMS experiment at the CERN LHC. Jinst, 803, S08004.

[8] Chatrchyan, S., Hmayakyan, G., Khachatryan, V., & CMS Collaboration. (2008). The CMS experiment at the CERN LHC. Journal of instrumentation, 3(8), S08004.

[9] Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proceedings of the IEEE, 86(11), 2210-2239.

[10] Cacciari, M., Salam, G. P., & Soyez, G. (2012). FastJet user manual. The European Physical Journal C, 72(3), 1-54.

[11] CMS collaboration. (2013). Identification of b-quark jets with the CMS experiment. Journal of Instrumentation, 8(04), P04013.

[12]]CMS Collaboration Collaboration, & Collaboration, C. (2013). Pileup Jet Identification. Tech. Rep. CMS-PAS-JME-13-005.

[13] Voss, H., Höcker, A., Stelzer, J., & Tegenfeldt, F. (2009, July). TMVA, the toolkit for multivariate data analysis with ROOT. In XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research (Vol. 50, p. 040). SISSA Medialab.

[14] CMS collaboration. (2012). Performance of CMS muon reconstruction in pp collision events at $\sqrt{s}$= 7 TeV. Journal of Instrumentation, 7(10), P10002.

[15] CMS collaboration. (2010). Electron reconstruction and identification at $\sqrt{s}$= 7 TeV. CMS-PAS-EGM-10-004.

[16] CMS collaboration. (2012). Performance of τ-lepton reconstruction and identification in CMS. Journal of Instrumentation, 7(01), P01001.

[17] Nason, P. (2004). A new method for combining NLO QCD with shower Monte Carlo algorithms. Journal of High Energy Physics, 2004(11), 040.

[18] Alwall, J., Herquet, M., Maltoni, F., Mattelaer, O., & Stelzer, T. (2011). MadGraph 5: going beyond. Journal of High Energy Physics, 2011(6), 128.

[19] CMS collaboration. (2012). CMS data preservation, re-use and open access policy. CERN Open Data Portal.

[20] CERN. (n.d). The Grid: A system of tiers. CERN Accelerating science. https://home. cern/about/computing/grid-system-tiers

[21] Gaillard, M. (2017). CERN Data Centre passes the 200-petabyte milestone.

[22] Lassila-Perini, K., Alverson, G., Cabrillo, I., Calderon, A., Colling, D., Hildreth, M., ... & Sonnenschein, L. (2014, June). Implementing the data preservation and open access policy in CMS. In Journal of Physics: Conference Series (Vol. 513, No. 4, p. 042029). IOP Publishing.

[23] CERN. (2018). CERN Open Data Portal. Opendata CERN. http://opendata. cern.ch/docs/about

[24] Wunsch, S. (2019). cms-opendata-analyses / AOD2NanoAODOutreachTool. Github. https://github.com/cms-opendata-analyses/AOD2NanoAODOutreachTool

[25] Bisong, E. (2019). Building machine learning and deep learning models on Google cloud platform (pp. 7-10). Berkeley: Apress.

[26] Burns, B., Beda, J., & Hightower, K. (2018). Kubernetes. Dpunkt.

[27] ATLAS and CMS Collaborations, LHC Higgs Combination Group. (2011). Procedure for the LHC Higgs boson search combination in Summer 2011, Technical Report ATL-PHYS-PUB 2011-11, CMS NOTE 2011/005, CERN.l

[28] d'Enterria, D., Ballintijn, M., Bedjidian, M., Hofman, D., Kodolova, O., Loizides, C., ... & CMS Collaboration. (2007). CMS physics technical design report: Addendum on high density QCD with heavy ions. Journal of Physics G: Nuclear and Particle Physics, 34(11), 2307.

Annexes

Annex A: Steps to perform the AOD2NanoAOD step in a Kubernetes cluster.

- First, go to Google cloud platform, and to the kubernetes engine section. https://console.cloud.google.com/kubernetes/
- Create a cluster with the default configuration, with 6 nodes of e2-standard-4 type. This means each node has 4 CPUs and 16 GB of RAM.
- Once the cluster is ready, start the Cloud SHEll, this is the online Google terminal, and login with a Google account using the following command:

```
gcloud auth login
```

- Establish connection with the cluster.
- Run the following command lines to install argo in the cluster. To manage the pods in parallel, It is used the Argo Workflows tool. This helps to easily manage a workflow in kubernetes.

```
kubectl create ns argo
kubectl apply -n argo -f https://raw.githubusercontent.com/argoproj/argo/
stable/manifests/quick-start-postgres.yaml
curl -sLO https://github.com/argoproj/argo/releases/download/v2.11.1/
argo-linux-amd64.gz
gunzip argo-linux-amd64.gz
chmod +x argo-linux-amd64
sudo mv ./argo-linux-amd64 /usr/local/bin/argo
```

- Give Argo sufficient rights to manage a workflow.

```
kubectl create clusterrolebinding YOURNAME-cluster-admin-binding
--clusterrole=cluster-admin --user=YOURMAIL
```

- Apply a small patch to the default argo config.

```
kubectl patch configmap workflow-controller-configmap -n argo --patch
"$(cat patch-workflow-controller-configmap.yaml)"
```

- Creating a disk to store workflow output on Google Kubernetes Engine.

```
gcloud compute disks create --size=300GB --zone=us-central1-c
gce-nfs-disk-1
```

- Create the PersistentVolume and the PersistentVolumeClaim with the right cluster IP (to get the cluster IP we can use the fourth command line).

```
kubectl apply -n argo -f 001-nfs-server.yaml
kubectl apply -n argo -f 002-nfs-server-service.yaml
kubectl apply -n argo -f 003-pv-pvc.yaml
kubectl get -n argo svc nfs-server |grep ClusterIP | awk ’{ print $3; }’
```

- Now, submit each workflow (for nine datasets) one by one.

```
argo submit -n argo WORKFLOW.yaml
```

- Merge the output files into the nine NanoAod datasets.

```
argo submit -n argo nanoaod-merge.yaml
```

- in order to access this files, we can mount a public IP. First, patch the config of the webserver.

```
kubectl create configmap basic-config --from-file=conf.d
```

- Deploy the fileserver: apply and expose the port as a LoadBalancer.

```
kubectl create -f deployment-http-fileserver.yaml
kubectl expose deployment http-fileserver --type LoadBalancer --port 80
--target-port 80
```

- Follow its status until get the EXTERNAL-IP

```
kubectl get svc
```

- Delete the index.html file in the pod in order to enable file browsing.

```
kubectl get pods
kubectl exec http-fileserver-XXXXXXXX-YYYYY -- rm
/usr/share/nginx/html/index.html
```

- Finally, Download the nine NanoAOD files from the external IP.

These steps were completed in a 26 hour time interval. Note that all configuration and workflow files used are in this github repository:

https://github.com/asdru30/CMSParallelJobKubernetes

Annex B: Complete set of graphs obtained.

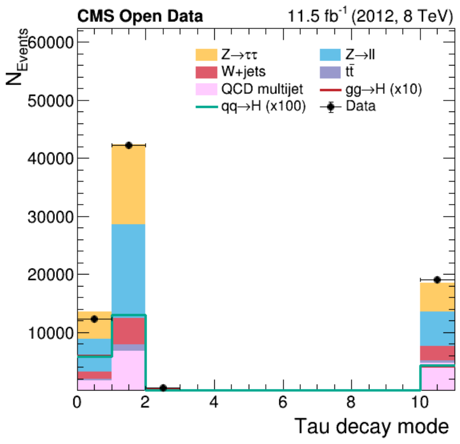

Figure 5.4: Observed τ decay mode distributions in the 8 TeV μτh channel.

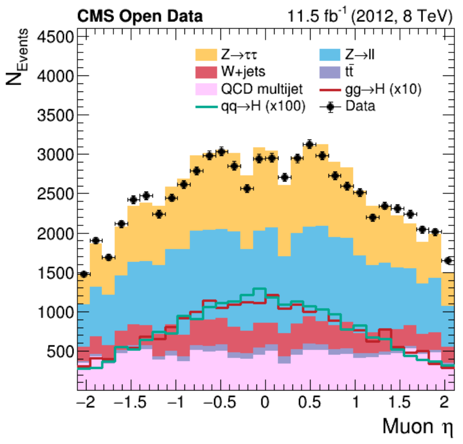

Figure 5.5: Observed muon pseudorapidity distributions in the 8 TeV μτh channel.

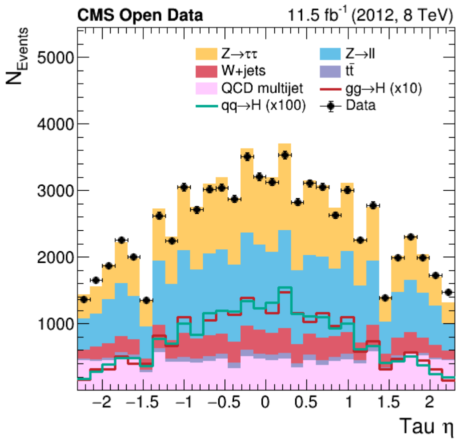

Figure 5.6: Observed τ pseudorapidity distributions in the 8 TeV μτh channel.

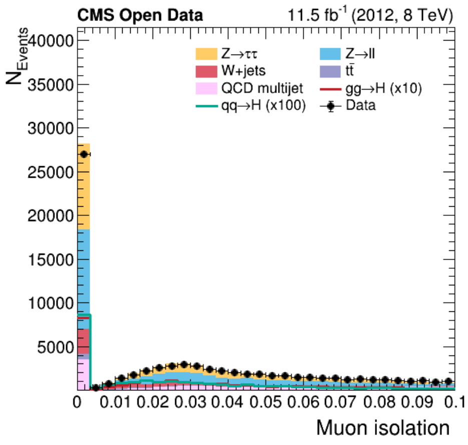

Figure 5.7: Observed muon isolation distributions in the 8 TeV μτh channel.

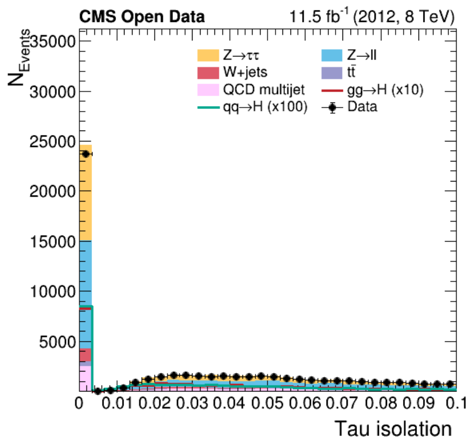

Figure 5.8: Observed τ isolation distributions in the 8 TeV µτh channel.

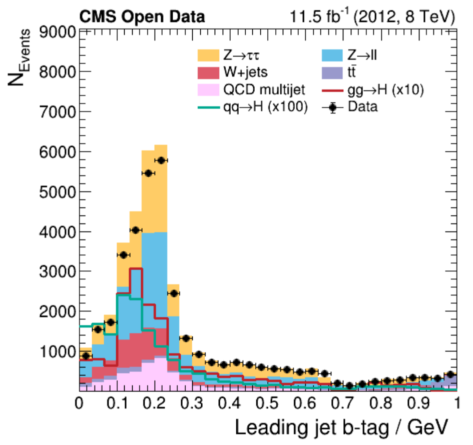

Figure 5.9: Observed leading jet b-tag distributions in the 8 TeV µτh channel.

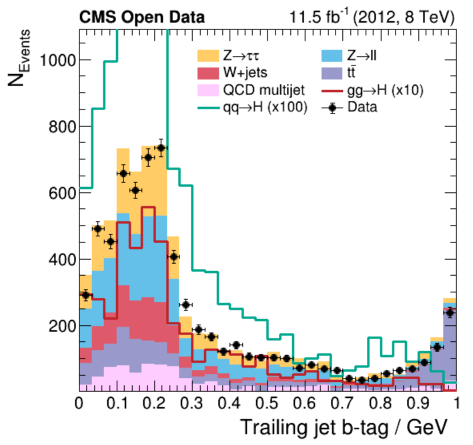

Figure 5.10: Observed trailing jet b-tag distributions in the 8 TeV µτh channel.

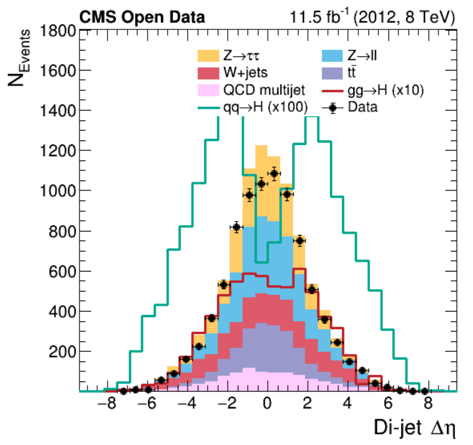

Figure 5.11: Observed di-jet $\triangle\eta$ distributions in the 8 TeV µτh channel.

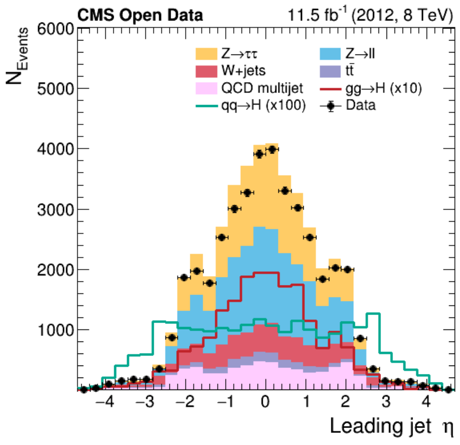

Figure 5.12: Observed leading jet pseudorapidity distributions in the 8 TeV μτh channel.

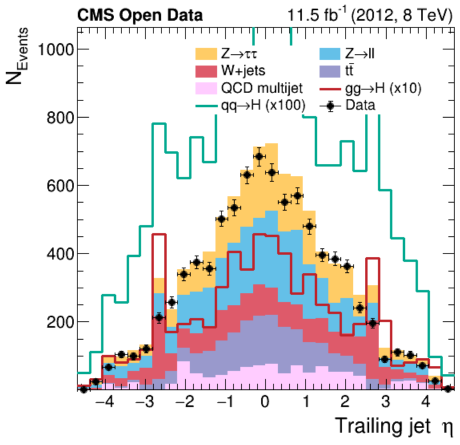

Figure 5.13: Observed trailing jet pseudorapidity distributions in the 8 TeV μτh channel.

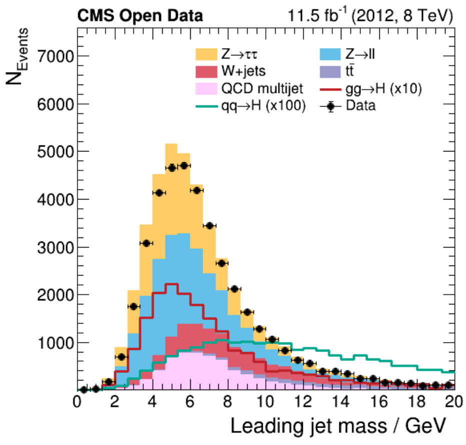

Figure 5.14: Observed leading jet mass distributions in the 8 TeV μτh channel.

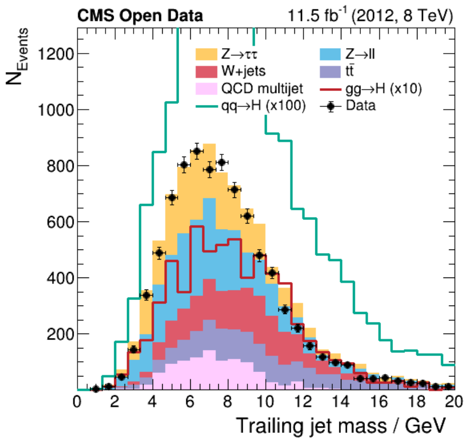

Figure 5.15: Observed trailing jet mass distributions in the 8 TeV μτh channel.

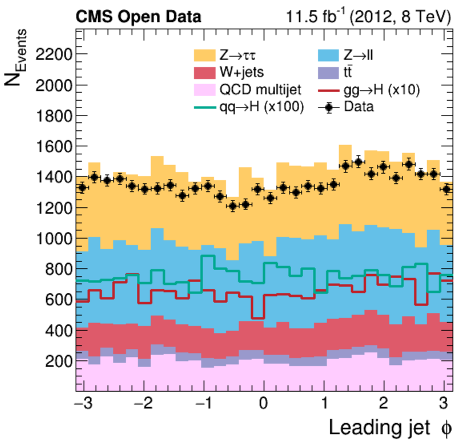

Figure 5.16: Observed leading jet ϕ distributions in the 8 TeV µτh channel.

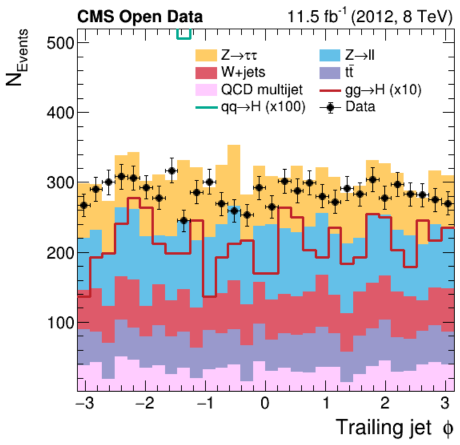

Figure 5.17: Observed trailing jet ϕ distributions in the 8 TeV µτh channel.

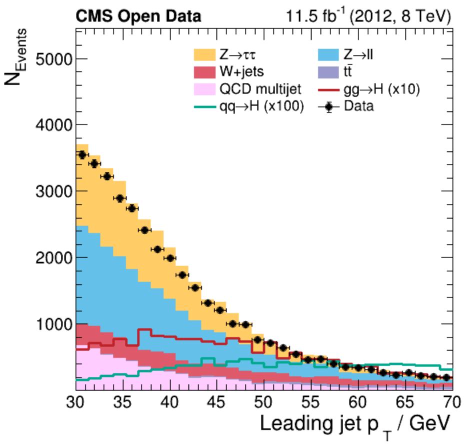

Figure 5.18: Observed leading jet p_T distributions in the 8 TeV µτh channel.

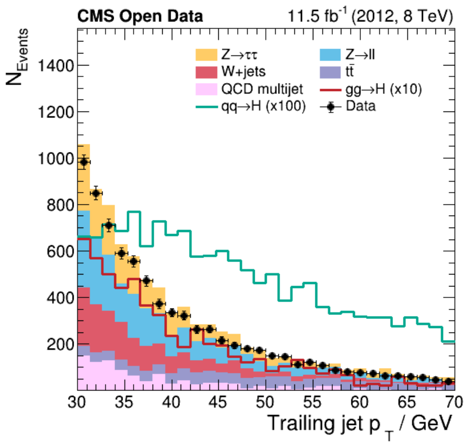

Figure 5.19: Observed trailing jet p_T distributions in the 8 TeV µτh channel.

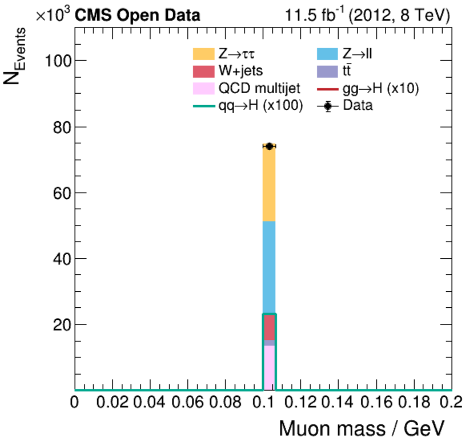

Figure 5.20: Observed muon mass distributions in the 8 TeV µτh channel.

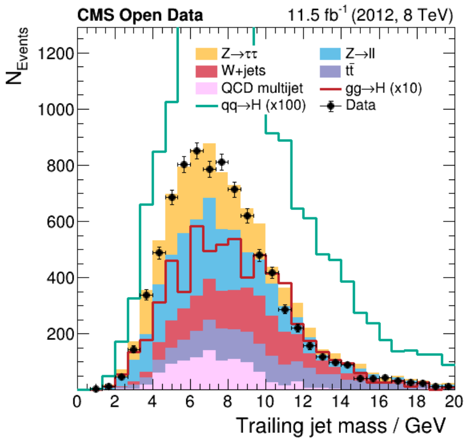

Figure 5.21: Observed τ mass distributions in the 8 TeV µτh channel.

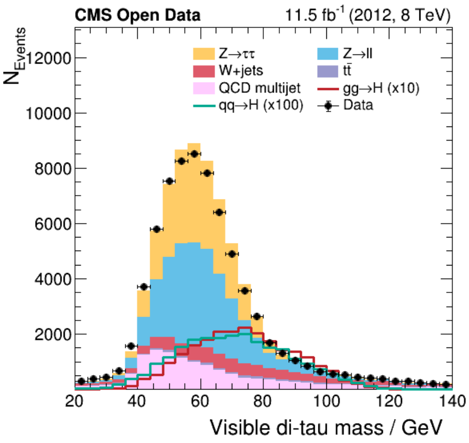

Figure 5.22: Observed visible di-τ mass distributions in the 8 TeV µτh channel.

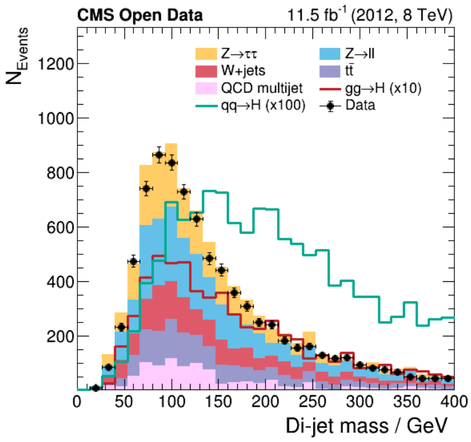

Figure 5.23: Observed di-jet mass distributions in the 8 TeV µτh channel.

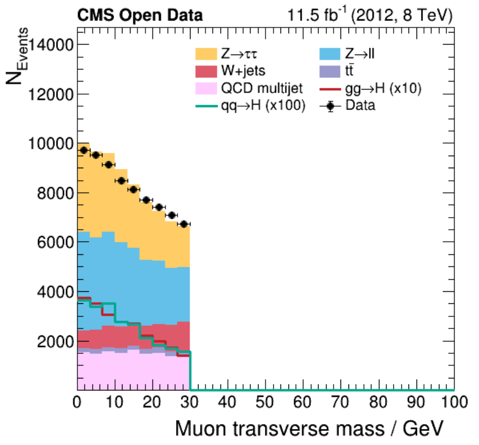

Figure 5.24: Observed muon transverse mass distributions in the 8 TeV μτh channel.

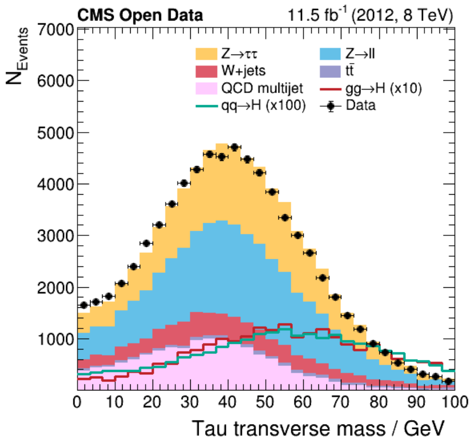

Figure 5.25: Observed τ transverse mass distributions in the 8 TeV μτh channel.

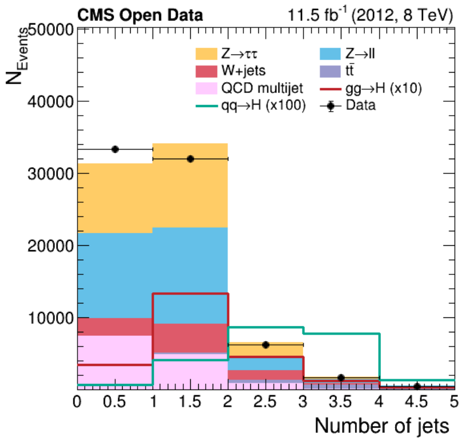

Figure 5.26: Observed jet number distributions in the 8 TeV μτh channel.

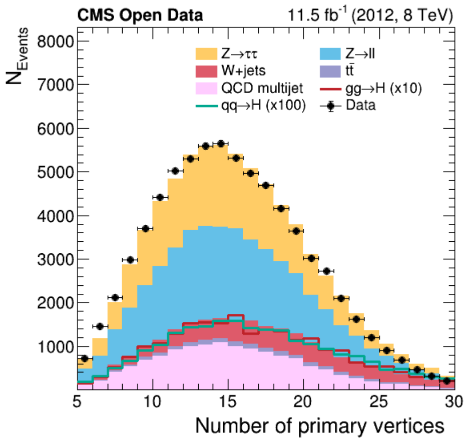

Figure 5.27: Observed primary vertice number distributions in the 8 TeV μτh channel.

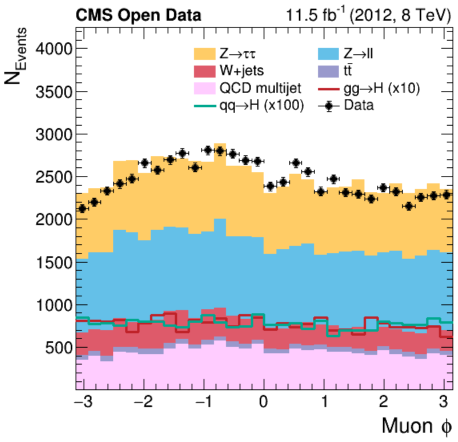

Figure 5.28: Observed muon ϕ distributions in the 8 TeV μτh channel.

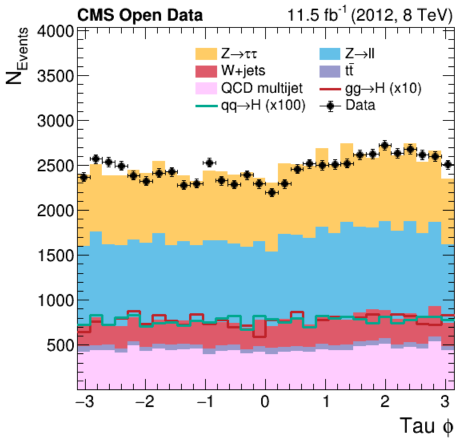

Figure 5.29: Observed tau ϕ distributions in the 8 TeV μτh channel.

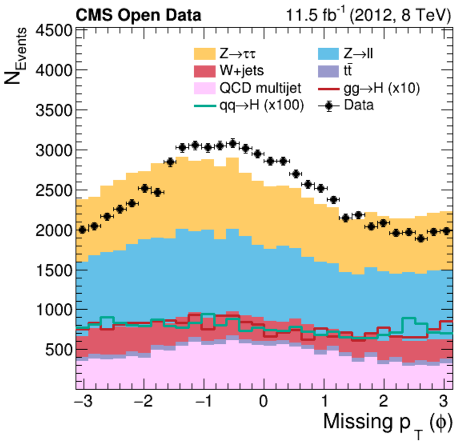

Figure 5.30: Observed missing transverse momenta distributions (as a function of ϕ) in the 8 TeV μτh channel.

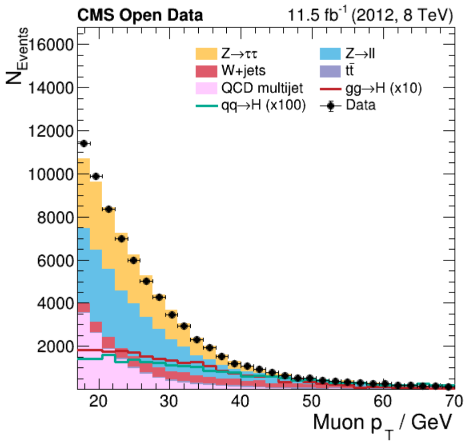

Figure 5.31: Observed muon transverse momenta distributions in the 8 TeV μτh channel.

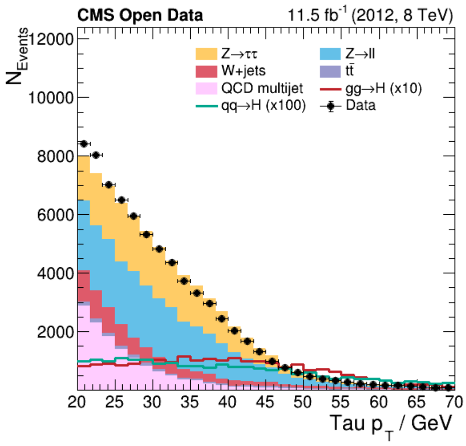

Figure 5.32: Observed τ transverse momenta distributions in the 8 TeV μτh channel.

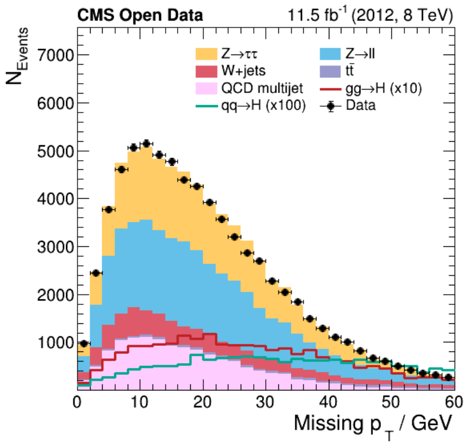

Figure 5.33: Observed missing transverse momenta distributions in the 8 TeV μτh channel.

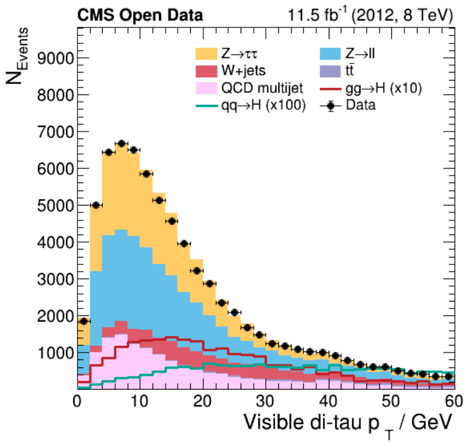

Figure 5.34: Observed visible di-τ transverse momenta distributions in the 8 TeV μτh channel.

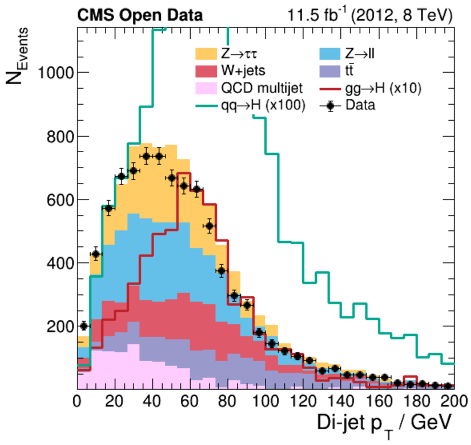

Figure 5.35: Observed di-jet transverse momenta distributions in the 8 TeV μτh channel.

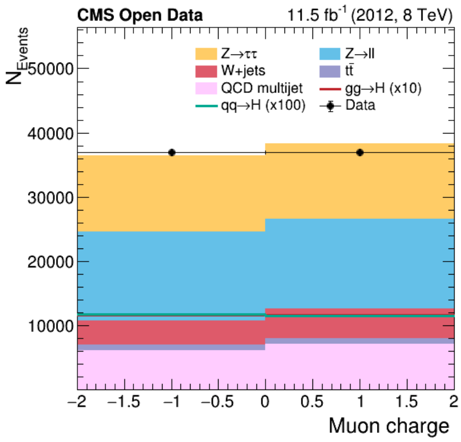

Figure 5.36: Observed muon charge distributions in the 8 TeV μτh channel.

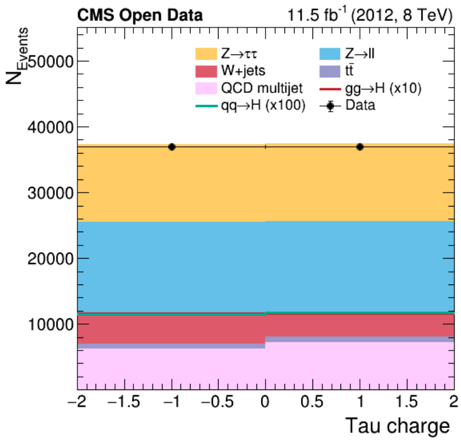

Figure 5.37: Observed τ charge distributions in the 8 TeV μτh channel.