Summer 2017

A search for the standard model Higgs Boson in the μ+μ- decay channel in PP collisions at √s=13 TeV with CMS, calibration of CMS Hadron forward calorimeter, and simulations of modern calorimeter systems

Viktor Khristenko
University of Iowa

A SEARCH FOR THE STANDARD MODEL HIGGS BOSON IN THE $\mu^+\mu^-$ DECAY CHANNEL IN PP COLLISIONS AT $\sqrt{S} = 13$ TEV WITH CMS, CALIBRATION OF CMS HADRON FORWARD CALORIMETER, AND SIMULATIONS OF MODERN CALORIMETER SYSTEMS

by

Viktor Khristenko

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Physics
in the Graduate College of
The University of Iowa

August 2017

Thesis Supervisor: Professor Yasar Onel

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Viktor Khristenko

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Physics at the August 2017 graduation.

Thesis Committee: _____
Yasar Onel, Thesis Supervisor

_____
Jane Nachtman

_____
Burak Bilki

_____
Darin Acosta

_____
Craig Pryor

_____
Markus Wohlgenannt

To my family, for the possibility to be standing where I am.

It is better to be than seem to be.

Valentin Pikul

ACKNOWLEDGEMENTS

First of all, I would like to express my deep gratitude to my advisor, professor Yasar Onel, for the possibility to be his student. In particular, I want to note the amount of experience, knowledge and connections I have made by staying at CERN for the past three years. It has been an incredible experience and would not be possible without his continuous support and advice.

I would like to thank Ugur Akgun for introducing me into the world of High Energy Physics. For me it has all started as a summer undergraduate student working on various detector simulations. My time at the University of Iowa and at CERN would not be the same without James Wetzel and David Southwick. From arguing about current politics and going out in Geneva to the actual physics discussions - it has been the one experience I will never forget.

I would like to thank Alexi Mestverishvili, who has made the best soup I have ever eaten in my life. For the friendship and kindness that he has shown to a poor student coming to Geneva to work on Compact Muon Solenoid Experiment.

I would like to thank one of the best detector physicists I have ever met - Burak Bilki, who did not just teach me how to perform simulations of calorimeters and set up a trigger system for the Fermilab Test Beam Experiment, but who has also been a good friend!

I would like to acknowledge professor Jane Nachtman, Ianos Schmidt, Paul Debbins and the rest of the Iowa group for the numerous discussions and help I have

received during my research.

I would like thank professor Darin Acosta for joining my committee and recognize my colleagues from the Higgs group: Darin Acosta, Sergei Gleyzer, Adrian Perieanu, Andrew Brinkerhoff, Andrea Marini, Andrew Carnes, Pierliugi Bortignon. I would like to acknowledge them not just for the possibility to join this wonderful analysis team, but, what is more important to me, it is through the interaction with these people that I have been able to finally understand, appreciate and enjoy the process of doing a physics analysis.

I would like to thank professors Craig Pryor and Markus Wohlgenannt, who agreed to join my thesis committee, provide continuous support for my research program and for overall cooperation with me on various logistics matters.

I would like to share the pleasure I have had to be part of the Physics Department at the University of Iowa for, first of, the opportunity to just be a member of this incredible group of people; and, secondly, for the support of the whole staff over these years. That has made a fundamental difference in my experience as a student.

I would like to acknowledge members of the HCAL personnel for the possibility of working side-by-side on such an exciting project, learning directly from the most experienced people in the field, and for the amazing barbecue parties organized over the years.

I would like to thank my undergraduate physics professors Steve Feller and Mario Affatigato for running Coe College Physics program as a family business!

Going across the globe, I want to recognize people who taught me not just how

to program or to solve differential equations, but how to actually think rationally and how to approach solving various problems, most importantly in life! I want to thank my high school computer science and mathematics teachers Konstantin Krivonogov and Vladimir Tupin.

I would like to thank my friends in Iowa: Eric Rodgers; Kris and Phil Diehl; Steve Eden; Marina Stankovich and Gary Dixon; the Postnikov family - for the friendship that I will try to cherish and carry over with me through my life.

Finally, I would like to mention my friends and family from Russia - my parents, uncles and aunts, my grandparents, my younger brother, my cousins, my best friends, my girlfriend Kate - who taught me how to enjoy simple things. These are the people who have always been behind my back and supported me to keep moving forward.

ABSTRACT

A search for the Standard Model Higgs Boson decaying to two muons in proton-proton collisions with the Compact Muon Solenoid experiment is performed. Building on top of the success of previous CMS analyses (CMS Run I campaign), results are presented using 35.9 fb^{-1} of data collected over the course of 2016 (CMS Run II campaign) at a center-of-mass energy of $\sqrt{s} = 13$ TeV.

During the Long Shutdown 1 of the Large Hadron Collider, the CMS detector underwent substantial hardware changes. The second topic discusses the process of calibration of the CMS Hadron Forward Calorimeter in preparation for collisions after LS1.

The final chapter discusses the process of building simulations of calorimeter systems. Walking through all the steps from geometry specification to readout definition the results for two standalone calorimeters are presented that have been proposed as potential replacements for respective CMS components.

PUBLIC ABSTRACT

Higgs Boson is the last missing piece of the most successful theoretical model of the elementary particle physics - Standard Model. The construction of the Large Hadron Collider and the experiments hosted by it was optimized with the search for the Higgs Boson in mind. The first part of this work focuses on trying to find the Higgs Boson in a dimuon decay channel with the Compact Muon Solenoid (CMS) detector.

Calorimeter is a device for measuring energy. In High Energy Physics, calorimeters are responsible for determining the energy of the particles passing through. Like any other measuring system, it needs to be calibrated to associate properly its response to the actual energy units. Second part discusses the calibration process of the Hadron Forward calorimeter of the CMS experiment.

Final chapter looks at simulations of calorimeters proposed for the future upgrades. Two different systems are built and evaluated. The focus is on obtaining basic performance characteristics.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Before moving on to the research topics of this thesis, here in the first chapter a brief outline of the structure of the thesis is discussed and a very general overview of the major topics is presented.

Modern Experimental High Energy Physics (HEP) is a science at scale from several points of view: the amount of money it requires to have invested, the physical dimensions of instruments that are built and number of people that collaborate on the same project. It is a unique field that brings in skills across various domains and provides a unique opportunity to get involved in several independent projects that, in the end, come together as a single scientific result. This thesis is no different - the topics discussed are completely independent and fall into separate categories, however they fill the phase space of the research program for the Compact Muon Solenoid Experiment.

In the second chapter, the results of the search for the Standard Model Higgs Boson decaying via two opposite-sign muons are presented. The analysis performed is an extension of previous work that has been performed earlier with the CMS Experiment during Run I campaign, however the new results benefit from increased center of mass energy and higher instantaneous luminosity.

In the second part, the CMS Hadron Forward (HF) Calorimeter will be introduced and the calibration procedure using a radioactive source will be discussed. During the Long Shutdown 1 (LS1) of the Large Hadron Collider (LHC), the HF

underwent a substantial electronics upgrade which required the energy scale of the calorimeter to be reestablished. The details of the experimental setup and analysis workflow will be presented. The success of the presented calibration technique and results has been demonstrated by the success of the CMS Run II data-taking and analysis campaigns.

The final chapter focuses on building simulations of calorimeters, evaluating basic performance characteristics and scaling them up to the production dimensions of a future experiment. The focus will be on two particular examples: High Granularity Calorimeter (HGC) and Shashlik plus Hadron Endcap System. Both systems were potential candidates for the future upgrades of the CMS Endcap and constitute ideal benchmarks for our simulations.

CHAPTER 2
A SEARCH FOR THE STANDARD MODEL HIGGS BOSON

2.1 Introduction and Motivation

The Standard Model is the leading theoretical description of elementary particle physics. It incorporates strong, electromagnetic and weak fundamental forces of nature and explains the basic constituents of matter. Very detailed treatment and specification of the Standard Model can be found in references [1, 2, 3]. In what follows, the answers to the following two questions are provided in a simplified and informal manner:

- What the Higgs Boson is and why the search is performed.
- Why probing the Higgs Boson decaying to two muons is important.

The Standard Model categorizes all of the elementary particles into two groups: bosons and fermions. Bosons with spin 1 (photons, gluons, $W^{\pm}$ and Z^0) represent the force carriers together with the Higgs Boson, with spin 0. Fermions are the main building blocks of matter and can be further subdivided into two subgroups: six quarks and six leptons. Quarks possess the strong "charge" (color) and, therefore, participate in the strong interaction and couple to the strong force carriers (gluons). On the other hand, leptons are colorless and therefore do not interact with gluons. The six leptons can be specified in the following way:

- Electron + electron neutrino
- Muon + muon neutrino

- Tau + tau neutrino

Neutrinos do not possess any electric charge, therefore they only participate in weak interactions, mediated by the $W^{\pm}$ and Z^0 bosons.

Mathematically, the Standard Model is a local gauge invariant Quantum Field Theory (QFT). Similarly to the Classical theory, it follows the Lagrangian formulation, by constructing a Lagrangian density and requiring it to satisfy a set of symmetries. To begin with, consider a Lorentz invariant Lagrangian density for a free 4-component Dirac Spinor $\psi(x)$ that describes a fermion:

$$\mathcal{L} = \bar{\psi} i\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi \tag{2.1}$$

and gauge transformations for the fermion field:

$$Global: \quad \psi \rightarrow e^{i\theta}\psi \tag{2.2a}$$

$$Local: \quad \psi \rightarrow e^{i\theta(x)}\psi \tag{2.2b}$$

where $\theta(x)$ is a function of space-time in Equation 2.2b. Now, for the case of the global transformation of the fermion field, it follows trivially that $\mathcal{L}$ remains unchanged. However, for the case of the local gauge transformation, the symmetry of the $\mathcal{L}$ does not hold:

$$\begin{aligned} \mathcal{L} &= \bar{\psi} e^{-i\theta(x)} i\gamma^{\mu}\partial_{\mu}(e^{i\theta(x)}\psi) - m\bar{\psi}\psi \\ &= \bar{\psi} e^{-i\theta(x)} i\gamma^{\mu}[ie^{i\theta(x)}\partial_{\mu}\theta(x) + e^{i\theta(x)}\partial_{\mu}]\psi - m\bar{\psi}\psi \\ &= \bar{\psi} i\gamma^{\mu}[\partial_{\mu} + i\partial_{\mu}\theta(x)]\psi - m\bar{\psi}\psi \end{aligned} \tag{2.3}$$

To remedy that, two additional items are defined; first, a new vector field is introduced, A^{μ}, which under the local gauge transformations, defined in Equation 2.2b,

transforms according to the Equation 2.4a. Second, in Equation 2.4b, the partial derivative is modified to accommodate the new vector field.

$$A^\mu \to A^\mu - \frac{1}{q}\partial^\mu\theta(x) \tag{2.4a}$$

$$D^\mu = \partial^\mu + iqA^\mu \tag{2.4b}$$

With the above definitions in mind, Equation 2.1 reduces down to the local gauge invariant definition of the $\mathcal{L}$:

$$\begin{aligned}
\mathcal{L} &= \bar{\psi}i\gamma^\mu[\partial_\mu + i\partial_\mu\theta(x)]\psi - m\bar{\psi}\psi \\
&\to \bar{\psi}i\gamma^\mu[\partial_\mu + iq(A_\mu - \frac{1}{q}\partial_\mu\theta(x)) + i\partial_\mu\theta(x)]\psi - m\bar{\psi}\psi \\
&= \bar{\psi}i\gamma^\mu[\partial_\mu + iqA_\mu]\psi - m\bar{\psi}\psi \\
&= \bar{\psi}i\gamma^\mu D_\mu\psi - m\bar{\psi}\psi
\end{aligned} \tag{2.5}$$

The obtained Lagrangian density describes a Dirac fermion in interaction with a vector field (the newly defined derivative hides the interaction term). To complete the picture, kinetic and mass terms are added into the Lagrangian:

$$\mathcal{L} = \bar{\psi}i\gamma^\mu D_\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} + m_A^2A^\mu A_\mu \tag{2.6}$$

The last step here is to note that the mass term of the introduced vector field does not satisfy the local gauge invariance under consideration:

$$\begin{aligned}
A^\mu A_\mu &\to (A^\mu - \partial^\mu\theta(x))(A_\mu - \partial_\mu\theta(x)) \\
&\to A^\mu A_\mu + \dots \\
&\neq A^\mu A_\mu
\end{aligned} \tag{2.7}$$

Therefore, the conclusion is that requiring local gauge invariance (Equation 2.2b) for a Dirac field results in the introduction of a massless vector field with the corresponding Lagrangian:

$$\mathcal{L} = \bar{\psi} i \gamma^{\mu} D_{\mu} \psi - m \bar{\psi} \psi - \frac{1}{4} F^{\mu\nu} F_{\mu\nu} \tag{2.8}$$

The theory described by the above Lagrangian density is called Quantum Electrodynamics (QED) and explains interactions between fermions and photons, electromagnetic force carries. It is important to point out that experimental observations confirm the fact that photons are massless, as is required by the QED. The symmetry that was imposed on the initial free $\mathcal{L}$ is the simplest possible form; $e^{i\theta(x)}$ is a complex scalar. QED is often called U(1) theory, precisely because a set of local gauge transformations, defined in Equation 2.2b, forms a U(1) group of unitary transformations.

The full group specification of the Standard Model is $\mathrm{SU}(3) \times \mathrm{SU}(2)_L \times \mathrm{U}(1)$. It is important to note that all of these groups define the transformations, application of which should leave our Lagrangian invariant, just like it was done above for QED. The guiding principle to generate the rest of the gauge bosons (for $\mathrm{SU}(3) \times \mathrm{SU}(2)$) is similar to U(1). Therefore, briefly consider the Yang-Mills theory (generated by the SU(2) group) which allows to introduce the weak bosons. SU(2) can be realized as a group of 2-dimensional unitary matrices with a determinant of 1. The corresponding

gauge transformations will then be defined as follows:

$$Global: \quad \Psi \rightarrow e^{i\boldsymbol{\theta}\cdot\boldsymbol{\sigma}}\Psi \tag{2.9a}$$

$$Local: \quad \Psi \rightarrow e^{i\boldsymbol{\theta}(\boldsymbol{x})\cdot\boldsymbol{\sigma}}\Psi \tag{2.9b}$$

where σ_i are the Pauli matrices. Define the Ψ to be a **doublet of Dirac Spinors**, which in the most general case can be mixed, therefore the Lagrangian for the free Ψ doublet will be (omitting the mass term):

$$\mathcal{L} = \bar{\Psi} i\gamma^{\mu}\partial_{\mu}\Psi \tag{2.10}$$

As for U(1), require the $\mathcal{L}$ to be invariant under local SU(2) gauge transformations. The result will be the introduction of three additional massless vector fields (according to the number of generators of the SU(2) group), which under local SU(2) symmetry group transform very similar to Equations 2.4. The differences are the direct consequences of the fact that SU(2) transformations are not commutative.

Now, first, define the left-handed and right-handed Dirac spinor components to be:

$$\psi_L = \frac{1}{2}(1-\gamma^5)\psi \tag{2.11a}$$

$$\psi_R = \frac{1}{2}(1+\gamma^5)\psi \tag{2.11b}$$

$$\tag{2.11c}$$

Then using the above definitions and the fact that $(\gamma^5)^2 = 1$, the fermion mass term can be rewritten as:

$$\bar{\psi}\psi = m\bar{\psi_L}\psi_R + m\bar{\psi_R}\psi_L \tag{2.12}$$

The point of the above factorization is that the Standard Model Electroweak theory imposes the invariance under local SU(2) gauge transformations, defined in Equation 2.9b, only for the left-handed spinors (the subscript of $SU(2)_L$ is there to signal this property of the SM). Therefore, the mass terms for fermions break $SU(2)_L$ local symmetry and, as a consequence, fermions should be massless. Moreover, the decomposition in Equation 2.12 is mathematically unsound, because ψ_L transforms according to SU(2) and ψ_R according to U(1).

At this point, the main consequences of the Electroweak part of the Standard Model, $SU(2)_L \times U(1)$ are manifested in:

- Fermions are massless, contrary to the observations
- Four gauge bosons (consider Electroweak bosons: γ, $W^{\pm}$ and Z^0) are massless, again contrary to the experimental observations where $W^{\pm}$ and Z^0 are massive.

The Higgs Mechanism is the approach to generate masses for the three gauge bosons and fermions by introducing a new field that lives in $SU(2)_L \times U(1)$ space into the Lagrangian with a particular choice of the potential function (factor out the SU(3), QCD theory). Therefore, start with a Lagrangian for a massless fermion with the electroweak interaction terms hidden inside the derivative matrix, D^{μ}, and introduce a new $SU(2)_L \times U(1)$ left-handed doublet, components of which are spin-0

complex fields, in the following way to preserve the original local gauge symmetry:

$$V(\Phi) = \mu^2\Phi^\dagger\Phi + \lambda(\Phi^\dagger\Phi)^2 \tag{2.13a}$$

$$\mathcal{L} = \mathcal{L}_{fermion} + \mathcal{L}_{new} \tag{2.13b}$$

$$= \bar{\Psi} i\gamma^\mu D_\mu \Psi + (D^\mu\Phi)^\dagger D_\mu\Phi - \mu^2\Phi^\dagger\Phi - \lambda(\Phi^\dagger\Phi)^2 \tag{2.13c}$$

For the specified potential function, the $\Phi = 0$ is no longer a minimum, but a continuous spectrum of ground states corresponding to the vacuum with non-vanishing expectation value is observed. From this point, the procedure to generate gauge boson masses is:

- Φ is a doublet of complex spin-0 fields $\rightarrow$ has four degrees of freedom (ϕ_1, ϕ_2, ϕ_3, ϕ_4)
- Perturb ϕ around a particular choice of vacuum: $\phi_4 = \sqrt{-\frac{\mu^2}{\lambda}} + h$, with the rest of components unchanged.
- Expand the $\mathcal{L}$ and select a particular gauge: unitary gauge.
- In unitary gauge, ϕ_1, ϕ_2 and ϕ_3 components of Φ will vanish, however the mass terms for the bosons will be recovered. All of these terms are generated from the kinetic Φ term.
- The perturbation field, h, around the selected vacuum is the real field with the mass $m_H = \sqrt{-2\mu^2}$ - Higgs field.

This procedure is called the Spontaneous Symmetry Breaking of the local $\mathrm{SU}(2)_L \times \mathrm{U}(1)$ gauge invariance, because, although the original Lagrangian in Equations 2.13 stays invariant under the symmetry transformation, the rotation in $\mathrm{SU}(2)_L \times \mathrm{U}(1)$,

the ground state of the system (a particular choice of vacuum) does change to a different state with the same energy.

So far, only masses for weak bosons were shown to be recovered by the introduction of the Higgs field, with a particular choice of vacuum. These terms come out of expanding the kinetic ϕ terms perturbatively around some ground state. In addition, terms involving interactions between the gauge bosons and Higgs are also generated at the same time and from the same kinetic expansion. In order to incorporate the interactions of the Higgs field with fermions, consider to add the following term to the previous $\mathrm{SU}(2)_L \times \mathrm{U}(1)$ Lagrangian:

$$\mathcal{L} = -y[\bar{\psi_L}\phi\psi_R + \bar{\psi_R}\bar{\phi}\psi_L] \tag{2.14a}$$

Remember that ϕ lives in the $\mathrm{SU}(2)_L \times \mathrm{U}(1)$ and therefore adding it will exactly contract the doublet indices and preserve the local symmetry (assuming the introduced field carries the right quantum numbers). Expanding ϕ around one of the vacuum states (applying the Higgs mechanism) will generate the fermion mass and the coupling of the Higgs real field to two Dirac fields:

$$m\bar{\psi}\psi \rightarrow y[\bar{\psi_L}\phi\psi_R + \bar{\psi_R}\bar{\phi}\psi_L] \tag{2.15a}$$

$$\rightarrow y\sqrt{-\frac{\mu^2}{2\lambda}}\bar{\psi}_{Dirac}\psi_{Dirac} + \frac{y}{\sqrt{2}}H\bar{\psi}_{Dirac}\psi_{Dirac} \tag{2.15b}$$

where ψ_{Dirac} is to emphasize that they are pure four-component Dirac spinors. The constant, y, is the Yukawa coupling.

The Higgs field is the fundamental aspect of the Standard Model, which would fall short explaining the experimental observations without it. Therefore, the search

for the introduced boson is of primary importance for the validity of the Standard Model. The search for the decay of the Higgs into muons is motivated by the following factors:

- Dimuon decay allows to test the Yukawa coupling constant to the 2^{nd} generation of fermions.
- At LHC, that is the only way to measure or constrain the Yukawa coupling for this generation.
- Branching fraction is small but testable, 0.00022 for 125 GeV Higgs boson, however it is much smaller for the 1^{st} generation of fermions which will remain untestable at LHC.

2.2 Compact Muon Solenoid Experiment

The aim of this section is to introduce the LHC Machine and the CMS detector by briefly reviewing the components that are pertinent to the analysis performed. Very detailed documents outlining the design and architecture of both the CMS and LHC can be found in the CMS Technical Design Report [4, 5] and LHC machine specification [6].

2.2.1 The LHC Machine

The Large Hadron Collider (LHC) is a particle accelerating and colliding complex located on the border between France and Switzerland. Figure 2.3 shows the geographical overview of the location of LHC and of several experiments that are currently hosted there: ATLAS [7], CMS, LHCb [8] and ALICE [9]. Each of these

experiments is located at the beam crossing point along the ring.

Figure 2.1: LHC Accelerator Complex.

The main constituent of the LHC is a 27 km circular tunnel located about 100 m underground and equipped with thousands of various superconducting dipole and

quadrupole electromagnets. The operating temperature of the collider electromagnets is about -271°C, which is maintained by a special liquid helium based cryogenic system. Superconducting magnets are used to focus the beam along the ring towards the corresponding interaction points.

The LHC beams are structured into individual colliding bunches, with 10^{11} protons per bunch and time separation of 25 ns, which is the period of the proton-proton collisions at the LHC. In total, it is possible to have at most about 3600 bunches, however only up to 2808 could be filled with protons. The rest, called Abort Gap, are used for the experiment's maintenance, calibration, or other accelerator-related work.

One of the most important characteristics of the collision environment specifically for the analysis is the amount of Pile Up (PU). Typically, it is represented in terms of the number of primary vertices per a given bunch crossing. The importance arises from the fact that simulation has to be properly accommodated by mixing simulated processes with additional Minimum Bias samples (samples which correspond to events with detector occupancies above some known threshold). Figure 2.2 shows the distribution of the number of primary vertices per event both for simulation and data.

2.2.2 The CMS Detector

The Compact Muon Solenoid, one of the two largest experiments hosted at the LHC, is a two-fold creature; first of all, it is a one of its kind general purpose particle

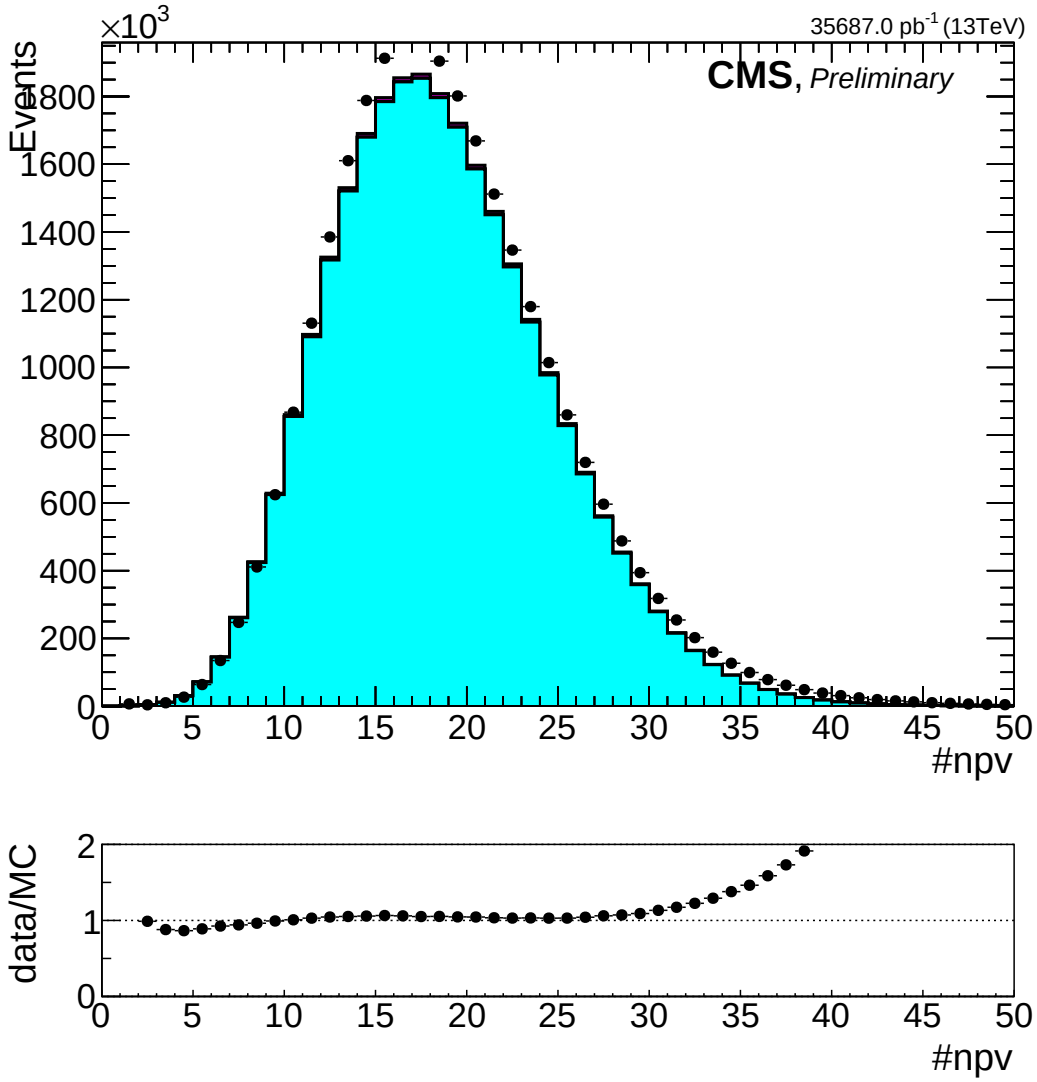


Figure 2.2: Pile Up distribution in terms of the number of primary vertices for data (black points) and simulation (filled area). Good agreement is observed with deviations near the tail of the distribution.

detector system, which consists of multiple components and whose primary objective is to record and reconstruct physics events that could help study the fundamental building blocks of nature. It acts like a camera, recording the footprint of the physics interactions which particle physicists use to deduce the underlying physics. The Compact Muon Solenoid is also an experiment, a collaboration of thousands of people whose combined effort made it possible to build and operate such a device. From people involved in data-taking at Point 5 and detector maintenance experts to people performing the analysis of recorded events and making sure that the collected data is of the highest quality - it is one of the best examples when group effort produces results of highest esteem.

The most important component of CMS that makes it stand out among other experiments is the superconducting solenoid with magnetic field of 3.8 T. The magnet is located just outside of the HCAL subsystem and plays a crucial role in the overall architecture of CMS. In particular, due to the strength of magnetic field, significant improvements are expected in the search for the Higgs Boson decaying via two muons due to the high resolution of the muon system together with the power of the magnet. The closest subsystem of CMS to the interaction point is the Silicon Tracker [10], whose primary objective is to measure the momenta of the charged particles passing through. What follows are Electromagnetic [11] and Hadronic [12] subsystems which respectively consist of several subdetectors with varying performance characteristics. Muon Systems are located just outside of the magnet and comprise different technologies depending on the η location [13]. In total, there are three different types of

gaseous systems used for muon chambers. Drift Tubes (DTs) together with Resistive Plate Chambers (RPCs) are located in the central (Barrel) region, a part of the CMS detector that expects lower occupancy due to the lower particle fluxes. In turn, in the region with higher neutron fluxes, Cathode Strip Chambers (CSCs) are used instead of DTs.

The Large Hadron Collider is not merely a challenging project from an engineering standpoint; it is also a data factory, one of its kind, that presents a unique challenge for the data processing and analysis domains. The amount of data that gets generated quickly becomes unmanageable and one has to be careful when selecting what to preserve and what to abandon. The collision rate is 40 MHz and the compressed size of just one event coming from CMS is on the order of 1 MB. That amounts to 40 TB/s and extrapolating up to a single hour of datataking it is at the order of 140 PB/h - obviously this becomes unfeasible very quickly. For the purpose of selection of events of interest, all HEP Experiments employ a sophisticated Trigger System, whose purpose is to select the physics events of interest. CMS has two layers for event triggering: Level 1 [14] and High Level Trigger (HLT) [15]. The Level 1 trigger system is a hardware-based processing system that is tightly integrated with the rest of the subsystems' electronics. Level 1 allows CMS to reduce the event rate from 40 MHz down to 100 KHz. This is achieved by applying basic selections at hardware level (integration with the subsystems' electronics results in zero copy processing) using reduced content information known as Trigger Primitives (TPs).

The 100 KHz output of Hardware Trigger System gets further reduced by the

High Level Trigger system, down to 1 KHz that gets actually written to disk. This content reduction is achieved by employing high level information produced via the reconstruction procedure. The CMS High Level Trigger System is a reasonable-sized High Performance Cluster (HPC) Complex which takes the output of Level 1 and acts as a filtering farm, preserving only events that should be stored to disk for offline processing. The software framework that gets used is the same one as for offline data analysis; however certain reconstruction steps are optimized for performance.

2.3 Higgs Boson at the LHC

The production of the Higgs Boson at the LHC is dominated by the gluon fusion process with the intermediate top quark loop. At the center of mass energy of 13 TeV and the hypothesized Higgs Boson mass near 125 GeV, the production cross section for this process is one order of magnitude larger than the Vector Boson Fusion (VBF) mode, the next highest contributing process. The main feature of VBF production is the presence of two forward jets in the event going into opposite ends (large separation in pseudo-rapidity coordinate among the two jets) of the CMS detector. The Higgstrahlung or Associated Production, the production of the Higgs Boson in association with either Vector Bosons or $t\bar{t}$, are the smallest contributing production processes considered in the search. The full list of the production processes (as CMS datasets) and corresponding cross sections is provided in the Section 2.5

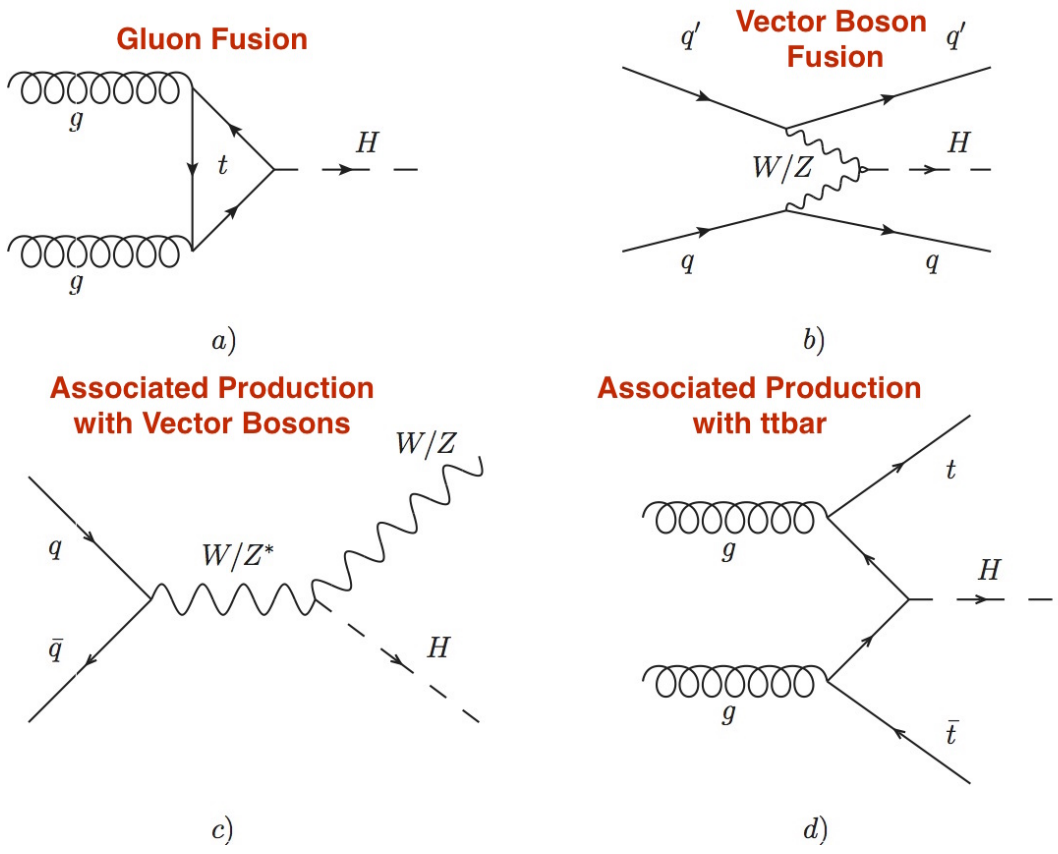


Figure 2.3: Feynman diagrams for the most dominant Higgs Boson production processes at the Large Hadron Collider. Gluon Fusion (a), Vector Boson Fusion (b), Associated Production with Vector Bosons(c), Associated Production with $t\bar{t}$ (d).

2.4 Previous Searches

On the 4^{th} of July 2012, the two largest experiments at the Large Hadron Collider, ATLAS and CMS, announced an observation of a new particle near the mass of 125 GeV. For that analysis, data collected at proton-proton collisions at $\sqrt{s}$ = 7 TeV and $\sqrt{s}$ = 8 TeV during the Run 1 production campaign was used. In total, the CMS experiment considered five decay channels: H $\rightarrow \gamma\gamma$, H $\rightarrow$ ZZ, H $\rightarrow$ W^+W^-, H $\rightarrow \tau^+\tau^-$, and H $\rightarrow b\bar{b}$. At the Higgs Boson hypothesis mass of 125.5 GeV, the combined significance of all of the five channels reached 5σ, which signals the discovery of a new particle near that mass. For the same mass hypothesis, the combination of just the first three channels (H $\rightarrow \gamma\gamma$, H $\rightarrow$ ZZ, H $\rightarrow$ W^+W^-) resulted in 5.1σ significance of the excess [16]. Higgs decay modes, H $\rightarrow \gamma\gamma$ and H $\rightarrow$ ZZ, are the drivers of the discovery due to the very good mass resolution. Furthermore, the combined measurement of the Higgs Boson mass by ATLAS and CMS experiments in the two most significant decay modes (H $\rightarrow \gamma\gamma$ and H $\rightarrow$ ZZ) was performed and found to be 125.09 GeV [17].

During the Run 1 analysis campaign, the decay of the Higgs Boson to two muons has been evaluated. Due to a very low branching fraction of the dimuon decay mode, 0.02%, which is approximately two orders of magnitude lower than the $\tau\tau$ decay, no excess of events above the expected background around the 125 GeV mass was observed [18]. Therefore, the upper exclusion limits were placed on the production cross section (times branching fraction). Results presented in this search will be further discussed and compared to Run 1 results in Section 2.13. The fact

that an observation of a new particle has been made around the mass of 125 GeV, allows one to significantly narrow the mass range when searching for the Higgs Boson in a dimuon final state.

2.5 CMS Datasets

At CMS, all of the accumulated data is organized in terms of datasets, which can be logically organized into key-value pairs. A key is the name of a dataset, which uniquely identifies it across the rest of the samples (dataset and sample are used interchangeably). At the same time, a key follows a standard naming convention, which allows to quickly resolve the most important characteristics of the actual data to which it points. A value, in turn, is a collection of data files that actually constitute the dataset and are used for the analysis.

Generally speaking, the CMS experiment produces two types of datasets: collision and simulation samples. Collision samples correspond to the actual data recorded from proton-proton collisions at the LHC. A typical collision sample name has three parts: stream identifier, timestamp and format of the data stored. Depending on the format, timestamp can either point to the actual period of datataking or the reconstruction date. The CMS assigns descriptive stream identifiers in order to provide an overview of the physical content of the events that go into a dataset. For instance, in this search, only events that have a good quality muon object are considered and a "SingleMuon" identifier signals exactly that. Furthermore, datasets could have identifiers like "SingleElectron" or "SinglePhoton", which would make them respectively

contain events with an electron or a photon.

Simulation samples correspond to the data produced by simulating the collisions environment (production processes and subsequent decays) and the response of the CMS detector. The most important use case of the simulation is to provide a modeling baseline with which collisions data will be compared. For the purpose of the search, simulation datasets can be further divided into the signal and background samples. Signal corresponds to the Higgs Boson production processes with its subsequent decay to two muons. Backgrounds, in turn, are basically all of the processes that can produce the same signature, the same final state. A typical unique simulation sample name, similar to collision sample name, has three parts: production process specifier, conditions identifier and the format of the data stored. The first part specifies a label that summarizes the actual production process and the Monte Carlo event generators used to perform the calculations of the Feynman diagrams. Conditions part carries a label to uniquely specify a campaign when a sample is produced, software versions used, and various calibrations and corrections applied.

2.5.1 Collision Datasets

For the purpose of the search, datasets with the total integrated luminosity of 35.9 fb^{-1} of CMS collision data collected over the course of 2016 datataking campaign are utilized and the full list of these samples is provided in the Table 2.1. The signature of the search is the presence of two opposite sign muons in the event; therefore, in order to pick up as many as possible events of interest, the choice for the stream identifier

is limited to either "SingleMuon" or "DoubleMuon". The choice of "SingleMuon" is dictated by the fact of having intrinsically higher efficiency of triggering a single muon rather than two muons per a given event. By selecting "DoubleMuon" trigger, for this particular search, all the events where only one muon triggers the HLT system would be thrown away. Therefore, considering that dimuon final state has a very low branching fraction for the Higgs Boson, the choice is made not to throw away the events.

Table 2.1: Datasets used for the search from proton-proton collisions recorded at the $\sqrt{s} = 13$ TeV by CMS at LHC in 2016.

Datasets	Int. Luminosity (fb^{-1})
/SingleMuon/Run2016B-03Feb2017_ver2-v2/MINIAOD	5.788
/SingleMuon/Run2016C-03Feb2017-v1/MINIAOD	2.573
/SingleMuon/Run2016D-03Feb2017-v1/MINIAOD	4.248
/SingleMuon/Run2016E-03Feb2017-v1/MINIAOD	4.009
/SingleMuon/Run2016F-03Feb2017-v1/MINIAOD	3.102
/SingleMuon/Run2016G-03Feb2017-v1/MINIAOD	7.540
/SingleMuon/Run2016H-03Feb2017_ver2(3)-v1/MINIAOD	8.606

2.5.2 Signal Datasets

For the purpose of testing various hypothetical Standard Model Higgs Boson masses, it is crucial to be able to build signal models for each hypothesis. In this analysis, samples with three different hypothetical Higgs Boson masses are used,

120/125/130 GeV, which allows a mass range [120, 130] GeV to be examined. The process of signal model construction and interpolation of parameters as a function of the Higgs Boson mass is discussed in further detail in Section 2.8. The choice of the mass range is driven by the evidence obtained from searches for the Higgs Boson in other final states, discussed in Section 2.4, where observations of a resonance near 125 GeV mass were made. Table 2.2 provides a summary of the CMS signal samples used along with cross section of each production process for the 125 GeV mass hypothesis. The Higgs signal production processes considered in this search are gluon fusion (ggH),

Table 2.2: Standard Model 125 GeV Higgs Boson Signal Datasets for 13 TeV. Dataset names for 120/130 GeV are omitted for brevity. Moriond 2017 conditions are used (omitted the conditions specification for brevity).

Datasets	σ (pb)
/GluGlu_HToMuMu_M125_13TeV_powheg_pythia8	48.58
/VBF_HToMuMu_M125_13TeV_powheg_pythia8	3.782
/WMinusH_HToMuMu_M125_13TeV_powheg_pythia8	0.5331
/WPlusH_HToMuMu_M125_13TeV_powheg_pythia8	0.851
/ZH_HToMuMu_M125_13TeV_powheg_pythia8	0.8839

vector boson fusion (VBF), Higgsstrahlung (VH). Production in association with top quarks ($t\bar{t}H$) has been generated privately. The Higgs MC samples are generated using POWHEG [19].

2.5.3 Background Datasets

As it has already been stated, background processes are the processes that result in the same final state (at least two muons in the event) as the Higgs Boson samples considered. For this analysis, all the processes that produce two muons in the final state, but not through their coupling to the Higgs field, are to be considered backgrounds. Table 2.3 provides a summary of the most dominant contributions among the background processes. The largest contributor is the Drell-Yan process, which constitutes approximately 90% of background events, and has a pair of leptons in the final state. The next-to-leading contributor is the $t\bar{t}$ production, with subsequent decay of top quarks into lighter bottom quarks and $W^{\pm}$ vector bosons further coupling to two fermions. These two mechanisms are responsible for more than 98% of background events contributing to the dimuon final state.

Table 2.3: Background Datasets. Moriond 2017 conditions have been used (omitted the conditions specification for brevity).

Dataset	σ (pb)
/DYJetsToLL_M-50_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	5765
/ST_tW_top_5f_NoFullyHadronicDecays_13TeV-powheg_TuneCUETP8M1	35.85
/TTJets_DiLept_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	85.656
/WJetsToLNu_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	61526.7
/WWTo2L2Nu_13TeV-powheg-herwigpp	10.481
/WZTo3LNu_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	4.712

Depending on the analysis strategy, the role of background simulation samples can be two fold. First, they are used for comparison with collision data, in particular to make sure that dimuon mass is well-modeled by the included backgrounds. The idea is to show that physical quantities of interest (dimuon mass, various kinematic variables) are in line with theoretical predictions. It is important to point out that both data and simulated samples will be further subject to exactly the same selections, further described in Section 2.6. Second, background datasets can be directly used for the hypothesis testing and statistical analysis of the presence of the signal in the data. In such a case, it is common to abbreviate this approach as simulation driven, because the simulated background dimuon mass distributions are directly used in the hypothesis testing. Another approach, commonly named data driven, is to build a model, similar to the construction of a signal model and discussed further in Section 2.9, that will be fit to the actual data, constrained and used to estimate the background yield.

This analysis follows a data-driven background estimation approach due to the low statistical power of simulated background samples. In other words, significant bin-to-bin fluctuations are present, especially for the categories with lower statistics, that would result in inadequate extraction of the upper limits. Therefore, the primary use of background datasets is to compare various kinematic variable distributions from data and theoretical predictions. Moreover, as it will be clarified in Section 2.7, depending on the categorization technique used, background samples will be further used for training a binary classification algorithm for the purpose of signal

discrimination.

Single top samples are generated with POWHEG, whereas $t\bar{t}$ samples and the multi-boson samples are generated either with MADGRAPGH [20] or AMC@NLO (Next to Leading Order) [21]. Spin effects in multiboson processes are simulated using MADSPIN. The parton shower and hadronization processes are modeled by the PYTHIA8 generator [22] with TuneCUETP8M1.

2.6 Event Selections

For the purpose of reconstruction of physics objects of interest, CMS utilizes a Particle Flow Algorithm [23], which aims to identify individual final state particles: photons, electrons, various hadrons, muons, etc. The main idea behind this algorithm is to use information not just from a particular subsystem (Muon, HCAL, ECAL, Tracker), but to combine and cross-reference features from various subdetectors: hits in the Tracker with Muon Stations, or clusters of energy depositions in ECAL, etc. In other words, this is an example of a sophisticated clusterization technique aimed to improve the resolution of energy and momentum reconstruction. For more information on the internals of the Particle Flow Algorithm, in particular in application to the CMS detector, consult the references [24], [25].

2.6.1 HLT Selections

The search requires one of the following Single Isolated Muon Triggers to fire per event:

- HLT_IsoMu24

- HLT_IsoTkMu24

The choice of the threshold is driven by the requirement of the availability of the trigger for the whole period of datataking with the lowest p_t threshold. These triggers require the event to have at least one isolated muon candidate with p_t above 24 GeV, with no explicit restriction on its pseudorapidity.

2.6.2 Primary Vertex Selections

Primary Vertex is the measured position, together with the associated uncertainty, of a vertex that corresponds to the proton-proton interaction. The identification and reconstruction of all primary vertices per a given event is done by first, selecting all the tracks that seem to originate from a common vertex, and then further performing a regression to actually identify the position of that vertex. One of the top requirements on selecting a proton-proton collision event is the presence of at least one valid Primary Vertex with the following conditions:

- $ndf \geq 4$ - number of tracks originating from this PV is greater than four
- $\rho < 2$ cm and $|Z| < 24$ cm - displacement along either Z-axis or in the transverse plan should be minimal w.r.t. the Interaction Point, defined as the origin.

2.6.3 Muon Selections

The muon candidates are reconstructed using the Particle Flow Algorithm by matching compatible track segments from the inner silicon tracker and the muon detectors [26]. In our analysis, muons are required to be within the pseudorapidity range $|\eta| < 2.4$ and with a minimum transverse momentum $p_t > 10$ GeV. Fur-

thermore, muons are constrained to have $\Delta\beta$-corrected relative isolation, defined in Equation 2.17, of $I_{rel}^{PF} < 0.25$. In the definition of the isolation variable p_T^{ch} is the charged hadron transverse sum of individual momenta, E_T^{γ} is the photon transverse energy sum, and E_T^{nh} is the neutral hadron transverse energy sum. The term p_T^{chPU} is the estimated transverse momentum of charged particles from pileup in the $\Delta R < 0.4$ cone (defined in Equation 2.16) and the factor of 0.5 is used to estimate the neutral pileup from the charged component.

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{2.16}$$

$$I_{rel}^{PF} = (p_T^{ch} + max(0, E_T^{\gamma} + E_T^{nh} - 0.5 * p_T^{chPU}))/p_t \tag{2.17}$$

Additionally, only "Medium Id" muons are selected - centrally provided and recommended by the CMS Physics Object Group [27, 28]. As it was mentioned previously, muons are reconstructed using clusters of hits from various subsystems (Tracker and Muon Stations); in other words a physics object muon is defined to be a composite cluster of combined hits from the Tracker and Muon Systems. It is important to understand the difference between the actual physical particles, which are muons, and the definition of a reconstructed muon! Moreover, when extracting the momentum of a muon, a regression procedure is performed, a fit for instance, from which it is possible to extract the goodness of fit or some other quality metric, using which one can judge the quality of the reconstructed candidate. Finally, "Muon

Id" is a set of criteria that defines a muon candidate of certain quality and allows to discriminate candidates based on this feature.

2.6.4 Muon Corrections

Given a very narrow theoretical width of the Higgs Boson, 4.2 MeV, the Higgs peak is dominated by the detector resolution which is on the order of several GeV. Therefore, to increase the sensitivity of the search, it is crucial to have the best possible dimuon mass resolution, both in data and Monte Carlo. Moreover, as it is going to be described further in the Section 2.8, it is necessary to correct for possible dimuon mass scale shifts.

There are two different sets of Muon corrections derived centrally within the CMS collaboration that are to be used separately: Rochester [29] and Kalman [30]. Both allow to correct the scale and smear the resolution. For the validation purposes, both sets of corrections are applied and the effects on Z scale and resolution are evaluated. In Figures 2.4 and 2.5 comparisons of uncorrected versus corrected results are shown for the scale and resolution, respectively. Both the Rochester and Kalman muon corrections successfully align Data with MC in terms of the Z peak. The impact of the two sets of corrections on the expected limits was evaluated and compared. It was found that the choice of one or another produces negligible difference for the expected limit, therefore final selections use the Rochester corrections.

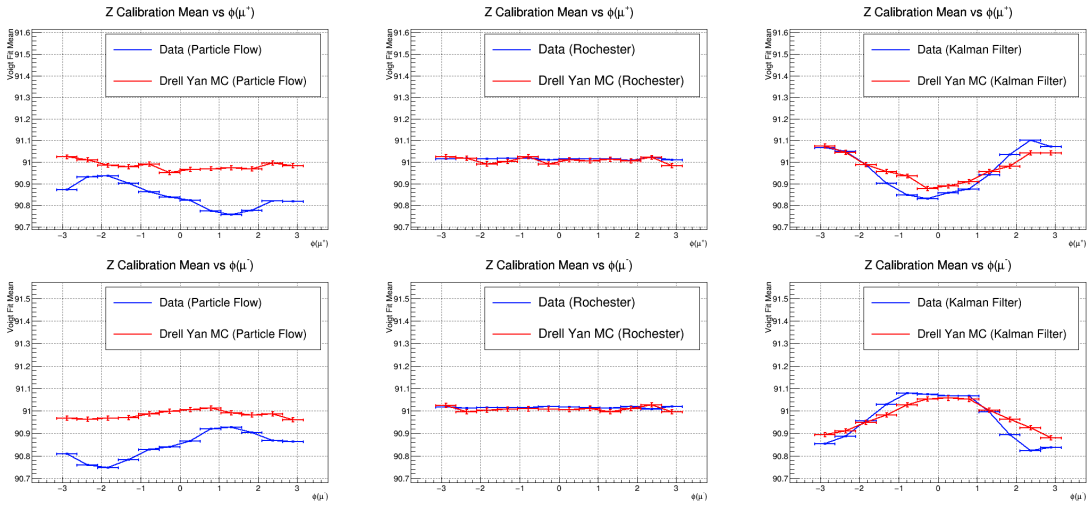


Figure 2.4: Comparing Uncorrected (left), Rochester (center) and Kalman (right) Corrections effects on the Z Scale (mass). Top (Bottom) three plots correspond to the Z mean vs phi of the first (second) muon from the candidate pair.

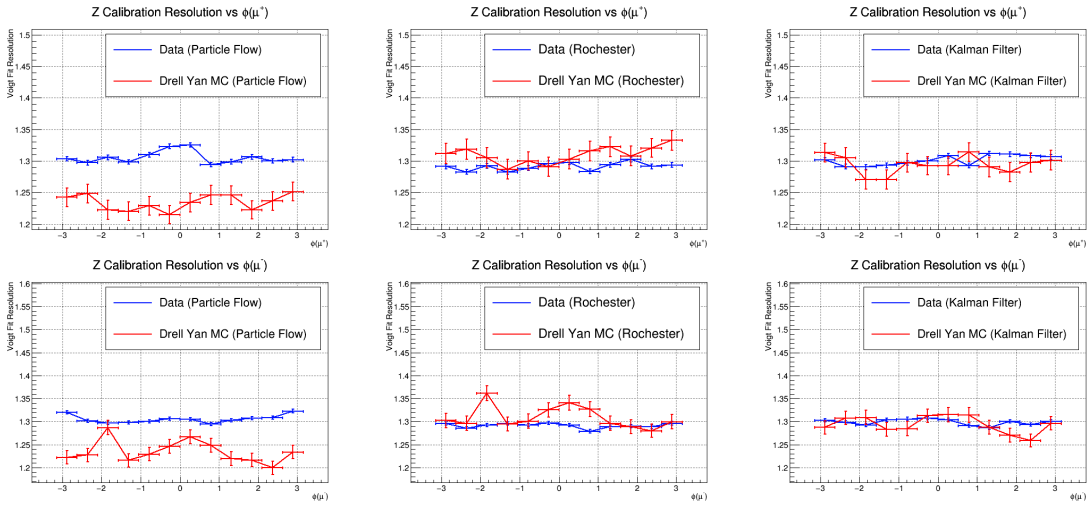


Figure 2.5: Comparing Uncorrected (left), Rochester (center) and Kalman (right) Corrections effects on the Z Resolution (width). Top (Bottom) three plots correspond to the Z width vs phi of the first (second) muon from the candidate pair.

2.6.5 Jet Selections

Jets are reconstructed from PF candidates with the anti-κ_T algorithm [31] with a distance parameter of 0.4 after rejecting the charged hadrons that are associated to pileup primary vertices. Jets are "cleaned" against muons - jets overlapping with the selected PF muons, $\Delta R < 0.4$, are not considered further in the analysis. The selected jets with a minimum p_t of 30 GeV and a maximum $|\eta|$ of 4.7 are corrected in energy in order to account for non-uniform detector response.

The PF jets are further required to fulfill the jet identification criteria. Candidates with $|\eta|$ less than 2.7 are required to contain at least one charged PF candidate and have a non-zero charged energy fraction, and a charged electromagnetic energy fraction less than 0.99. The neutral and photon energy fractions must be less than 0.99. Jets in the region $2.7 < |\eta| < 3.0$ are required to have a neutral electromagnetic fraction of greater than 0.01 and neutral hadron fraction less than 0.98 while jets with pseudorapidity above 3.0 are required to have more than ten neutral particles and a neutral electromagnetic energy fraction less than 0.9.

For reconstructed jets with $|\eta| < 2.4$, the combined secondary vertex (CSV) b-tagging algorithm [32] is used to discriminate against the $t\bar{t}$ background process. The CSV medium operating point is chosen as the baseline for the search analysis and corresponds to a b-tagging efficiency of $60 - 65\%$ while the misidentification rate for light quarks such as u, d, s, and gluon jets is at the order of 1%.

2.6.6 Selections Summary

Cuts and selections can be summarized for the search as follows:

- At least 1 Primary Vertex passing the PV Selections
- At least 1 of the HLT Paths fired
- At least 2 opposite sign muons which pass Muon Selections. If more than 1 pair - choose the 2 candidates with the highest p_t. That is the Higgs Candidate.
- At least 1 muon from the Higgs Candidate pair with $p_t > 26$ GeV and is matched to the HLT Path that fires the event with $\Delta R < 0.1$.
- Filter out the jets that do not pass the Jet Selections, but do not require a certain number of them at this stage.

2.6.7 Validation

After applying all of the specified selections, a set of events is selected that is going to be further used in order to maximize the sensitivity of the search. However, before proceeding to the next section, validation of basic kinematic distributions needs to be considered for the inclusive set of events.

First, in Figure 2.6, the inclusive dimuon mass distribution is presented without any muon momentum corrections applied. Notice the presence of a discrepancy near the Z mass (≈ 91 GeV) peak. It is the result of a discrepancy in both the mass scale and resolution.

Second, in Figure 2.7, distributions of various kinematic variables after passing the selections are presented. Overall, data is well modeled; however, there are some

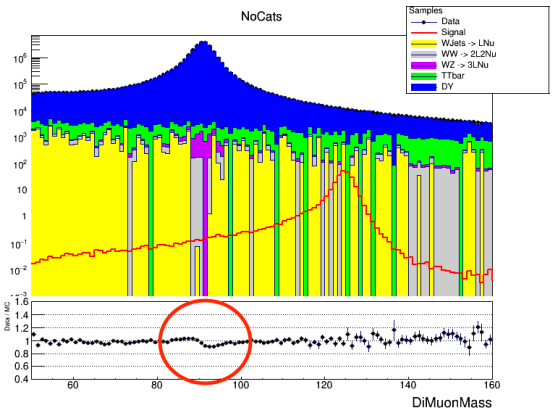


Figure 2.6: Inclusive Dimuon Mass Distributions without any Muon Corrections. Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line. The discrepancy around the Z mass peak of 91 GeV is circled in red.

deviations in particular for the p_t of the dimuon system. Up to 40% discrepancy is observed; however, only for a small region around 50 GeV. For the muon p_t, the observed deviation happens only for the transverse momentum values below the 26 GeV threshold that is being used for the muon matching the HLT. It is crucial to point out that although there are deviations in the modeling of some of the kinematic distributions, the modeling of the dimuon mass distribution shows very good agreement between the data and the simulated backgrounds.

Furthermore, in Figure 2.8, the inclusive dimuon mass distributions are shown with Rochester and Kalman corrections respectively applied to both Data and Monte Carlo. Notice that the discrepancy in scale and resolution is no longer present around the Z mass ($\approx$ 91 GeV) peak. Overall, comparing collision data with the simulated

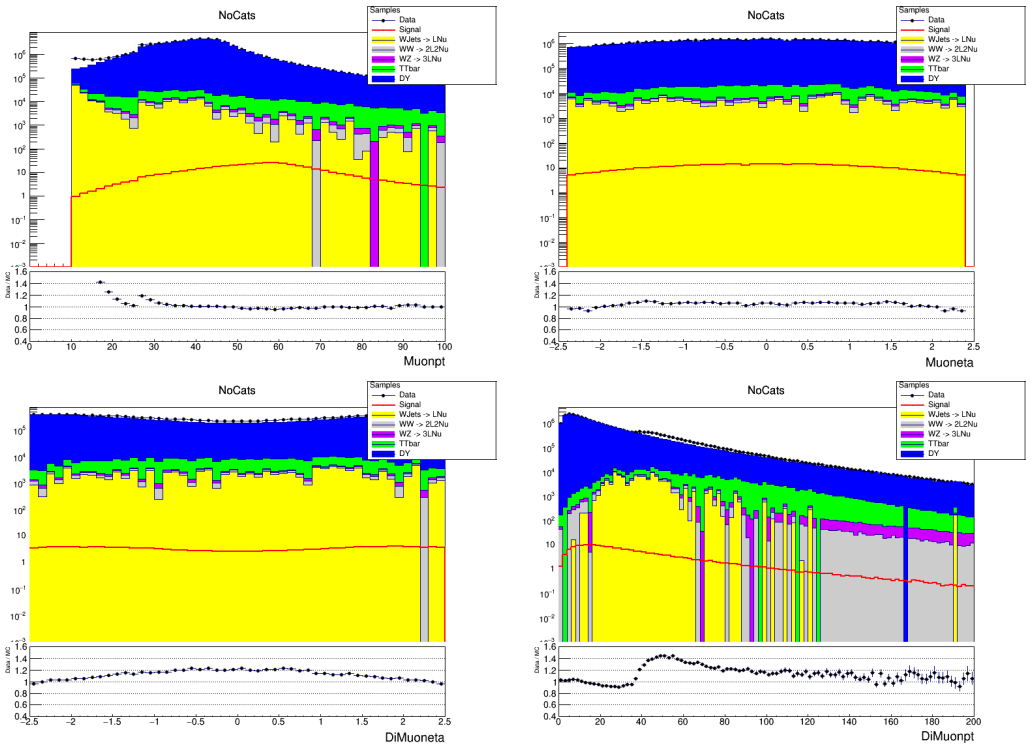


Figure 2.7: Inclusive Kinematic Distributions: muon p_t (top left), muon η (top right), dimuon η (bottom left) and dimuon p_t (bottom right). Discrepancy between data and simulated backgrounds is observed for muon p_t and dimuon p_t.

backgrounds, the dimuon mass distribution is well modeled.

2.7 Event Categorization

For the purpose of increasing the sensitivity of the analysis, the next step performed is the categorization of events into different subclasses. The basic idea is to utilize the knowledge about two important characteristics of the search needs: the physics of interest; i.e., differences between the production processes, and the fact that muon momentum resolution worsens with the increasing η coordinate of a track (barrel vs endcap regions of the CMS detector). Using this information will allow to substantially increase the significance of the Standard Model signal over the expected background.

2.7.1 Baseline Categorization

During the CMS Run I analysis campaign [18], the categorization procedure used was optimized to separate VBF-like events from the rest by requiring the presence of at least two jets passing the Jet Selections and going into the opposite ends of the detector, and employing the large Missing Transverse Energy (MET) selection. Moreover, the muon momentum resolution is significantly better in the Barrel region than in the Endcap, therefore the space is subdivided in the η variable by defining:

- Barrel: $|\eta| < 0.8$
- Overlap: $|\eta| \geq 0.8$ & $|\eta| < 1.6$
- Endcap: $|\eta| \geq 1.6$ & $|\eta| < 2.1$

Furthermore, once a Higgs candidate pair is constructed, each muon is tagged accord-

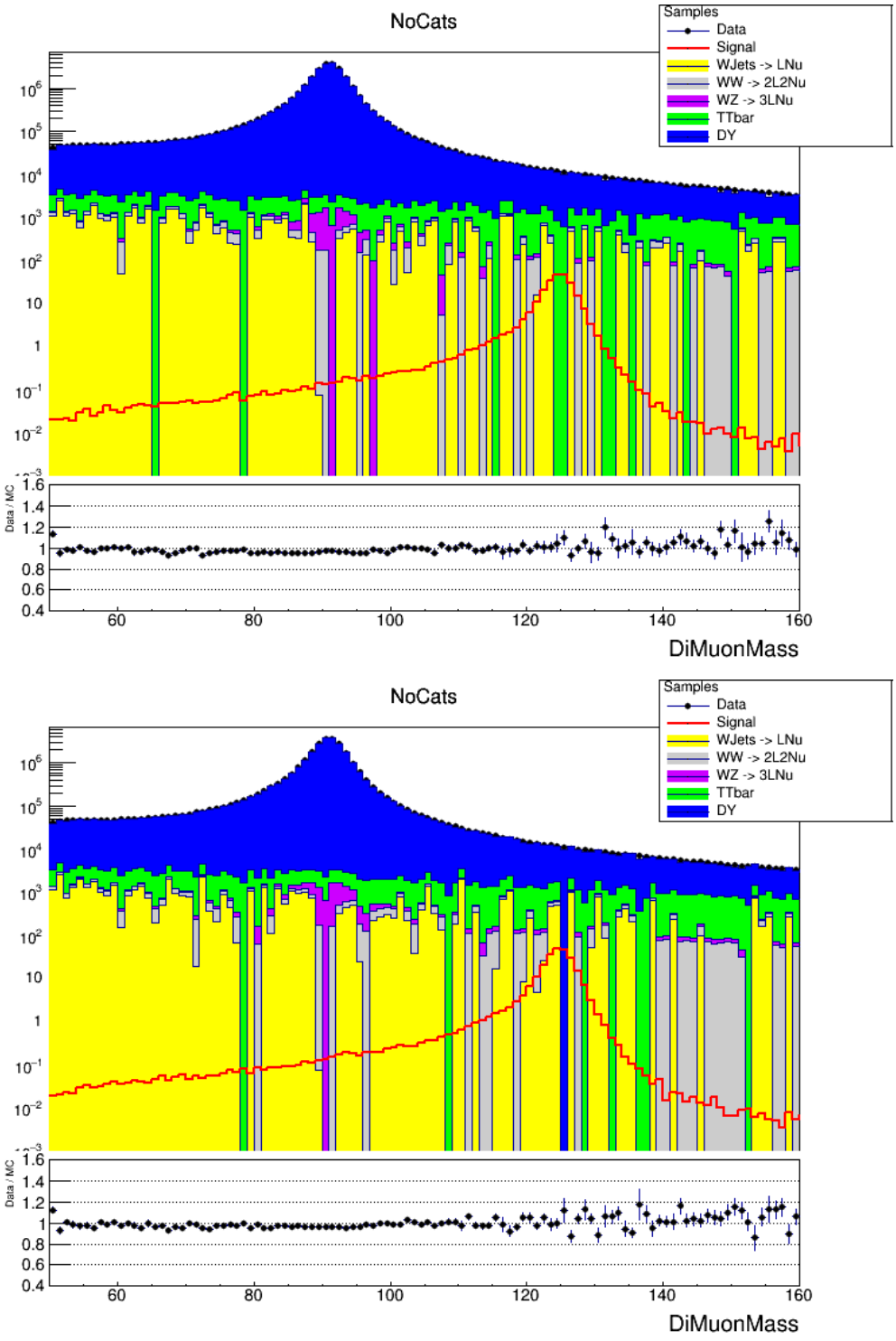


Figure 2.8: Inclusive Dimuon Mass Distributions. Rochester (Left) and Kalman (Right) Corrections applied. Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

ing to the region of the detector in which it was reconstructed. The exact selections applied in the baseline categorization procedure are:

- $njets \geq 2$ & $p_t^{j1} > 40$ GeV & $p_t^{MET} < 40$ GeV - require at least two jets with thresholds on p_t of the leading jet and Missing Energy Transverse.
  - "VBFTight" Category: $m_{jj} > 650$ GeV & $|\Delta\eta| > 3.5$ - category targeting the Vector Boson Fusion production mechanism. The signature is the presence of two jets with a large invariant mass and going into opposite ends of the CMS detector.
  - "GFTight" Category: if not VBFTight & $m_{jj} > 250$ GeV & $p_t^{\mu\mu} > 50$ GeV. In this category, we target Gluon Fusion production together with the events that do not pass the VBF-like selections.
  - "GFLoose" Category: here everything is collected that does not pass the VBFTight and GFTight selections.
- $njets \leq 1$ - the biggest portion of the events will have only 1 or 0 jets total per event (that pass previously discussed selections).
  - "01JetsTight" Category - $p_t^{\mu\mu} \geq 25$ GeV. The only discriminating variable we are left with at this point is the p_t of the dimuon system that typically has higher values for the signal events.
    - "01JetsTightBB": $\mu_{1,2}$ are Barrel
    - "01JetsTightBE": μ_1 is Barrel & μ_2 is Endcap
    - "01JetsTightBO": μ_1 is Barrel & μ_2 is Overlap

    - "01JetsTightEE": μ_1 is Endcap & μ_2 is Endcap
    - "01JetsTightEO": μ_1 is Endcap & μ_2 is Overlap
    - "01JetsTightOO": μ_1 is Overlap & μ_2 is Overlap
  - "01JetsLoose" Category - Left over events fall into this category.
    - "01JetsLooseBB": $\mu_{1,2}$ are Barrel
    - "01JetsLooseBE": μ_1 is Barrel & μ_2 is Endcap
    - "01JetsLooseBO": μ_1 is Barrel & μ_2 is Overlap
    - "01JetsLooseEE": μ_1 is Endcap & μ_2 is Endcap
    - "01JetsLooseEO": μ_1 is Endcap & μ_2 is Overlap
    - "01JetsLooseOO": μ_1 is Overlap & μ_2 is Overlap

Figures 2.9- 2.13 show dimuon mass distributions for all of the terminal categories.

2.7.2 Greedy Categorization

Another approach taken was to optimize the manually constructed categorization tree, discussed in the previous section, in a more systematic way by first training and applying a Boosted Decision Tree (BDT) for the purpose of signal to background discrimination. Then, use the BDT output as a new input feature for optimizing the categorization tree using a greedy optimization procedure.

In the first step, the BDT for the signal to background discrimination has been trained using both signal and background Monte Carlo samples. The BDT implementation from the ROOT [33] framework has been successfully used. Given

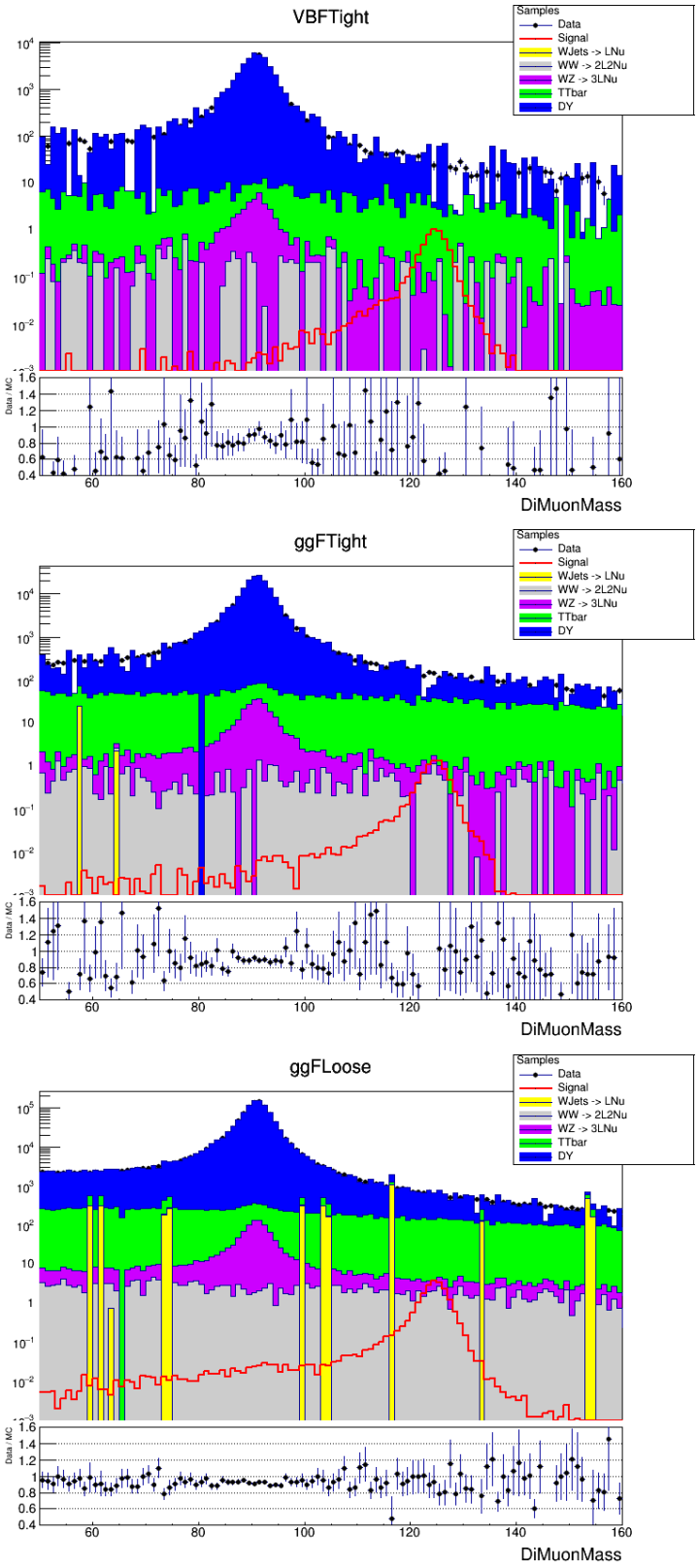


Figure 2.9: Dimuon Mass Distributions for VBFTight (Top), GFTight (Middle) and GFLoose (Bottom) Categories. Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

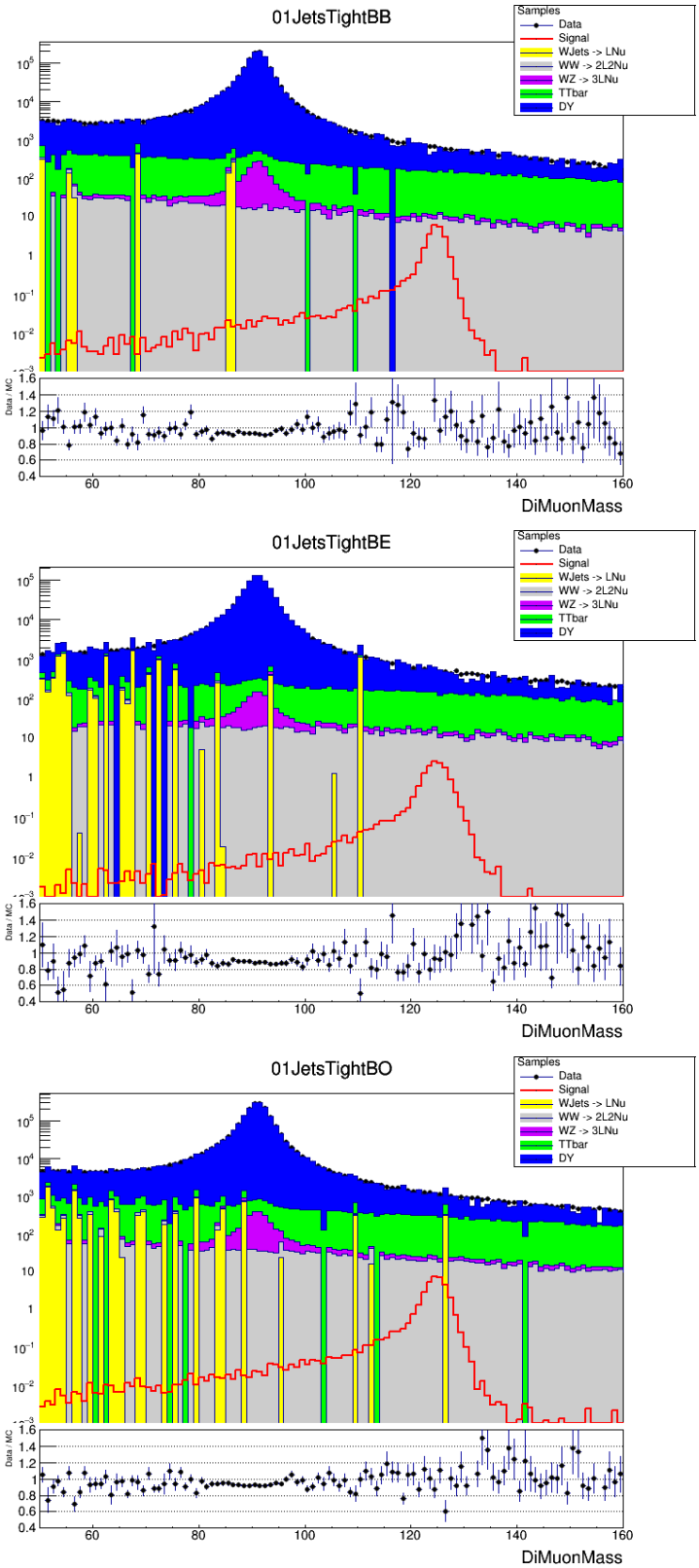


Figure 2.10: Dimuon Mass Distributions for 01JetsTight Categories: BB (Top), BE (Middle) and BO (Bottom). Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

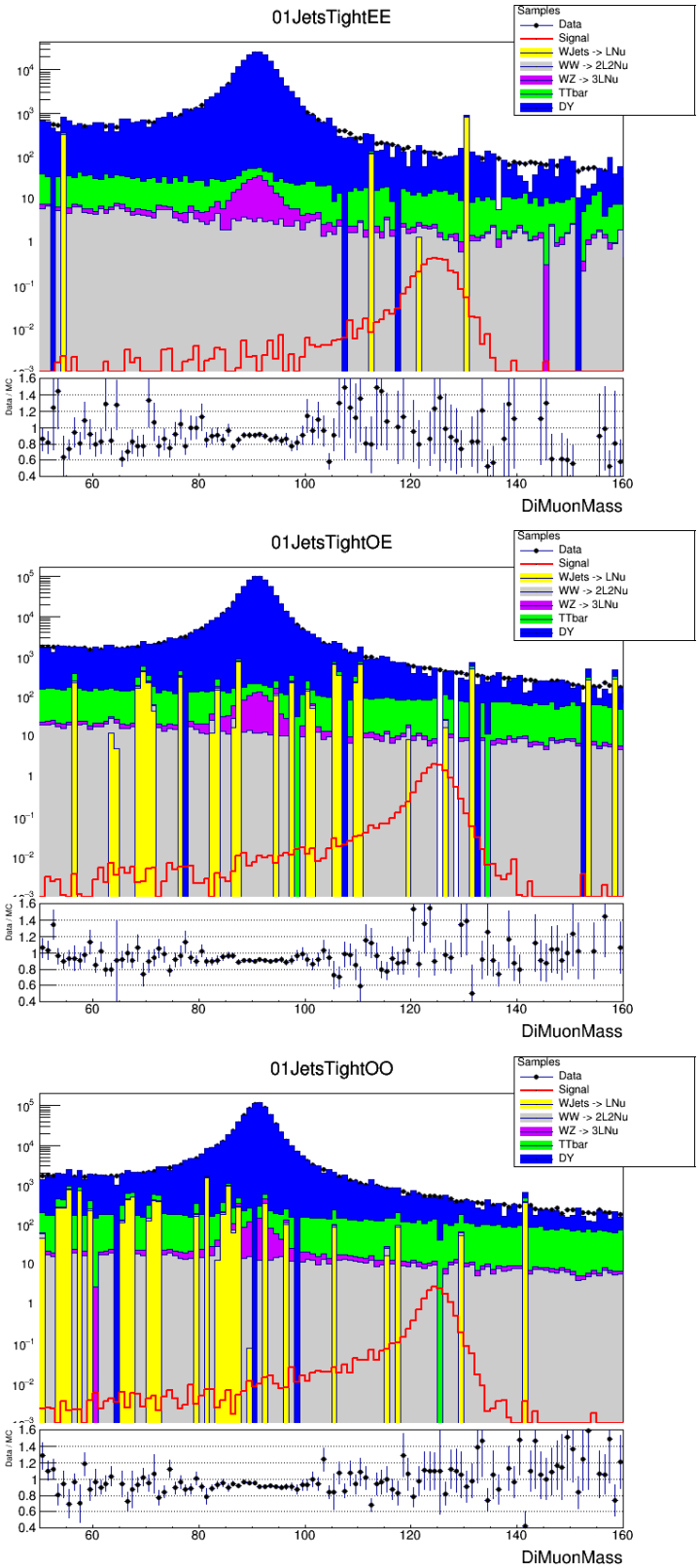


Figure 2.11: Dimuon Mass Distributions for 01JetsTight Categories: EE (Top), OE (Middle) and OO (Bottom). Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

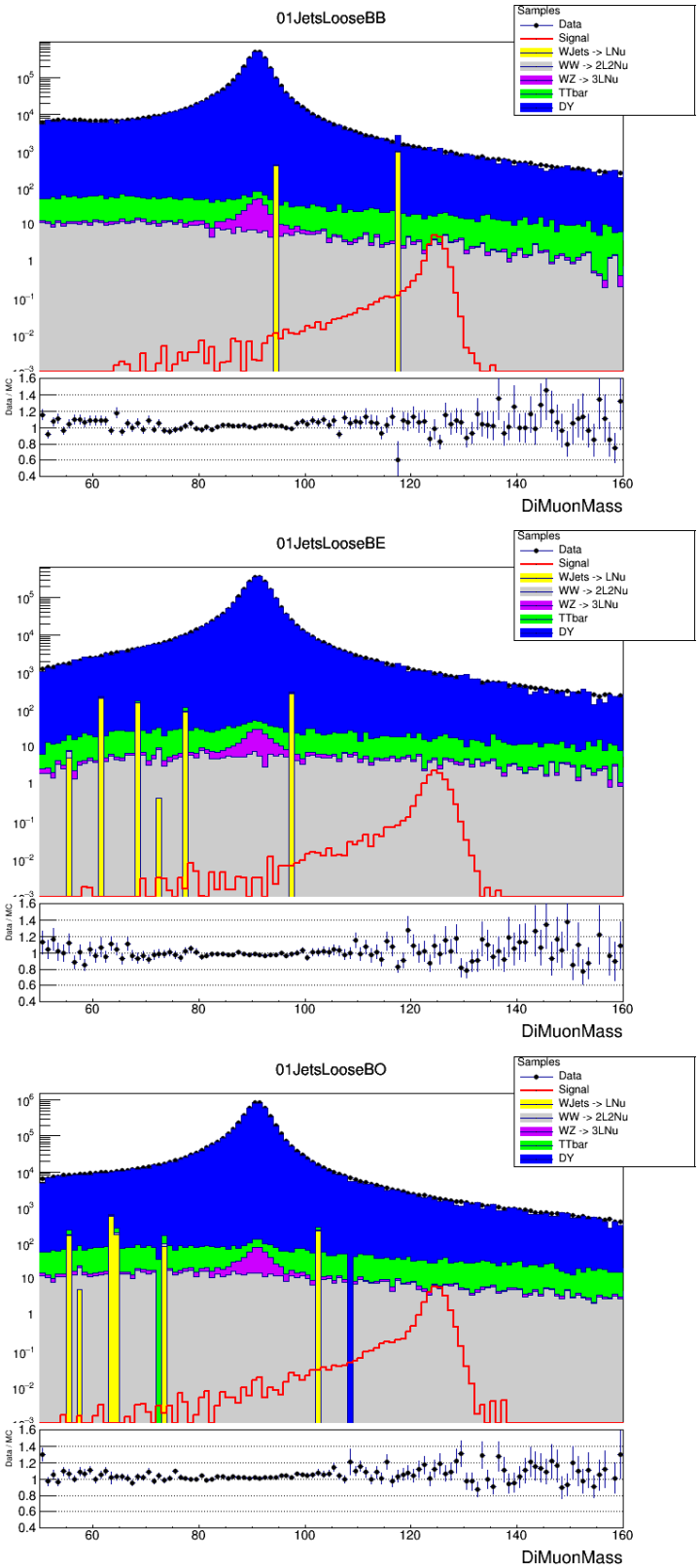


Figure 2.12: Dimuon Mass Distributions for 01JetsLoose Categories: BB (Top), BE (Middle) and BO (Bottom). Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

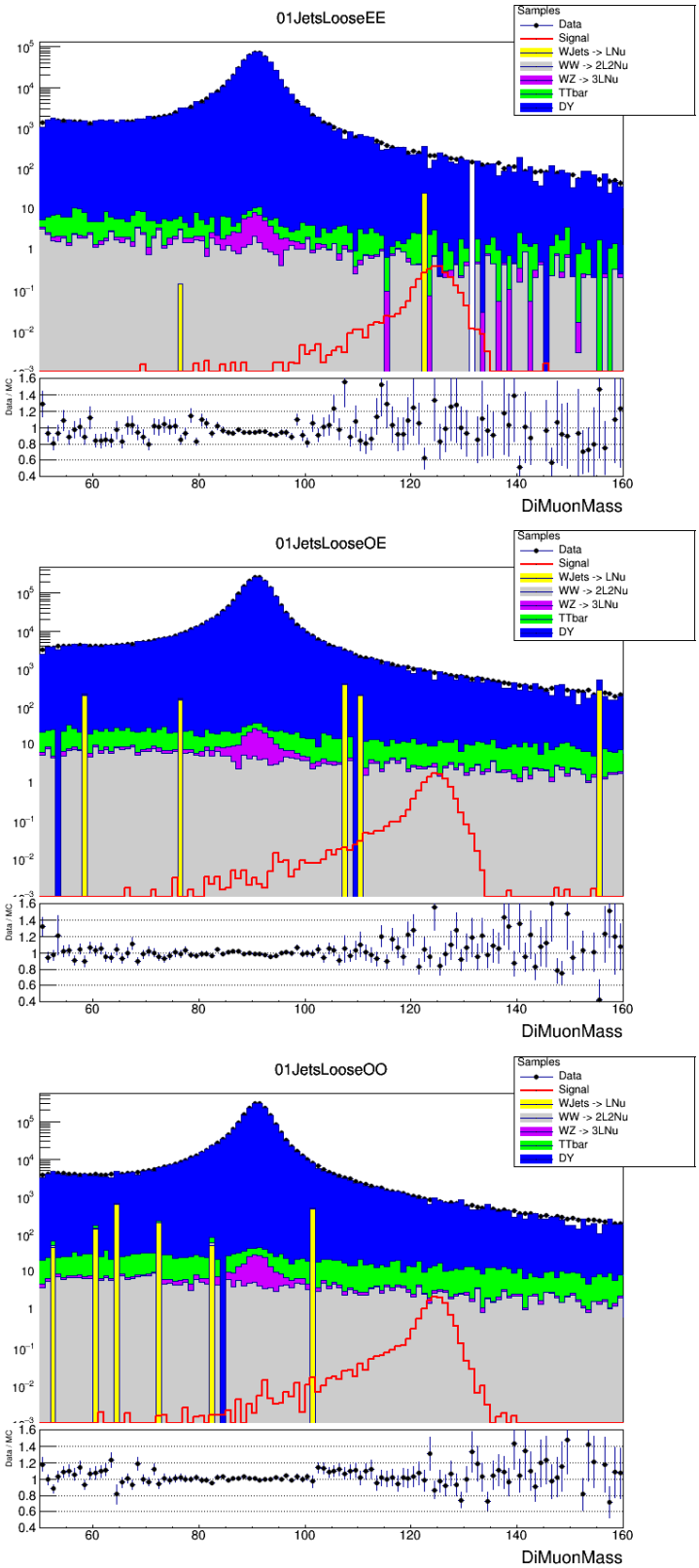


Figure 2.13: Dimuon Mass Distributions for 01JetsLoose Categories: EE (Top), OE (Middle) and OO (Bottom). Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

that background samples are not utilized for the statistical analysis, all of the available MC background events are employed for the purpose of training, cross-validation and testing. However, in terms of the Higgs signal, the even-numbered events are used for the categorization optimization. The rest is reserved for the statistical analysis. The following kinematic variables are used as input features for the BDT:

- The $p_t^{\mu\mu}$ of the dimuon system
- The $\eta_{\mu\mu}$ of the dimuon system.
- The $|\Delta\eta|$ between the muons from the Higgs candidate.
- The $|\Delta\phi|$ between the muons from the Higgs candidate.
- The η_{jj} of the dijet system. 2 highest p_t jets are used.
- The m_{jj} of the dijet system. 2 highest p_t jets are used.
- The $|\Delta\eta|$ between the 2 highest p_t jets.
- The number of Central jets with $|\eta| < 2.4$
- The number of Forward jets with $|\eta| > 2.4$
- The number of b-tagged jets - jets passing the CSVv2 medium b-tag working point
- The $\not{E}_T$

As it follows from the above list, there is no dependence on the Higgs candidate mass among the features. In other words, given the above features, one can not deduce the mass value of the dimuon system. Mass independence is an important criteria of this discrimination technique.

After the training and cross-validation to tune the hyper parameters of the decision tree itself, the results for the test dataset are shown in Figure 2.14, where the BDT output distributions for both training and testing subsets are overlaid and are in a very good agreement. Furthermore, the ROC curve from Figure 2.14, which is an indicator of the efficiency of the selection and rejection at the same time, shows a definite deviation from the diagonal (no discrimination). It follows that even with 0% efficiency rejecting background, the signal classification is not 100% efficient. Finally, given that a pure binary classification is not performed, the BDT output score is preserved for the next stage.

The next step of the Greedy Categorization procedure is to optimize the categorization tree selections and perform the actual splitting of events into different subsets. For that purpose, a simple decision tree is used with a custom metric, signal significance, and with the following two features:

- $\max(\eta_{\mu_1}, \eta_{\mu_2})$ - max η among the muons from the Higgs Candidate pair
- Discriminating BDT score

Again, only part of the signal has been used for the training, whereas all of the background events have been passed through the custom decision tree.

Before moving forward with the actual algorithm, a few definitions are provided. First of all the objects to be used are the dimuon mass histograms as in Figure 2.15. For a given mass distribution, define a signal significance (S_{node}) according to Equation 2.18. $N^S_{c,i}$ and $N^B_{c,i}$ are the yields (number of entries) for that particular histogram and for that particular bin. Note that all the bins in the distribution are

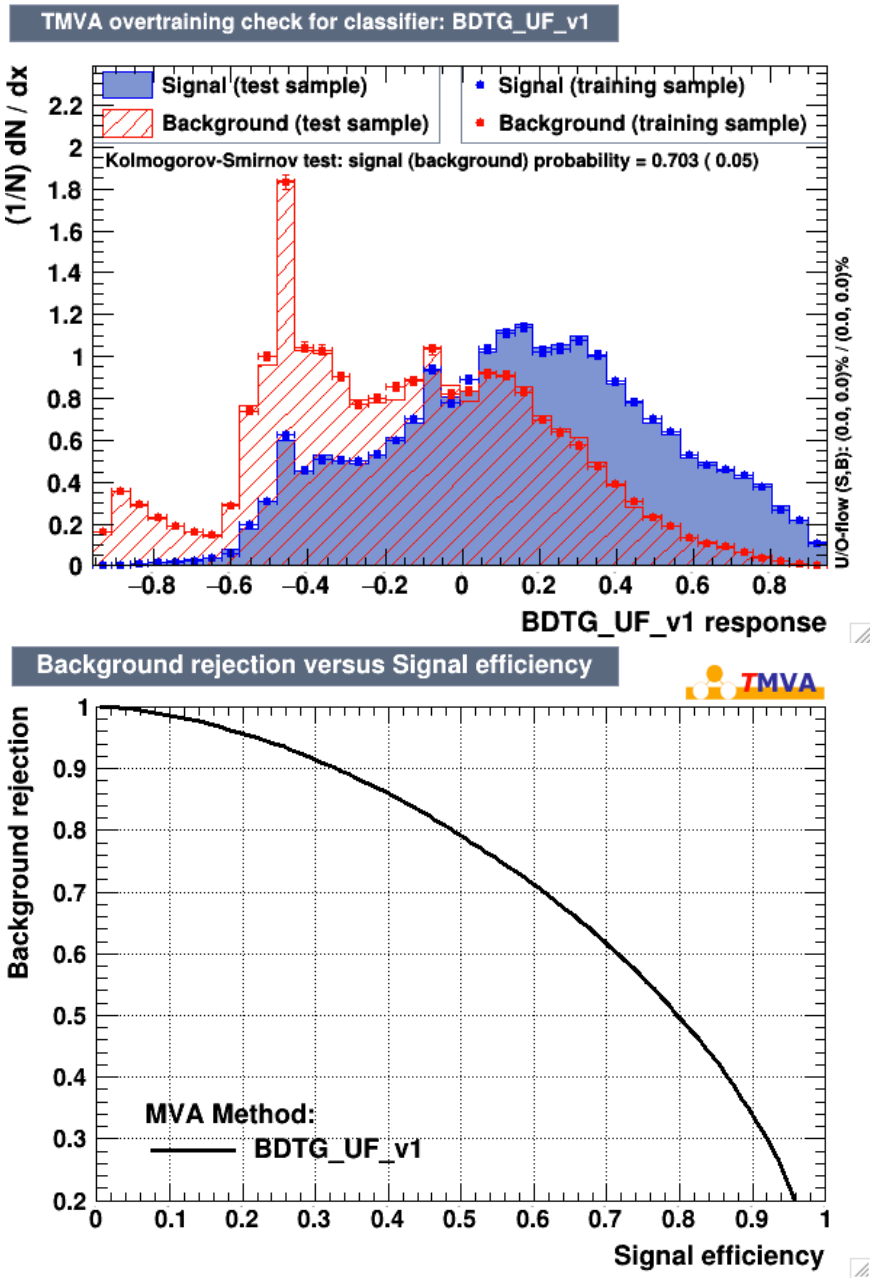


Figure 2.14: BDT score distribution (Top) and Receiver Operating Curve, ROC, (Bottom) - a kind of True Negative/Positive selection efficiency indicator.

treated separately and only those that fall in the [120 GeV, 130 GeV] mass range are used. Furthermore, the metric, based on which the splitting is optimized, is provided in Equation 2.19, where S^2_{left} and S^2_{right} are the significances of the two histograms created by a given split. Roughly speaking, the split that gives the maximum gain across all of the possible splits is to be selected.

$$S^2_c = \sum_{c,i} (N^S_{c,i})^2 / N^B_{c,i} \tag{2.18}$$

$$G = S^2_{left} + S^2_{right} - S^2_c \tag{2.19}$$

After providing all of the important definitions to work with, the actual procedure for the tree splits optimization can be summarized as follows:

- Stage 1. Start out with the inclusive set of all events - a node.
- Stage 2. Greedily scan through the features and select the split:
  - Scan through all of the possible values of $\max(\eta_{\mu_1}, \eta_{\mu_2})$ as split's candidates. Save the split that produces maximum gain.
  - Scan through all of the possible values of the BDT score. Again, save the split that maximizes the gain for the second feature.
  - Select the split which maximizes the gain: $\max(\mathrm{G}_\eta, \mathrm{G}_{score})$
- Stage n. Repeat Stage 2 recursively for each of the new nodes until the tree reaches the limit on the number of categories or runs out of splitting conditions.

The procedure just described is an example of a greedy algorithm, because the full parameters' phase space optimization is not performed as it is computationally infea-

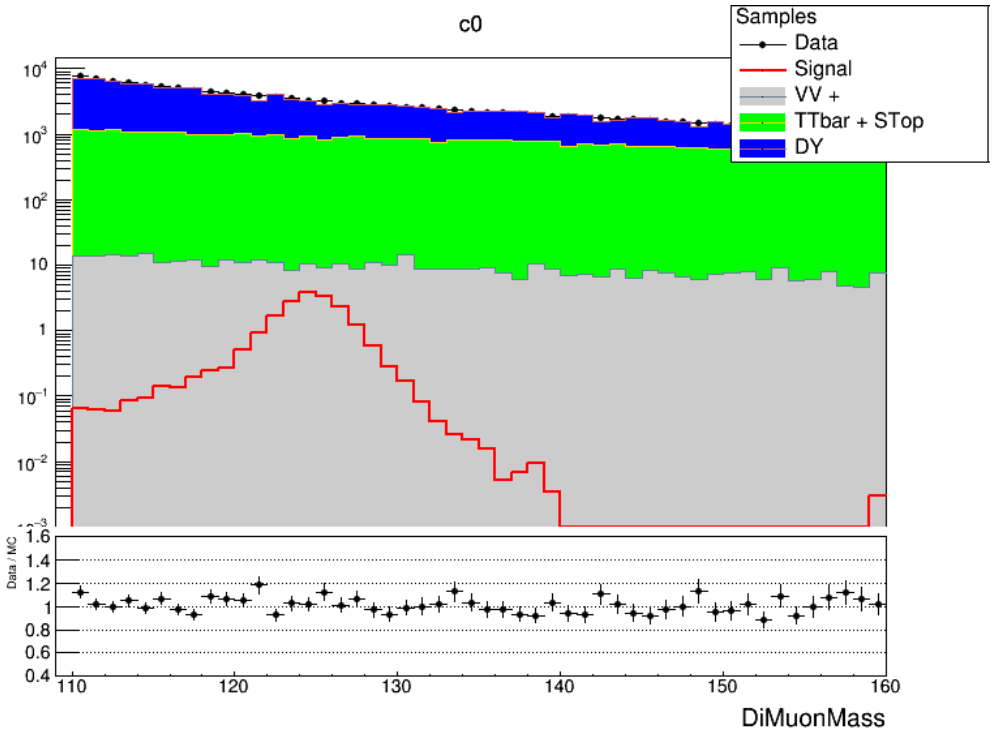


Figure 2.15: Mass Distribution for Category "c0" - lowest Discriminating BDT score category. Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

sible. Rather, the metric of interest is optimized greedily, step by step iterating over all of the available features.

The outcome of this technique is 13 new categories with the categorization tree shown in Figure 2.16. Integer labels are assigned based on the trained sensitivity of a particular subset, going from the category with the lowest sensitivity, "c0", to the one with the highest, "c12".

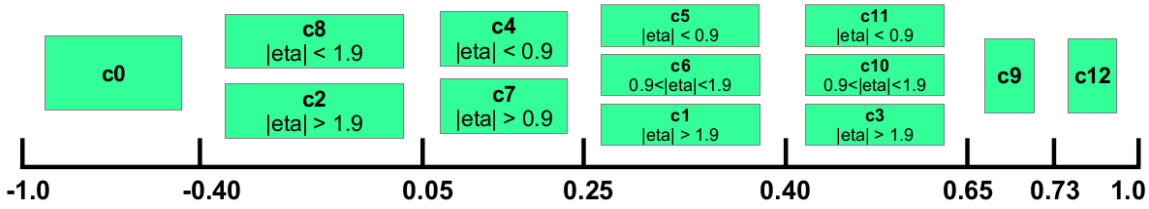


Figure 2.16: Greedy Categorization Tree.

The physical content of these categories can be observed in Figure 2.17, which shows the composition of the Higgs Boson production processes for each category. For instance, it can be deduced that the most sensitive category, "c12", has a large fraction of the VBF events (almost on par with Gluon Fusion), compared to less sensitive categories. This is due to the fact that the Vector Boson Fusion production mode has two additional jets in the final state and therefore has additional variables to discriminate upon.

Figures 2.18 - 2.21 summarize the resulting dimuon mass distributions. The modeling of the dimuon mass agrees well between data and simulations.

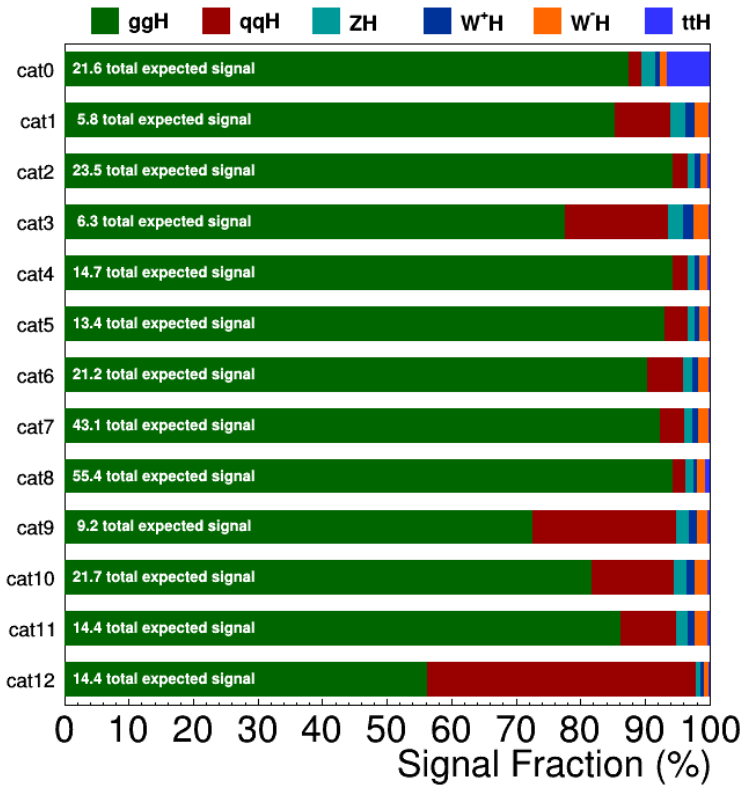


Figure 2.17: Composition of the signal production processes per category.

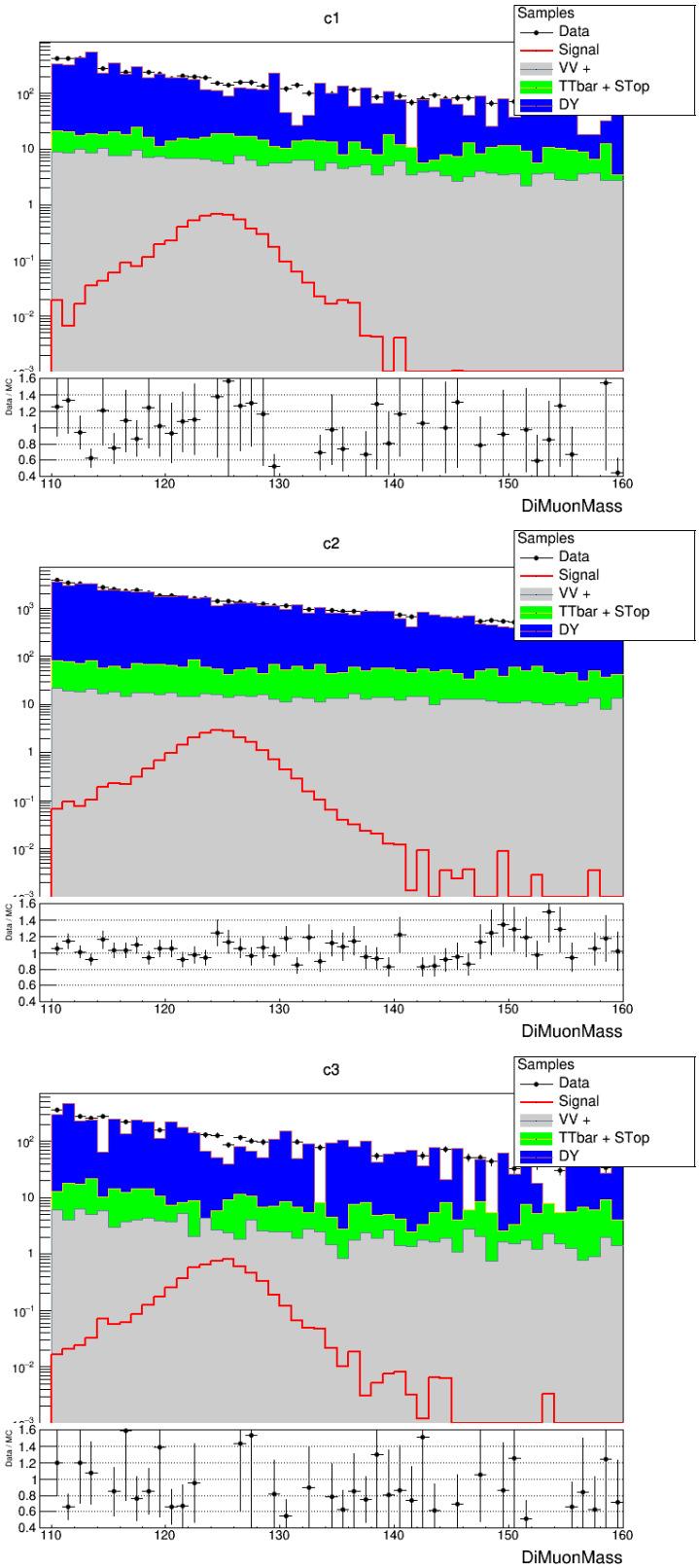


Figure 2.18: Dimuon Mass Distributions for c1-c3 subsets from the Greedy Categorization. Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

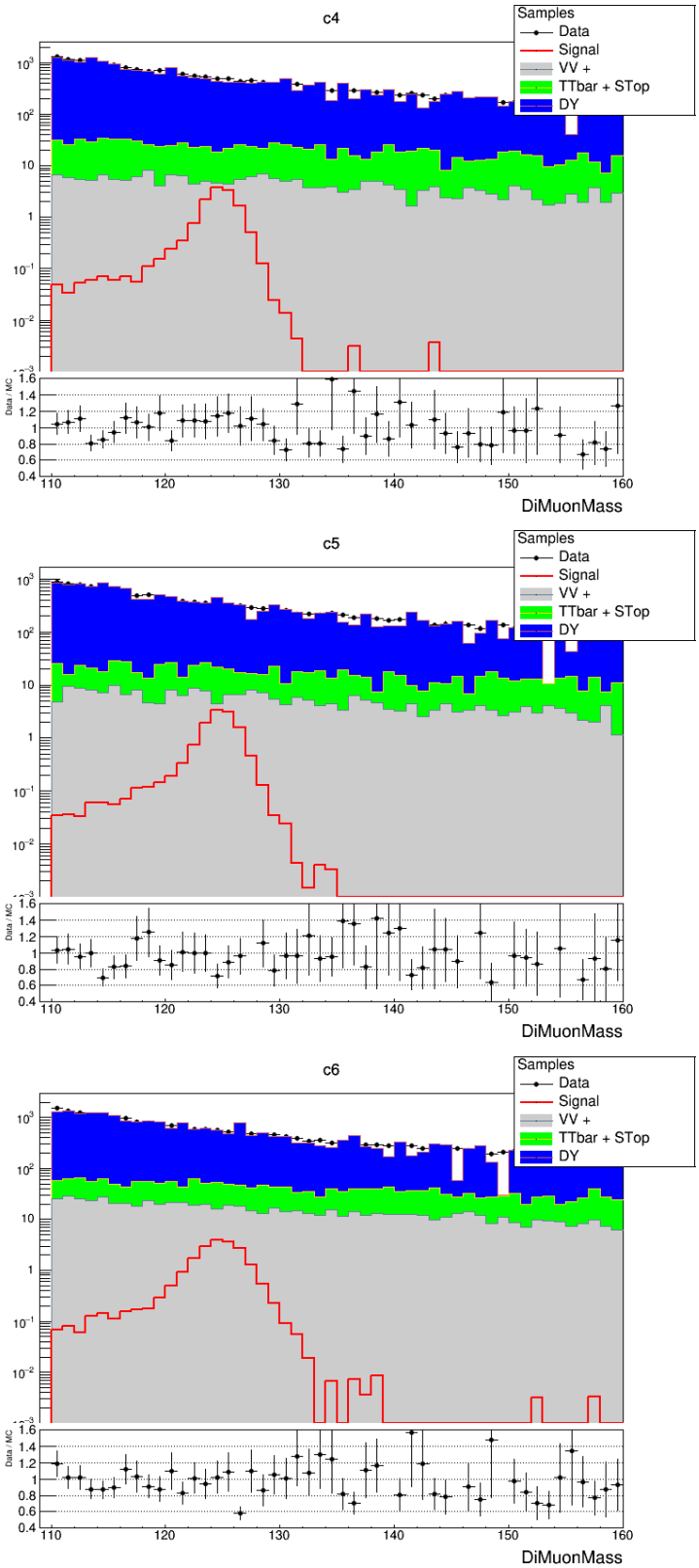


Figure 2.19: Dimuon Mass Distributions for c4-c6 subsets from the Greedy Categorization. Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

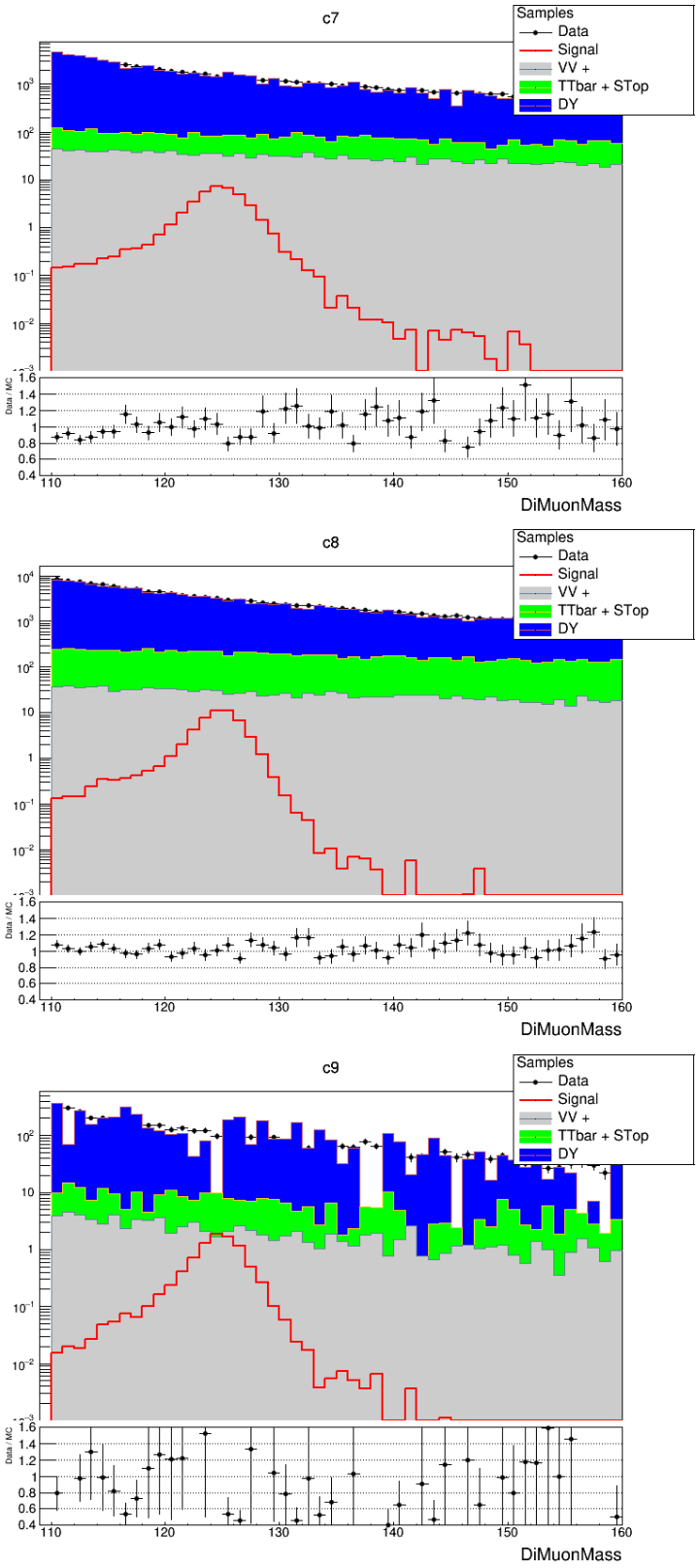


Figure 2.20: Dimuon Mass Distributions for c7-c9 subsets from the Greedy Categorization. Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

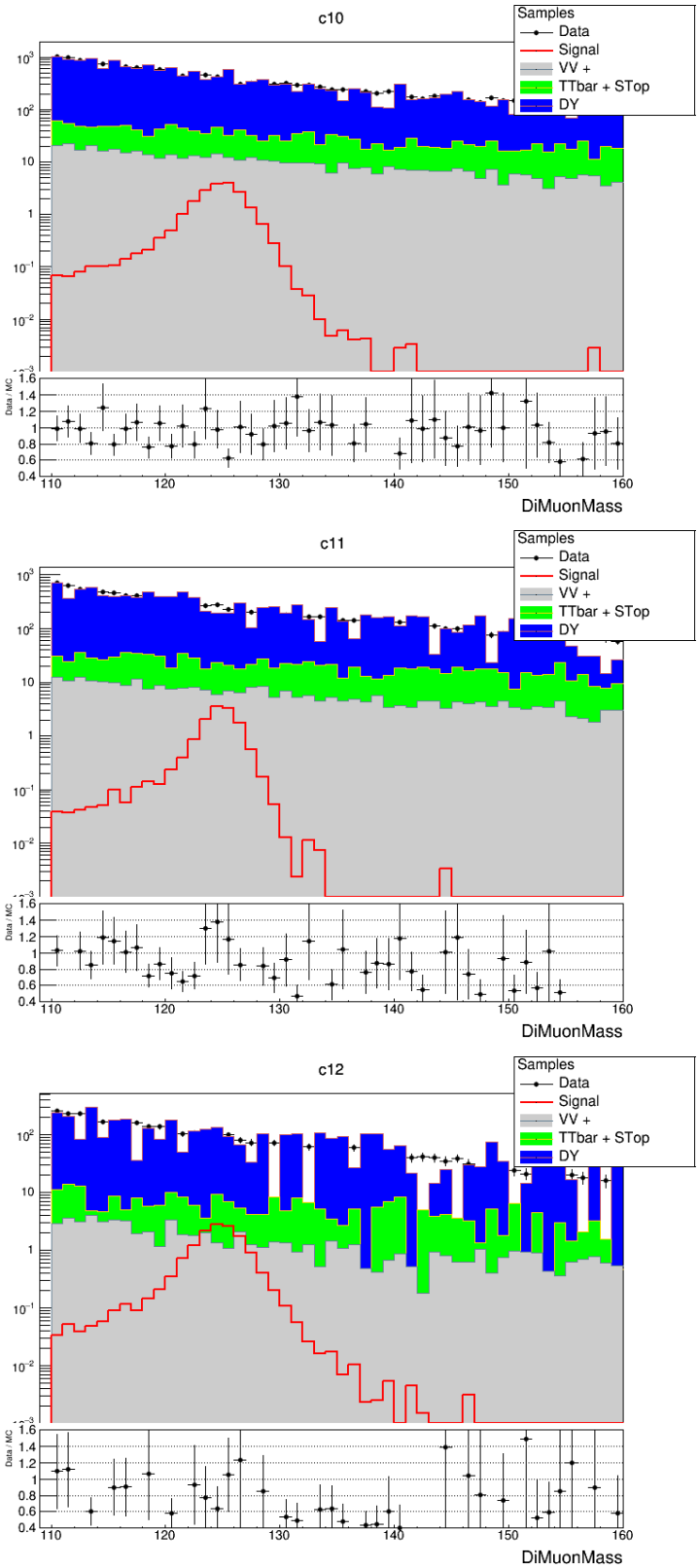


Figure 2.21: Dimuon Mass Distributions for c10-c12 subsets from the Greedy Categorization. Comparison of data (black points) and simulated backgrounds (stacked). 125 GeV Higgs Boson signal is shown as a red solid line.

Figure 2.22 provides a comparison of the data with MC for the kinematic variables, input features, that are used for the signal discrimination. Figure 2.23 shows the same features but compared for the most sensitive category, "c12". Overall, good agreement between data and simulations is observed, especially for the dimuon mass distributions.

Table 2.4 provides a summary of the statistical properties of dimuon mass shapes provided in Figures 2.18 - 2.21. One of the important characteristics of the search, although not obvious, is the signal width, provided in terms of Full Width at Half Maximum (FWHM). It is an important feature because the smaller the width is, the more pronounced the signal is and the less influence the shape of the function used for the background estimation will have on the Standard Model Higgs signal.

2.8 Signal Modeling

To be able to draw conclusions about possible excess of events due to the Standard Model Higgs Boson, a model aiming to explain the to be observed excess needs to be built. Given that the excess is motivated as a bump-like structure near the Higgs Boson Mass, it is natural to model this excess via a composition of Gaussian functions. Each category and production process has been treated individually and only depends on the Higgs Boson Mass.

Both Double (Equation 2.20) and Triple (Equation 2.21) Gaussian forms were tested. The main reason for extending the form up to triple Gaussian (w.r.t. Run I) is to be able to pick up both the possible mass shift due to Finite State Radiation (FSR)

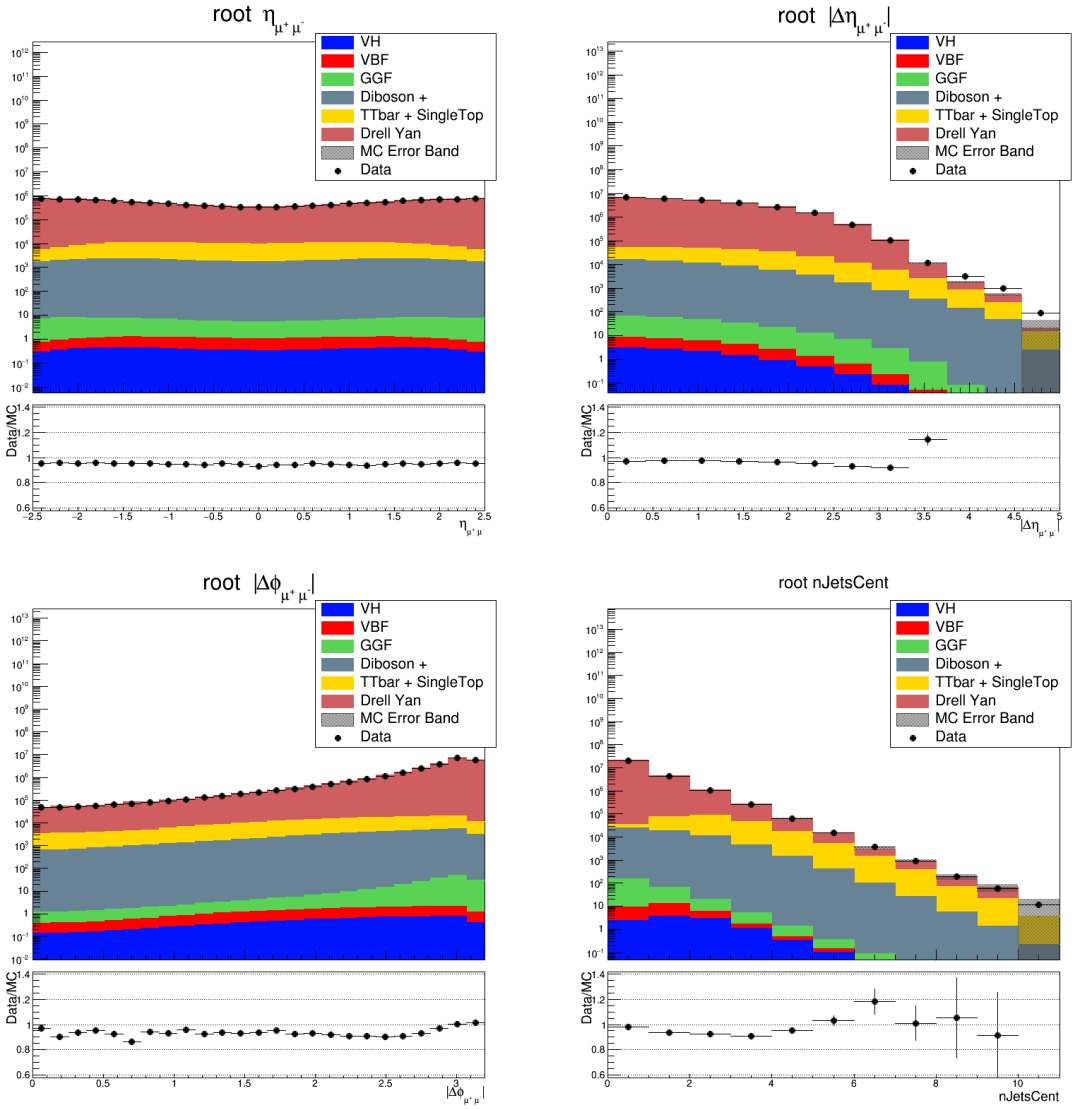


Figure 2.22: Data/MC agreement for the BDT input variables before categorization. η of the dimuon system (top left), $|\Delta\eta|$ of the dimoun system (top right), $|\Delta\phi|$ of the dimuon system (bottom left), number of jets in the central region (bottom right).

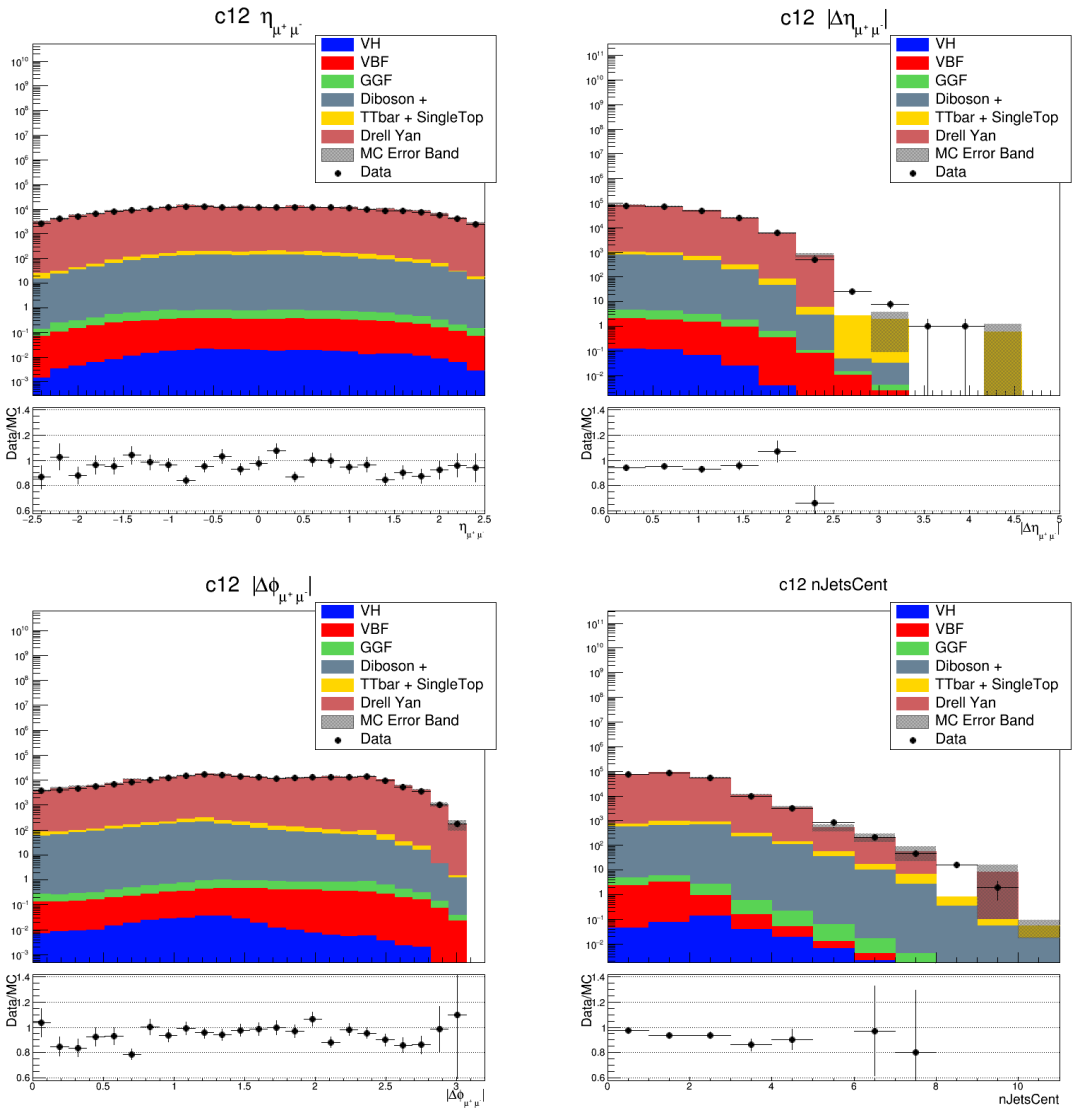


Figure 2.23: Data/MC agreement for the BDT input variables for the most sensitive category. η of the dimuon system (top left), $|\Delta\eta|$ of the dimoun system (top right), $|\Delta\phi|$ of the dimuon system (bottom left), number of jets in the central region (bottom right).

Table 2.4: Comparison of Signal and Background Yields for greedily optimized categories.

Category	Signal FWHM (GeV)	N^S	N^B	$\mathrm{N^S}/\sqrt{\mathrm{N^B}}$
c0	4.4	13.4826	14187.6	0.113193
c1	5.8	3.52888	970.875	0.113255
c2	5.8	14.2112	8248.64	0.156473
c3	6.0	3.82012	779.594	0.136818
c4	3.0	9.92363	1609.08	0.24739
c5	2.8	8.5825	1028.08	0.267671
c6	4.0	14.1224	2371.09	0.290025
c7	4.2	26.3575	6904.16	0.317211
c8	3.6	35.6493	11187.4	0.337043
c9	4.0	6.12457	451.281	0.288305
c10	4.0	13.9349	1667.59	0.34124
c11	3.0	9.59255	765.226	0.346768
c12	4.0	9.58541	420.983	0.467173

and accommodate the broadening due to detector resolution effects. Moreover, the Triple Gaussian form is defined using recursive coefficients to make sure that there are no negative contributions from any of the gaussians.

$$S(x, m_{\mathrm{H}}, \theta) = f\mathcal{N}_1(x, \mu_1, \sigma_1) + (1-f)\mathcal{N}_2(x, \mu_2, \sigma_2) \tag{2.20}$$

$$S(x, m_{\mathrm{H}}, \theta) = f_1\mathcal{N}_1(x, \mu_1, \sigma_1) + (1-f_1)\left(f_2\mathcal{N}_2(x, \mu_2, \sigma_2) + (1-f_2)\mathcal{N}_3(x, \mu_3, \sigma_3)\right) \tag{2.21}$$

where $\mathcal{N}(x, 0, 1)$ is a normal distribution, $\mu_i(m_{\mathrm{H}}, \theta), \sigma_i(m_{\mathrm{H}}, \theta)$ are respectively the mean and sigma of each Gaussian, x is the reconstructed invariant mass of the two muons ($m_{\mu\mu}$), and θ is the list of parametric nuisances.

For each category and each production process (ggH, qqH, WPlusH, WMi-

nusH, ZH, ttH), all of the signal parameters are derived by fitting a given model Double (Equation 2.20) or Triple (Equation 2.21) Gaussian to the dimuon invariant mass spectrum obtained from the MC Higgs signal samples being subject to the same event selection as data and the same categorization procedure. From now on, the choice of Greedy categorization is implicitly assumed.

For the purpose of testing various Higgs Boson Mass hypotheses, three mass points ($m_{\mathrm{H}} = 120\,\mathrm{GeV}$, $125\,\mathrm{GeV}$, $130\,\mathrm{GeV}$) were used, which allow to interpolate in between and probe any mass in the range $[120, 130]\,\mathrm{GeV}$. For a given category and a production process, the procedure to extract model parameters from the signal MC goes as following:

- Start with the invariant mass spectrum for m_{H} of $120\,\mathrm{GeV}$ and perform a binned maximum likelihood fit using initial default parameters.
- Proceed to next point in mass (m_{H} of 125 GeV), by using the same fitted resolution (σ_i) from the previous fit ($m_{\mathrm{H}} = 120\mathrm{GeV}$) and shifted scale ($\mu_i$), by the mass difference, parameters as initial guesses. Perform the binned max likelihood fit again and extract the parameters.
- Proceed to the m_{H} of $130\,\mathrm{GeV}$ and perform the same procedure as for $125\,GeV$.
- This procedure can be applied to any number of mass points.
- Each parameter μ_i, σ_i, f_i can be now interpolated across the mass points, using a spline function.
- At this point, all the parameters (σ_i, μ_i, f_i) are functions of the floating Higgs mass (m_{H}).

Figure 2.24 shows examples of the individual fits of different Higgs Boson mass hypotheses for the most sensitive category, "c12" with the dominant Higgs Boson production mechanism, Gluon Fusion. In both cases, the functional form is not normalized to the expected yield at 36 fb^{-1} at this stage.

Figure 2.25 presents an example of the interpolation of the model parameters for the most sensitive category, "c12". Parameters from the individual fits shown in Figure 2.24 are used and interpolated as a function of the hypothetical Higgs Boson mass.

The last missing piece for the signal model construction is the normalization. Given a dimuon invariant mass distribution for a particular category for a particular production process, the expected yields can be expressed as in Equation 2.22.

$$\text{Yield} = \mathcal{L}\,\sigma(pp \to H)\,\mathcal{B}(\text{H} \to \mu\mu)\,\varepsilon A \tag{2.22}$$

The production cross sections ($\sigma(pp \to H + X)$) for each process and the branching ratio of the Higgs boson to decay into a muon pair ($\mathcal{B}(\text{H} \to \mu\mu)$) are taken from the Yellow Report 4 [34] as centrally provided by the CMS Higgs Combination Group. Efficiency times acceptance (εA) is computed using the MC sample information for a particular production process as:

$$\varepsilon A = \frac{w_{category}}{w_{total}} \tag{2.23}$$

where $w_{category}$ is weighted number of events in that category and $w_{generated}$ is the total number of events generated.

Finally, Figures 2.26 - 2.29 show the results of the interpolation algorithm for

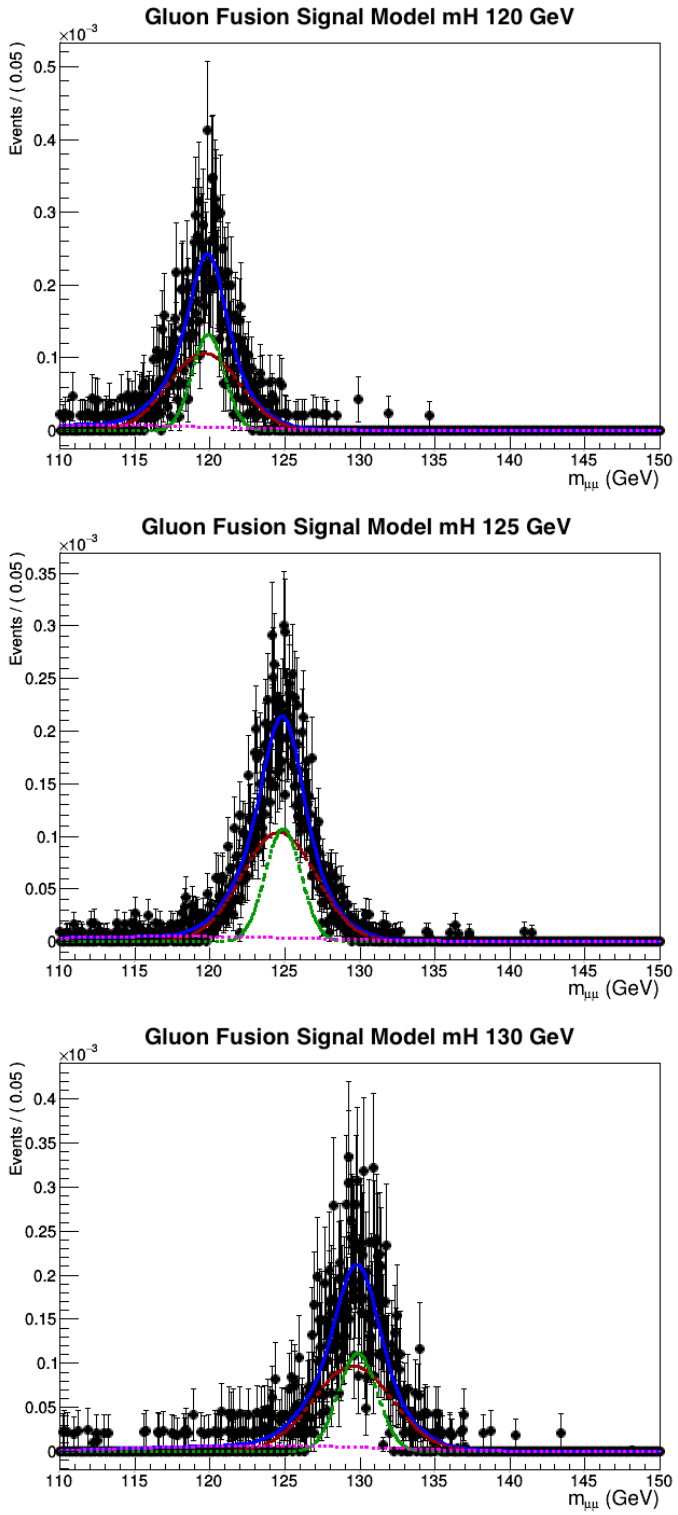


Figure 2.24: Examples of fits for the individual dimuon mass distributions of signal MC for the Gluon Fusion production mechanism for the most sensitive category, "c12". 120 GeV signal (Top), 125 GeV (Middle), and 130 GeV (Bottom).

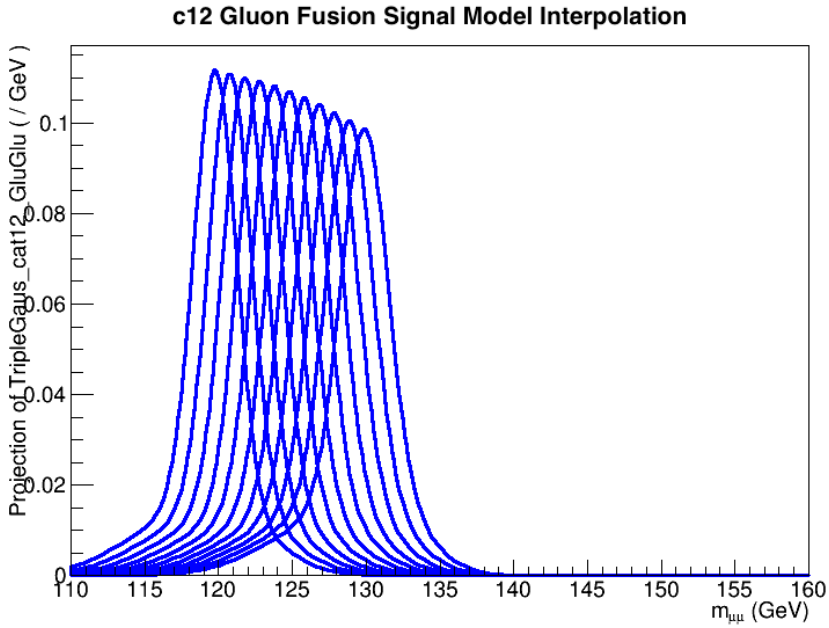


Figure 2.25: Example of the results of the Signal Interpolation for the Gluon Fusion production mechanism for the most sensitive category, "c12".

the 2 dominant Higgs Boson production processes for all of the categories (skipped "c12"): Vector Boson Fusion and Gluon Fusion.

2.9 Background Modeling

Given the highly dominant nature of the background processes, it is important to properly model the background shape as that is the biggest contributor for any physical quantity of interest to be computed. In general, background can be described by means of a falling exponential with some polynomial contribution.

The background is mainly composed of Drell-Yan and top pair production ($t\bar{t}$). MC samples for such processes have been used, but they are not suited for directly

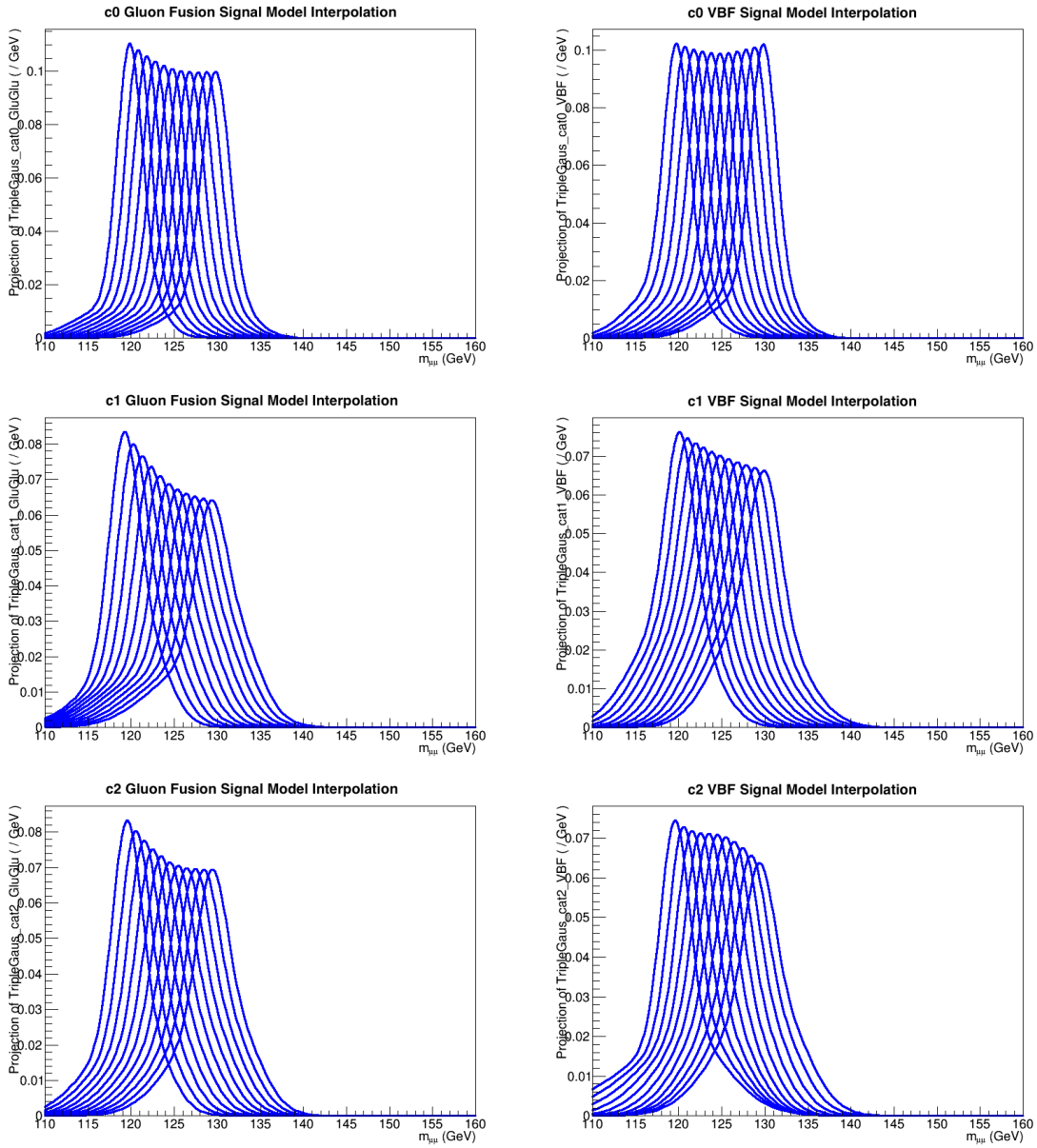


Figure 2.26: Signal Model Interpolation. Models for 120/125/130 GeV built, fitted, and parameters are interpolated as functions of the Higgs Boson mass. Gluon Fusion (left column) and Vector Boson Fusion (right column). "c0" (top row), "c1" (middle row) and "c2" (bottom row).

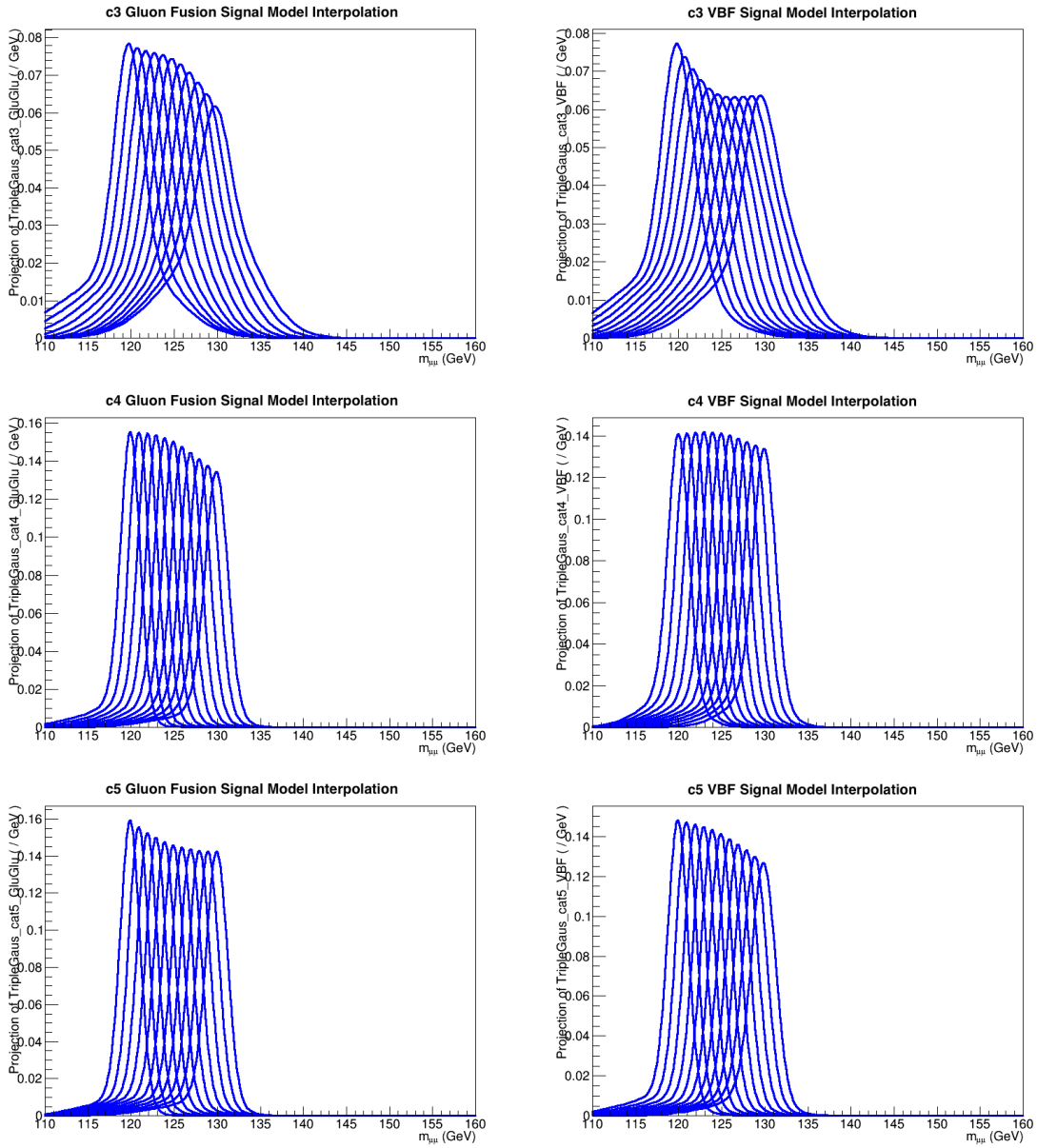


Figure 2.27: Signal Model Interpolation. Models for 120/125/130 GeV built, fitted, and parameters are interpolated as functions of the Higgs Boson mass. Gluon Fusion (left column) and Vector Boson Fusion (right column). "c3" (top row), "c4" (middle row) and "c5" (bottom row).

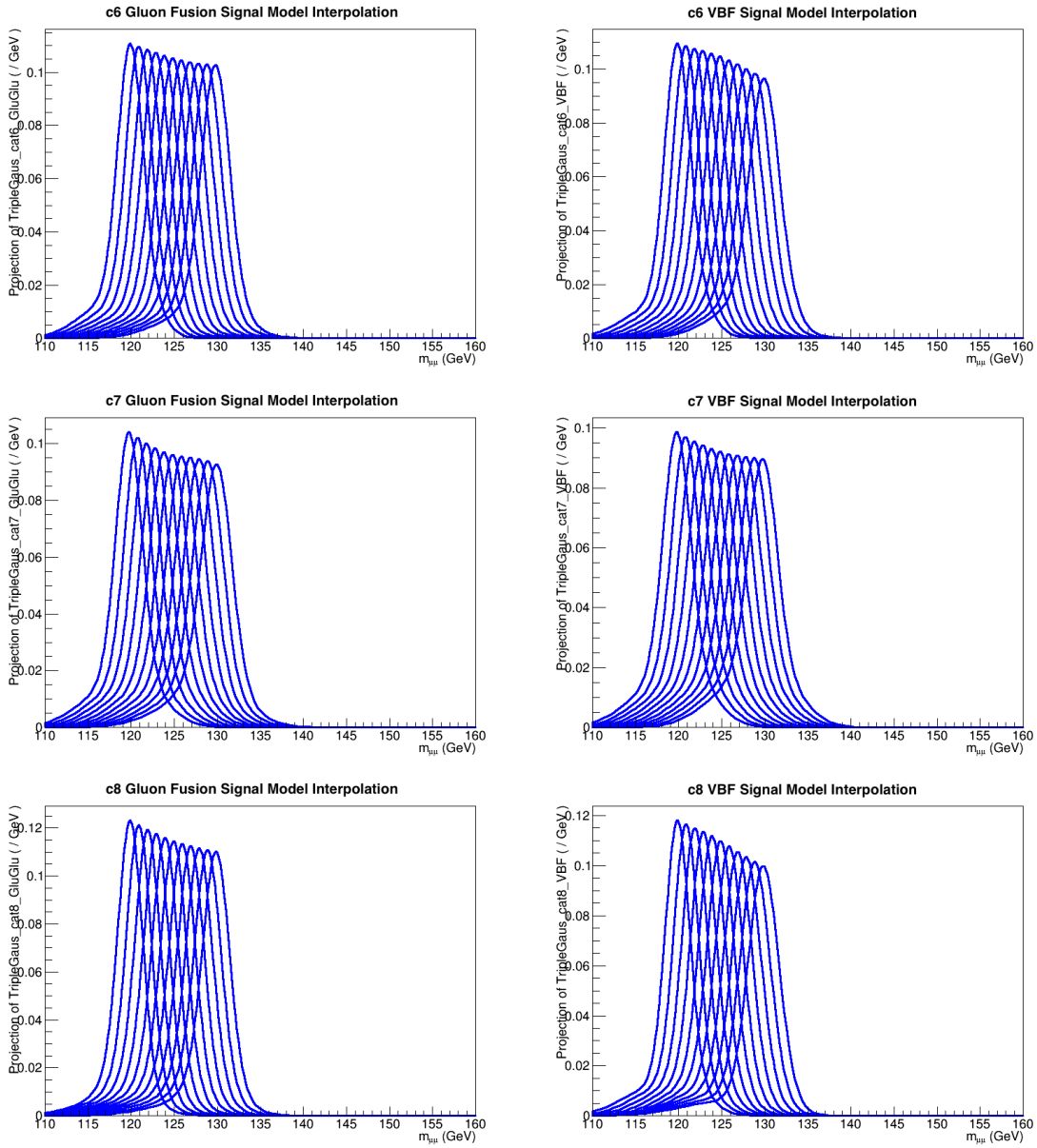


Figure 2.28: Signal Model Interpolation. Models for 120/125/130 GeV built, fitted, and parameters are interpolated as functions of the Higgs Boson mass. Gluon Fusion (left column) and Vector Boson Fusion (right column). "c6" (top row), "c7" (middle row) and "c8" (bottom row).

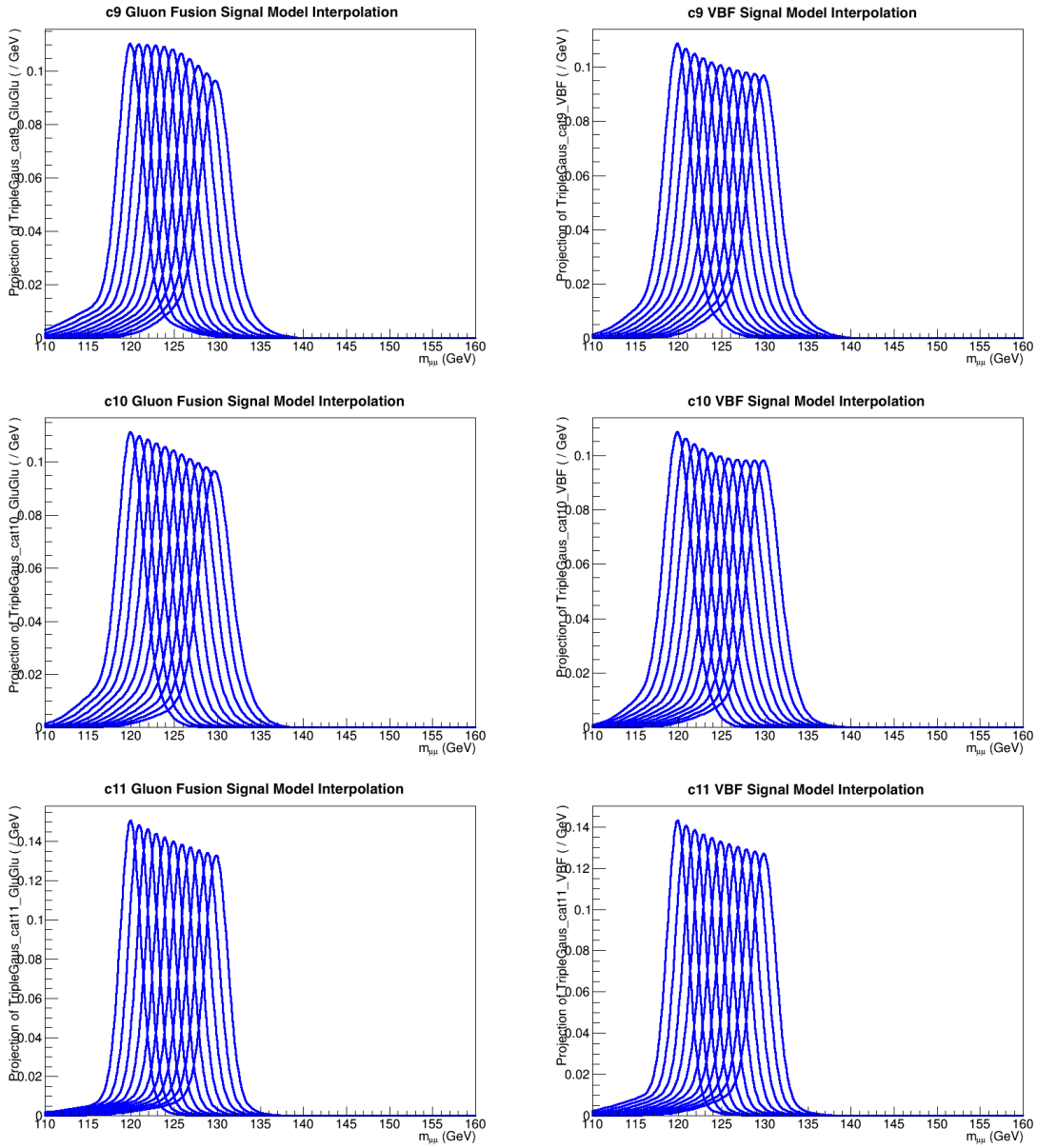


Figure 2.29: Signal Model Interpolation. Models for 120/125/130 GeV built, fitted, and parameters are interpolated as functions of the Higgs Boson mass. Gluon Fusion (left column) and Vector Boson Fusion (right column). "c9" (top row), "c10" (middle row) and "c11" (bottom row).

modeling background because they have little statistical power with respect to the data in the region of interest as well as large statistical uncertainties due to extra terms in QCD or EW expansions (high p_t), as well as resummation (low p_t), pdf and scale uncertainties, and unknown modeling of the long range correlation uncertainties. Therefore, the modeling of background is treated using smooth functional forms.

For the purpose of modeling, two classes of models are identified for consideration: physics motivated models and general purpose series able to describe any smoothly falling functional form (polynomials, sum of exponentials). The first class contains several pdfs whose functional forms are driven by the knowledge of background processes contributing the most (Drell-Yan, $t\bar{t}$). The functions considered are summarized in Equations 2.24 - 2.27.

$$\text{ExpPolynomial:} B(x) = e^{a_1 x + a_2 x^2} \tag{2.24}$$

$$\text{BWZ:} B(x) = \frac{e^{ax}\sigma_z}{(x-\mu_z)^2 + (\frac{\sigma_z}{2})^2} \tag{2.25}$$

$$\text{BWZRedux:} B(x) = \frac{e^{a_2 x + a_3 x^2}}{(x-\mu_z)^{a_1} + (\frac{2.5}{2})^{a_1}} \tag{2.26}$$

$$\text{BWZGamma:} B(x) = f\frac{e^{ax}\sigma_z}{(x-\mu_z)^2 + (\frac{\sigma_z}{2})^2} + (1-f)\frac{e^{ax}}{x^2} \tag{2.27}$$

These shapes have been derived and validated fitting the FEWZ (NNLO generator [35]) generated mass shapes and fitted values for the parameters have been used as initial guesses for the modeling procedure. The second class of functions considered comes from general families which are, in principle, capable of describing any functional form by incorporating more and more terms. Several families are considered: polynomials in the Bernstein basis, power law, and sum of exponentials; summarized

in Equations 2.28 - 2.31.

$$\text{Bernsteins:} B(x) = \sum_{i=0}^{n} \alpha_i [\binom{n}{i} x^i (1-x)^{n-i}] \tag{2.28}$$

$$\text{SumExponentials:} B(x) = \sum_{i=1}^{n} \beta_i e^{\alpha_i x} \tag{2.29}$$

$$\text{SumPowers:} B(x) = \sum_{i=1}^{n} \beta_i x^{\alpha_i} \tag{2.30}$$

$$\text{LaurentSeries:} B(x) = \sum_{i} \alpha_i x^i \tag{2.31}$$

The background modeling procedure involves the construction of an envelope (RooMultiPdf [36]) of functions that can either be used individually or altogether (discrete profile method [37]) for the purpose of fitting the signal strength or setting the upper limit. For the physics-motivated functions, the binned maximum likelihood fit is performed to provide sensible initial guesses and the model gets inserted into the envelope. Although the discrete profile method allows to naturally take into account the number of free parameters in the fit, the computational time it takes scales linearly with the product $n_{\text{cat}} \times n_{\text{family}}$, since it performs one fit for all possible combinations of choices of functional forms in all categories. In order to minimize the time it takes to perform the regression, the number of families in the envelope is reduced by selecting an appropriate order for a family using a procedure known as F-Test, **the Fisher Test**, at 95% Confidence Level. For a given category and for a particular family, the actual algorithm of the order selection for a particular family can be summarized as follows:

- H_0 hypothesis: order n is the true order. To reject this hypothesis, one has to be at least 95% confident rejecting it, in other words: $p - \text{value}(\chi^2, ndf) < 5\%$

- Perform the background only fit for orders n and $n+1$ to the data
- Compute χ^2_n and χ^2_{n+1} respectively. $\chi^2 = -2\log\mathcal{L}$
- Compute the difference in the number of degrees of freedom: $\Delta NDF = NDF_{n+1} - NDF_n$
- $\chi^2 = \chi^2_{n+1} - \chi^2_n = -2\Delta\log\mathcal{L}$. Note, that $\chi^2_1 + \chi^2_2$ is also distributed as a χ^2, but with $NDF = NDF_1 + NDF_2$.
- Compute the χ^2 p-value.
- For p-value less than 5%, reject n and move on to test n+1.
- For p-value greater than 5%, stop and select order n for this category, for this functional family.

Figures 2.30 - 2.33 show the dimuon mass distributions together with various analytic background functions fitted to the data. The binned maximum likelihood fit is performed. The order of a family is chosen using the F-Test procedure described above and varies depending on the category.

2.10 Systematic Uncertainties

All of the systematic uncertainties that come into play in this search, except for just one, are due to the potential effects of mismodeling of the Standard Model Higgs signal. In general several types of potential sources of problems in our model to correct for can be identified: the shape of the SM Higgs Boson signal, the yield of the signal, and various experimental and theoretical uncertainties.

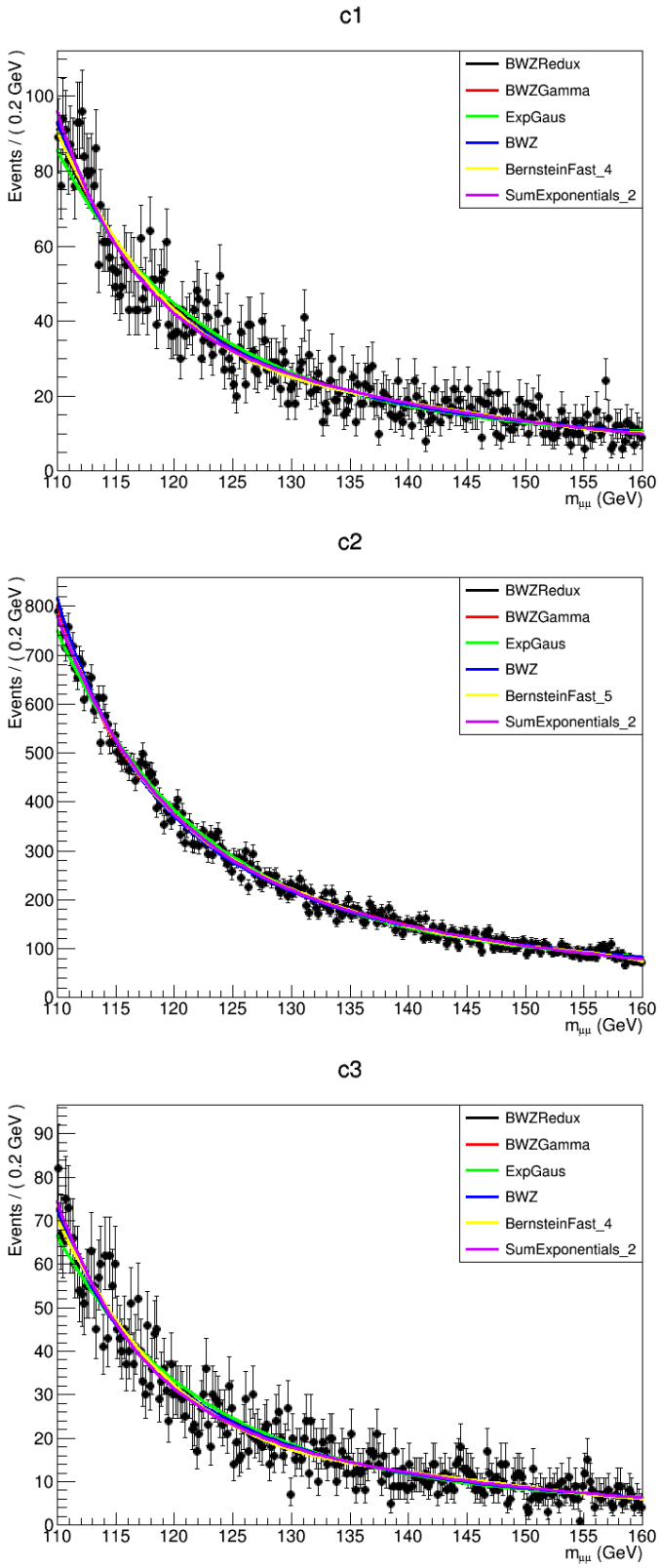


Figure 2.30: Background Modeling envelope of analytic functions fitted to the data (c1 - c3).

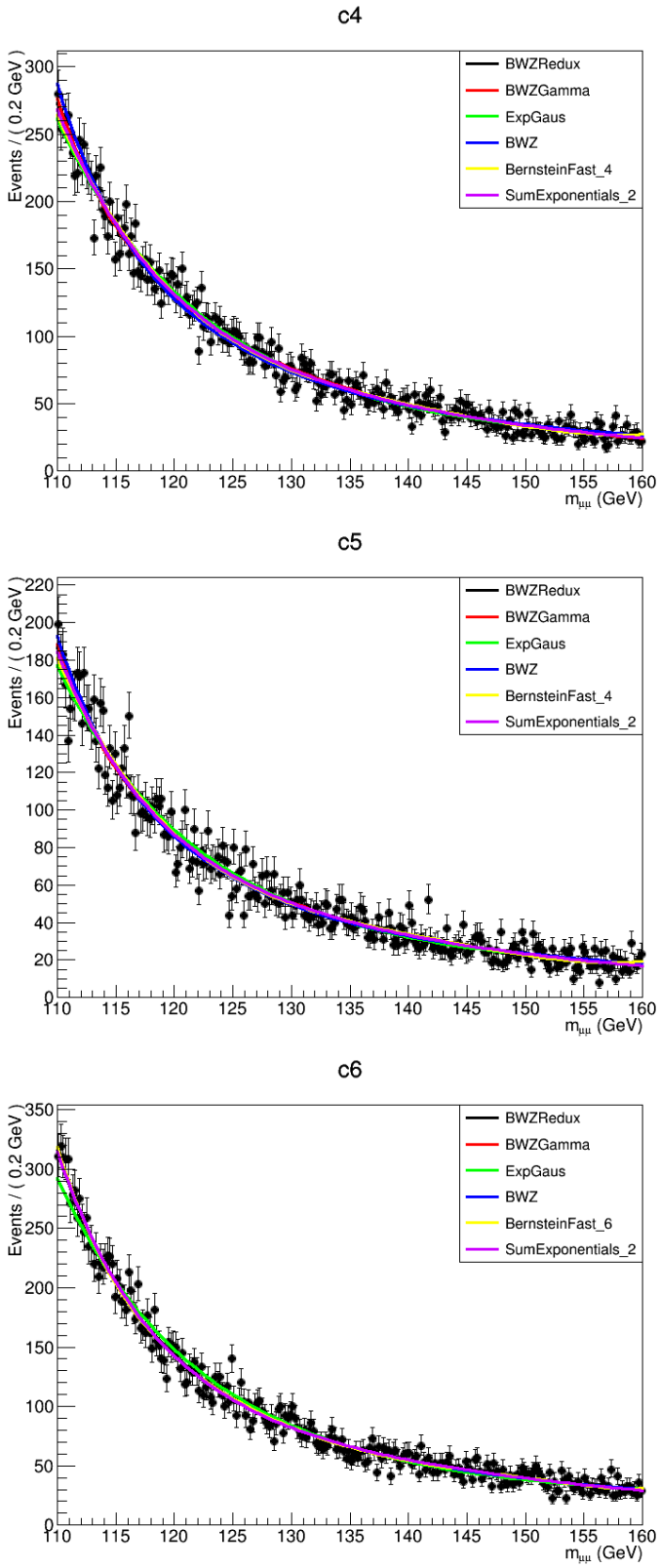


Figure 2.31: Background Modeling envelope of analytic functions fitted to the data (c4 - c6).

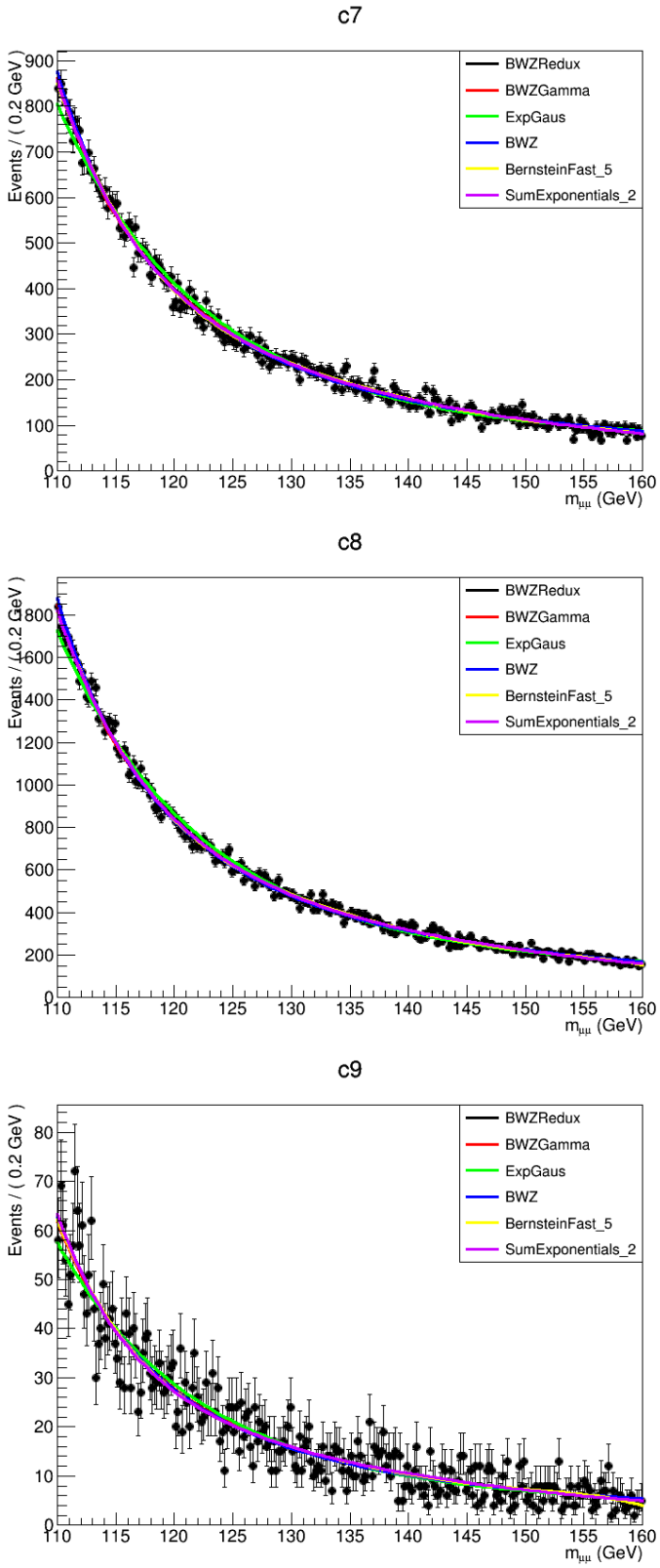


Figure 2.32: Background Modeling envelope of analytic functions fitted to the data (c7 - c9).

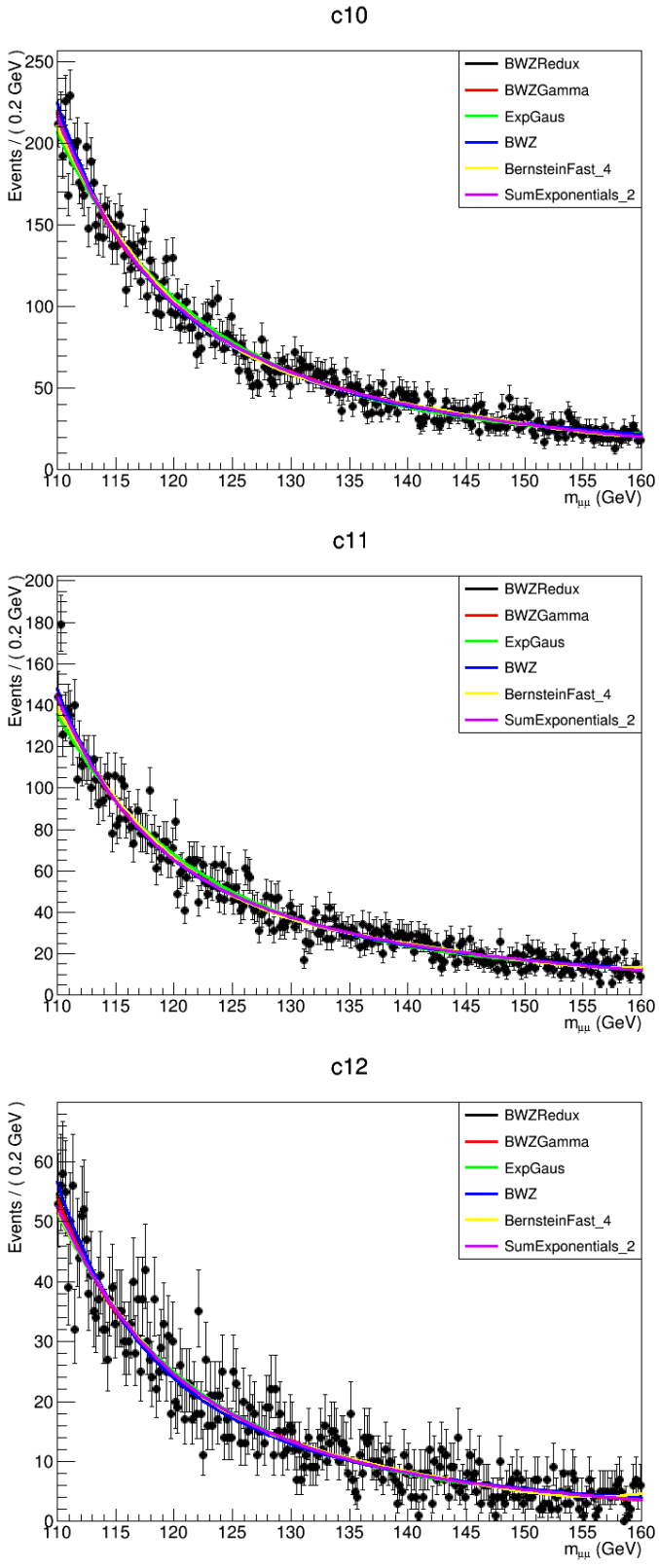


Figure 2.33: Background Modeling envelope of analytic functions fitted to the data (c10 - c12).

2.10.1 Shape Uncertainties

Nuisances responsible for the shape systematic uncertainties are used to address the potential mismatch of the expected signal peak in simulations with the data. There are two possible sources of such mismatch: the mean (the scale) of the signal mass distribution could be shifted or the width (the resolution) could be different.

- **Muon Scale**, Up to 5% of the muon momentum. The potential effect is to change the position of the Higgs signal with respect to its nominal value.
- **Muon Resolution**, up to 10%. There are two important outcomes of this mismodeling. One is that the width of the signal can be different from what it appears to be in the data. And second, as a consequence of the first, the wider the signal is, and noting how much more dominant the background is for the search, the easier it is for background to "eat" the signal peak. In other words, some analytic functional forms are quite flexible in a way that they can contribute to the peaky structure of the signal, causing the improper modeling of the Standard Model Higgs signal and invalid extraction of exclusion limits.

The second point is of particular importance and will be addressed further in the next Section 2.11.

2.10.2 Category Migration

The following nuisances added to the model are primarily responsible for the migration of events across the categories. It has been verified within the statistical power of the MC that no major shape distortions are present.

- **Jet Energy Scale**, up to 6%. After applying the jet energy corrections, the energy scale of the jet is varied as a single nuisance parameter. Given that there are categories with the dominant VBF production process contribution, this nuisance can result in substantial yield variations, especially in the sensitive categories. This uncertainty is the dominant experimental uncertainty in the measurement; i.e. the uncertainty with the largest impact on the parameter of interest among those listed.
- **Jet Energy Resolution**, up to 1 - 2%. The jet energy resolution can also induce category migration due to the non-linearly falling spectrum of the jet transverse momentum. It has a modest impact with respect to the scale and it amounts to up to a few percent across categories.
- **Pileup Reweighting**, up to 1 - 2%. The Pileup Reweighting procedure uses the "minimum bias" cross section to extract the estimated amount of pileup in data. There are mainly two effects of PU: it can reduce the efficiency of the muon selection by an increasing of the amount of hadronic activity close by, and it can promote random clusters of energy to jets.
- **b jet efficiency**, up to 1%. After correcting the efficiency of the b-jets passing the medium working point, uncertainties are assigned. Due to the fact that b-jets are vetoed in the most sensitive categories in order to suppress the $t\bar{t}$ contribution, this uncertainty yields to $\simeq 1\%$ variations.
- **b jet fake rate**, up to 1%. Similar to the above but corrects the efficiency of light-flavored jets to fake a b-jet ($\simeq 1\%$).

- **MC f/r scale**, up to 6%. Factorization and renormalization scale are varied up and down by a factor of two in the MC using the MC weights present in the official production. The most extreme variations are excluded from this accounting ($r = 0.5, f = 2$ and $r = 2, f = 0.5$). This uncertainty yields up to $\simeq$ 6% category migration and does not account for the sample total normalization.
- **MC pdf**, up to 2 - 3%. Parton Distribution Functions (PDFs) are varied using the NNPDF3.

2.10.3 Rate Uncertainties

Rate uncertainties come into play to correct for the possible mismatch in the normalization (the total yield) of the Standard Model Higgs signal. First, nuisances related to the theoretical aspects are assigned and account for where they are relevant for the pdf, scale, and changes of the total rate of the Higgs boson production. They are reported in the Yellow Report 4 [34] and are summarized below:

- **Branching Ratio**, up to 1.7%. Branching ratio of the Higgs boson decaying to a pair of muons ($\mathcal{B}(\mathrm{H} \to \mu\mu)$). Applies to all of the signal production modes.
- **ggH cross section**, 5%. Derived from N3LO, applies only to the gluon-gluon fusion production.
- **qqH cross section**, 2.2%.
- **ZH cross section**, -3.5%, $+4.1\%$.
- **WH cross section**, 2 %.

- **ttH cross section**, −9.9 %, +6.8 %.

Furthermore, there are several nuisances taken into account that come from using experimentally measured quantities that directly affect the total yield of the Higgs signal.

- **Luminosity**, 2.5%. Luminosity measurement has been performed centrally by the respective CMS group.
- **Lepton Scale Factors**, 2%. Lepton scale factors provided centrally by the respective CMS working group.

2.11 Higgs Combination

To perform statistical tests in order to extract physical quantities of interest, the Higgs Combination Package [38] is employed. It allows to perform various tests for each category individually (building a likelihood for a single category) and to combine all of the categories simultaneously as well (single likelihood still, but bins from all of the categories will be used for regression). Given that a search for the SM Higgs Boson is performed, the main objective is to improve the exclusion limits that are to be placed on the Standard Model Higgs Boson signal strength with the data collected during Run II in 2016. In this section the Higgs Combination Package is discussed; it will is be utilized to derive the exclusion limits.

At this stage, the parametric model of our signal has been built (in terms of the floating Higgs Boson mass), the Fisher Test for various families has been performed and the actual envelope of analytic functions that will be used further has been

selected. The next steps can be summarized as follows:

- Build the workspaces and datacards for each category in a proper format as specified by the Higgs Combination Package. A datacard is a textual representation of the model, which specifies what signals, backgrounds, data samples are to be used.
- Validate models using Asimov dataset by injecting a signal of some known strength. Asimov dataset is a sample distribution generated from a given analytical form without applying any Gaussian or Poisson randomization applied to the sampled points.
- Perform the relevant Statistical Tests. The main interest is in the exclusion limits computed using the Asymptotic Approximation [39].

2.11.1 Datacards and Workspaces

The Higgs Combination Package uses datacards and workspaces where models are defined to perform the tests. A datacard outlines our models, yields from data, signal and background processes. In the datacard, all of the uncertainties (nuisances) are listed that can modify the yields of the models. Workspaces are the containers for the programmatically specified signal and background models. The detailed list of inputs that are used in the combination is the following:

- Mass shapes for the actual data for each category.
- Background models for each category. All of the models come inside an envelope (RooMultiPdf) as described earlier. The overall normalization for the

background yield as well as the parameters of each functional form are left completely floating, unless they are meant to be constrained by the model.

- Signal models for each category and for each production process. All model parameters are parametric in the Higgs Boson mass. The purpose is to test various mass hypotheses. It is crucial to note that all of the Signal Model parameters are frozen after performing the procedure described in the earlier Section 2.8. However, they are parametric in the Higgs Boson mass. The Standard Model Higgs Boson yield is fixed and the tests performed with respect to this yield - probing for deviations.
- Nuisances (systematic uncertainties) that influence our measurement (integrated luminosity, pile-up, etc...) have to be considered. Multiplicative nuisance parameters are provided that can modify the Standard Model Higgs yield.

2.11.2 Validity Tests

The very first test of the validity of the model to be used is to perform the tests against the Asimov Dataset [39]. The tests performed can be summarized in the following procedure:

- Given the signal and background models, generate an Asimov dataset with a certain signal strength, μ of 1. This is to be done for each category individually.
- Perform the signal plus background fit to the generated Asimov dataset,
- The fitted signal strength, μ, should return the value that has been injected, 1. If substantial variations are observed, it is a clear sign of an issue with the

model. One of the issues observed during such tests was the situation when background was too flexible and was able to "eat" some of the signal under the Higgs peak. The solution is to increase the granularity of the mass distributions. Figure 2.34 provides an example of a single category test against the Asimov dataset. The mass spectrum shown on the left side of the figure shows the generated dataset overlaid with the fit that was performed with the function from which the dataset was generated. On the right side, the profiling of the metric that is minimized during the fit, $-2\Delta \log \mathcal{L}$, versus signal strength is shown. Very good agreement is observed.

2.12 Results

The Expected 95% Confidence Level upper exclusion limits on the Higgs Boson production cross-section have been derived for the Greedy Categorization. Asymptotic approximation has been used for the actual computation, by following the procedure described in the previous section. For the final results, Rochester corrections have been used; however no significant differences were observed among the two possible sets of corrections. The Triple Gaussian analytic form has been used for the modeling of the Standard Model Higgs Boson signal. For the background, the envelope of analytic functions was used and accounts for the potential bias among the forms across all of the categories.

Table 2.5 summarizes the results for the 95% CL upper limits on the signal strength as a function of the Higgs mass. Eleven different mass hypotheses have

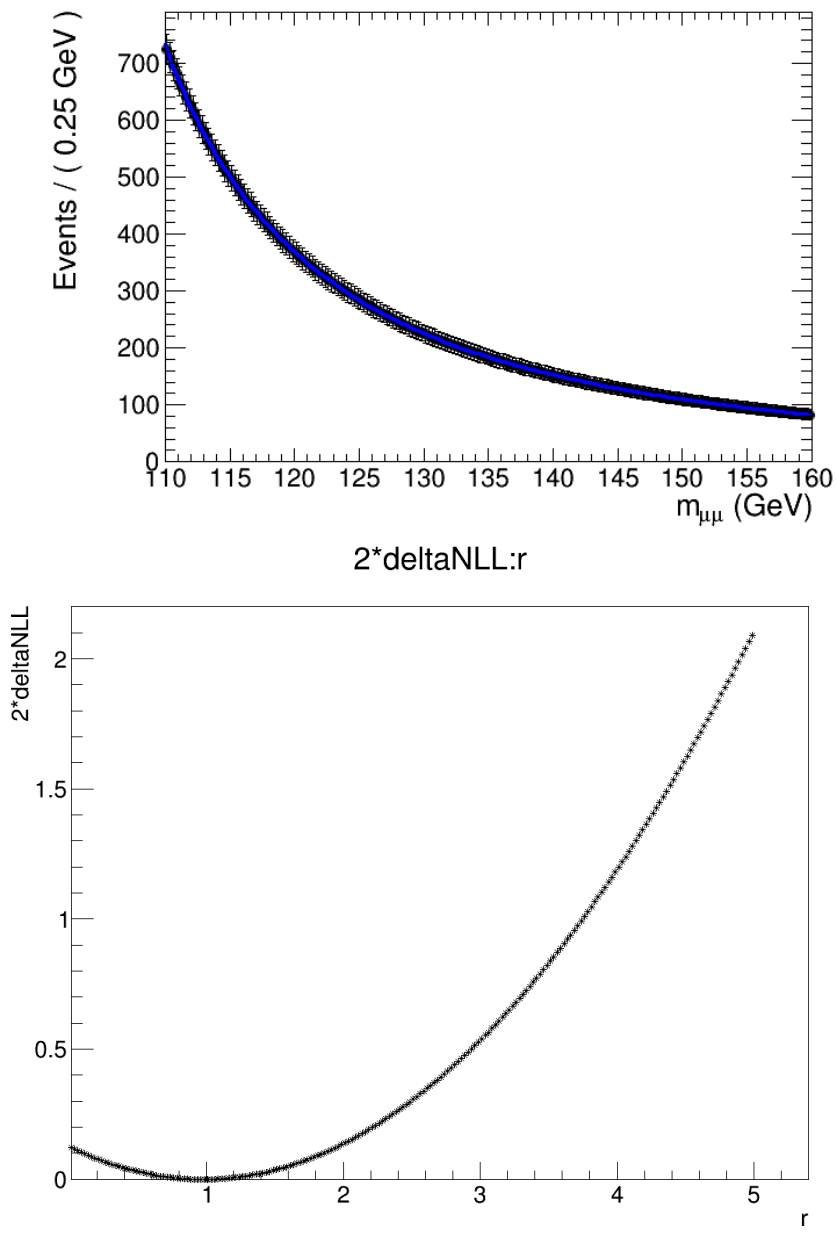


Figure 2.34: The generated Asimov dataset fitted with the same functional form with which it was generated (top). The $-2\Delta \log \mathcal{L}$ versus μ (Higgs Combination Package uses r for μ) profiling (bottom). The minimum is very close to 1 (> 0.99), which confirms the validity of the model. The plot is asymmetric and shifted to the positive values of the signal strength on purpose as negative values are not physical.

been tested, in the range [120, 130] GeV with a step of 1 GeV, and interpolated in between. The observed upper limit on the signal strength does not exhibit significant deviations and is within 1σ from the expectation.

Table 2.5: 95% C.L. Upper Limits on the Standard Model Higgs Boson Signal Strength, $\sigma/\sigma_{SM}(h \to \mu\mu)$.

m_{h} [GeV]	Expected Limits					Observed limit
	-2σ	-1σ	median	1σ	2σ	
120	1.08	1.44	2.02	2.84	3.84	1.90
121	1.07	1.44	2.01	2.83	3.82	1.50
122	1.07	1.43	1.99	2.83	3.82	1.63
123	1.07	1.43	1.99	2.83	3.85	2.28
124	1.07	1.43	2.01	2.84	3.87	2.92
125	1.08	1.44	2.02	2.87	3.91	2.77
126	1.10	1.47	2.05	2.91	3.97	2.37
127	1.12	1.49	2.09	2.98	4.04	2.13
128	1.15	1.52	2.13	3.03	4.09	2.06
129	1.17	1.56	2.18	3.09	4.18	1.94
130	1.20	1.60	2.23	3.16	4.27	1.82

Figure 2.35 graphically shows the dependence of the exclusion limit on the hypothetical Higgs mass. Figure 2.36 shows the 95% CL upper limits on the Higgs Boson Branching Fraction, $\mathcal{B}(\mathrm{H} \to \mu\mu)$, decaying via 2 muons. The Standard Model Higgs Boson production cross-section is assumed.

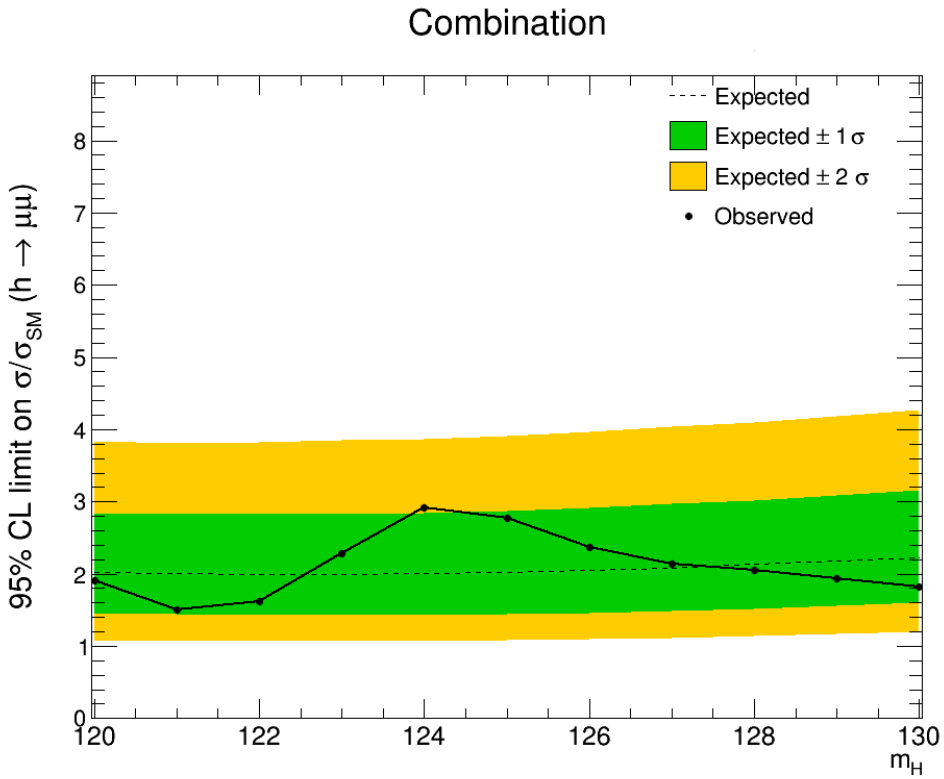


Figure 2.35: 95% C.L. Upper Limits on the Standard Model Higgs Boson Signal Strength as a function of mass. 11 mass hypotheses are tested and interpolated in between the [120, 130] GeV mass range.

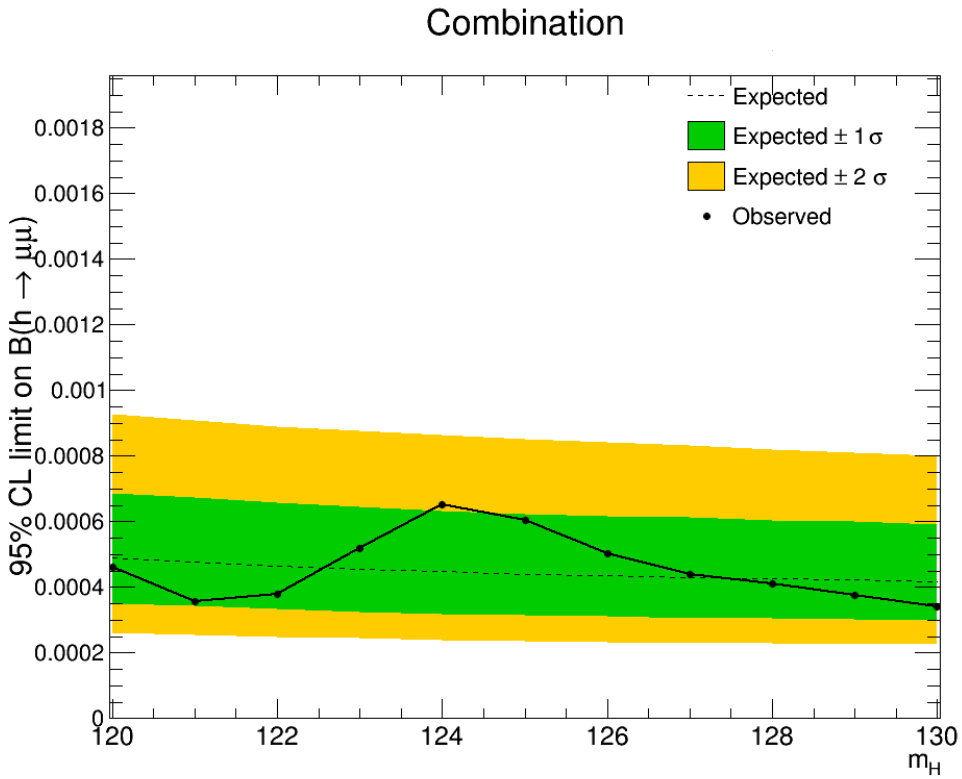


Figure 2.36: 95% C.L. Upper Limits on the $\mathcal{B}(\mathrm{H} \rightarrow \mu\mu)$ as a function of mass. The Standard Model Higgs Boson production cross section is assumed. Eleven mass hypotheses are tested and interpolated in between the [120, 130] GeV mass range.

2.13 Conclusions

In this chapter, a search for the Standard Model Higgs Boson in a dimuon final state has been discussed. Data collected with the Compact Muon Solenoid Experiment from proton-proton collisions at $\sqrt{s} = 13$ TeV over the course of the 2016 Run II datataking campaign has been used and corresponds to an integrated luminosity of 35.9 ± 0.9 fb^{-1}. The observed upper limit on the rate of production of a Higgs Boson has been set for a [120, 130] GeV mass range. For the 125 GeV Higgs Boson, the observed (expected) limit is 2.77 ($2.02^{+0.85}_{-0.58}$) $\times$ SM for 13 TeV. This corresponds to the observed limit of 0.0006 on $\mathcal{B}(\mathrm{H} \to \mu\mu)$ for 125 GeV Higgs Boson.

In comparison with Run I results, where the combined observed (expected) limit for the 7 and 8 TeV data was found to be 7.4 ($6.5^{+2.8}_{-1.9}$) $\times$SM, [18], the search presented in this chapter demonstrates more than a factor of two improvement in the 95% CL upper limit. The new results allow to further constrain the Yukawa fermion coupling parameter of the Standard Model.

CHAPTER 3
CALIBRATION OF THE CMS HADRON FORWARD CALORIMETER

3.1 Introduction and Motivation

During the First Long Shutdown (LS1) of the Large Hadron Collider (LHC), the Hadron Forward (HF) Calorimeters have undergone a substantial hardware upgrade: the new HF Photo Multiplier Tubes (PMTs) have been installed and the part of the readout electronics has been replaced. Therefore, the detector had to be recommissioned in preparation for the Run II datataking campaign. In order to establish the Calibration Coefficients (CCs) for the HF Calorimeter with the new PMTs, three sourcing campaigns have been performed: HF- October 2013, HF+ April 2014, and HF- July 2014.

In this chapter the procedure and results of the performed calibration are presented.

3.2 Description of the HF Calorimeters

The HF calorimeters in the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) cover a large pseudorapidity range, $3 \leq |\eta| \leq 5$, and thus significantly improve jet detection and the missing transverse energy resolution which are essential in top quark production studies, Standard Model Higgs, and all SUSY particle searches [40, 41].

CMS has two hadronic forward calorimeters, HF- and HF+, each being a hollow cylindrical steel structure with an inner radius of 12.5 cm (to accommodate

the beam pipe) and an outer radius of 130.0 cm with the front face located 11.15 m from the center of CMS. Both calorimeters are wrapped by hermetic radiation shielding of 40 cm thick steel, 40 cm of concrete, and 5 cm of polyethylene, with a large steel plug in the back of each detector.

Figure 3.1 displays the cross sectional view of HF calorimeter. Each HF side is azimuthally subdivided into 18 20° wedges composed of 5 mm thick grooved steel plates that perform as absorbers. The plate grooves run parallel to the beam line and are spaced 5.0 ± 0.1 mm center-to-center, each housing a single optical fiber. Each fiber consists of fused-silica, also referred to as quartz, core of 600 µm diameter, layered to an outer diameter of 630 µm with polymer hard-clad, and surrounded to a final diameter of 800 µm with protective acrylate buffer. Each wedge's fibers are bundled to form 24 towers, each with 0.175 x 0.175 ($\Delta\eta$ x $\Delta\phi$) angular span, with exception in angular span for the 2 towers formed closest to the beam pipe. The bundled fibers are held in ferrules that illuminate one end of air-core light guides penetrating shielding of steel, lead, and polyethylene necessary for the readout boxes. The air-core light guides are hollow tubes inlined with highly reflective metal-coated sheets and are coupled to photomultiplier tubes. During LS1 R7525 PMTs were replaced by Hamamatsu R7600 PMTs.

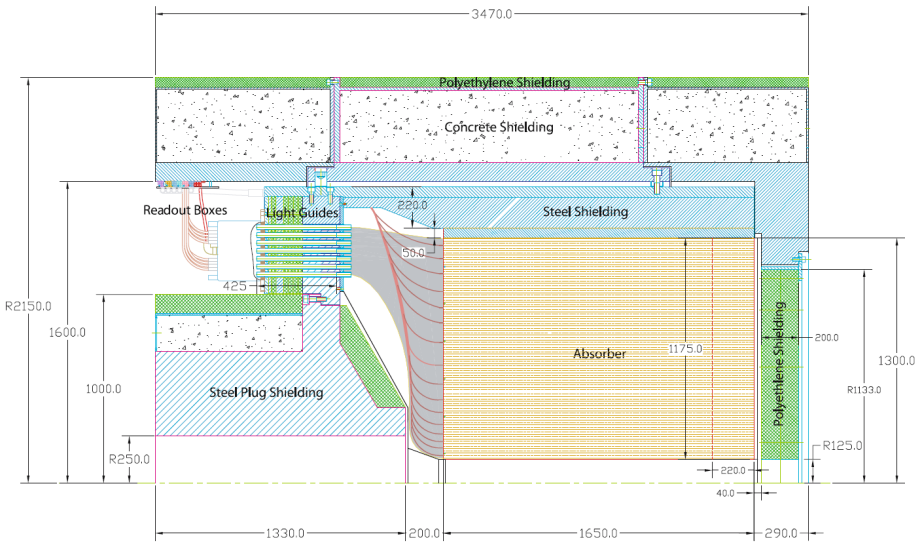


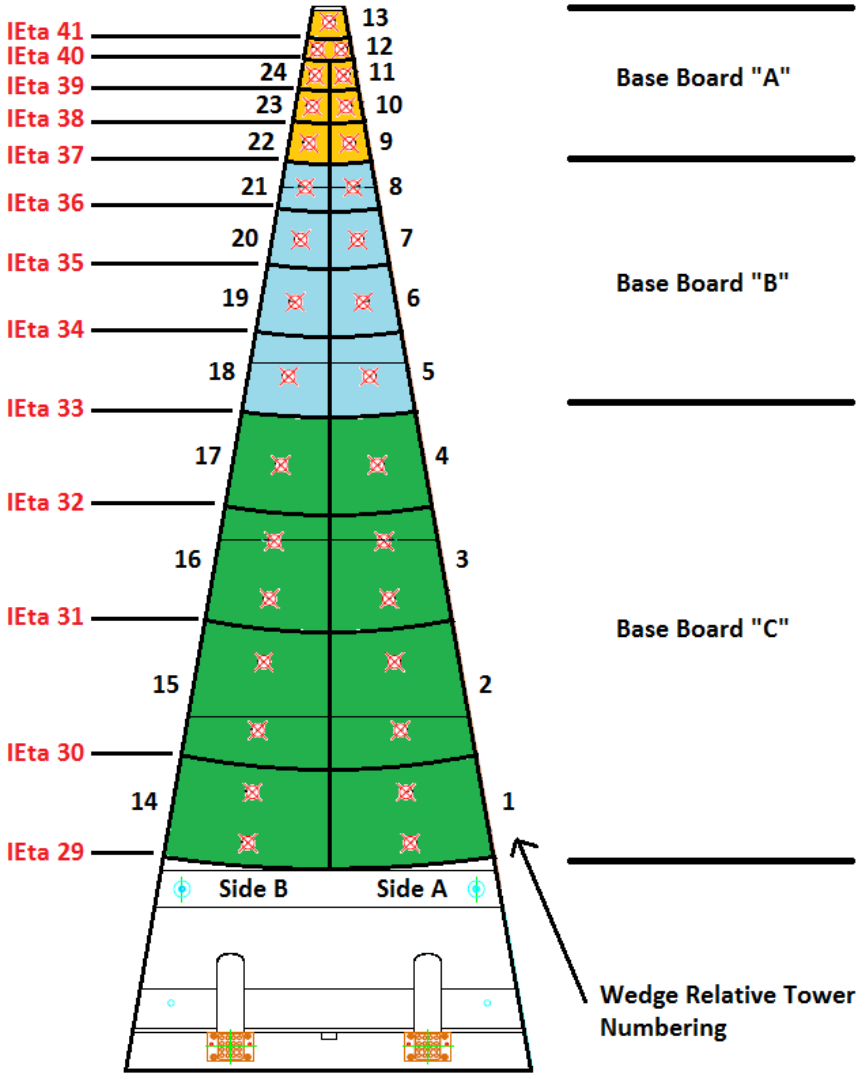
Figure 3.1: Cross sectional view of the HF Calorimeter. IP is to the right.

Cherenkov light forms the signal from charged shower particles above the Cherenkov threshold (E $\geq$ 190 KeV for electrons), optimizing the HF calorimeter's sensitivity to the electromagnetic component of showers [42]. Light incident upon the core-cladding interface above the critical angle (71°) contributes to the signal, although a small fraction of light is captured in the numerical aperture (NA = 0.33 $\pm$ 0.02) and only half reaches the PMT photocathode. Since the calorimeter is most sensitive to the electromagnetic component of showers, two different lengths of optical fibers are inserted into the absorbers and connected to separate PMTs. Half of all fibers in each calorimeter tower extend the full depth of the absorber (165 cm $\approx$ 10 λ_I), while the other half starts 22 cm from the front end. These depths are chosen because a large fraction of energy from electrons or photons are deposited within the

first 22 cm of the absorber's front end, while energy from hadronic showers can be deposited throughout the absorber. The naming convention for these two depths of fibers is EM (or L) for the long fibers collecting the total signal and H (or S) for the short fibers collecting the signal from beyond the first 22 cm. Therefore, with 24 towers per wedge and 2 PMTs per tower, a single readout box housing 24 PMTs services half a wedge (10°).

In order to uniformly collect data throughout each wedge between EM and H fiber bundles, each absorber groove is filled with alternating EM and H fibers. For the purpose of calibrating the energy readout from the calorimeter, a groove in the center of each tower has its corresponding EM or H fiber replaced with a hollow 15 gauge stainless steel tube with inner and outer diameters of 0.97 mm and 1.32 mm, respectively, referred to as a source tube. Due to the geometry and varying dimensions of certain towers, there can be more than one groove housing a source tube in order to calibrate sufficiently, especially for higher η towers.

Figure 3.2 displays the transverse segmentation of a single wedge with all 24 towers and 31 source tubes. It can be seen in Figure 3.2 that the geometry of each tower changes significantly between high η index and low η index; therefore, it is expected that the energy deposited into each tower by the passing radioactive source varies.



(a)

Figure 3.2: Diagram of a single HF calorimeter wedge.

The radioactive source excites a small region of the absorber within its immediate vicinity - 90% of the signal output originates from a region within 3 cm of the

source, with the closest optical fibers to the source contributing most of the output. When performing radioactive source calibration, each tower is analyzed separately. Therefore, geometric correction factors must be applied to the signal output from each tower to account for the energy that escapes into neighboring towers. Monte Carlo techniques are used to account both for the energy leakage of a given tower as well as the relative position of the source tube with respect to the optical fibers (the source tubes replaced either an EM or H type optical fiber). Table 3.1 lists the geometric correction factors for each tower's energy containment with respect to the type of optical fiber and the source tube replaced.

3.3 Infrastructure for Radioactive Source Calibration

3.3.1 Source Driver System

To perform calibration of the calorimeter with a radioactive source, a specialized source driver system has been developed. It includes devices capable of inserting a long thin wire, approximately 11 m long, tipped with a point-like radioactive source into the HF source tubes embedded in the calorimeter. As the radioactive source moves through the HF absorber, gamma rays are emitted and consequently create Compton electrons, which in turn can generate Cherenkov photons inside the quartz fibers if they exceed the Cherenkov threshold. With the new HF PMTs, the quantum efficiency is such that 30% of the photons reaching the PMT cathode face will actually be converted into the readout signal.

The thin wire, referred to as the source wire, used by the system is made

Table 3.1: Geometric correction factors for each tower's energy containment for a radioactive source passing through a given source tube. The value depends on what source tube contains the radioactive source and whether there is a match between the type of optical fiber that the source tube replaced.

Tube	$g(LS/SL)$	$g(LL/SS)$	Unc.
1(14)A	0.917	0.840	0.017
1(14)B	0.909	0.833	0.017
2(15)A	0.945	0.866	0.009
2(15)B	0.938	0.859	0.009
3(16)A	0.925	0.846	0.013
3(16)B	0.908	0.829	0.015
4(17)	0.930	0.851	0.002
5(18)	0.889	0.810	0.003
6(19)	0.847	0.772	0.004
7(20)	0.789	0.713	0.006
8(21)	0.715	0.633	0.018
9(22)	0.646	0.572	0.006
10(23)	0.597	0.516	0.013
11(24)	0.498	0.425	0.023
12A(B)	0.413	0.331	0.056
13	0.588	0.510	0.023

of stainless steel and has inner and outer diameters of 0.406 mm and 0.711 mm, respectively. The end of the wire that penetrates the HF absorber, the front end, is melted shut, shaped into a bullet nose, and chemically plated to reduce friction. The point-like radioactive source is inserted into the opposite end of the wire, the back end, and held in place against the front end by a fine steel piano wire. The outer source wire and inner piano wire are then crimped together at the back end in order to fix the position of the radioactive source.

A Lexan polycarbonate reel, belt driven by a DC reversible electric motor, is

used to insert or retract the source wire into or out of the calorimeter's source tubes. The calorimeter's source tubes are coupled to acetal plastic tubing to mediate the transfer of the source wire from the source driver into the absorber. The transition between the plastic tube and the metal source tube typically involves small-angle conical holes in brass that channel the source wire to the source tube.

The driver system also contains an additional electric motor that functions to select the source tube into which to direct the source wire, an action referred to as indexing. The position of the radioactive source, referred to as the reel position, is provided by an optical rotary encoder read out by industrial batch counters. Typical speeds at which the radioactive source may be inserted or retracted by the driver are between 5 and 15 cm/s. Figure 3.3 displays the source driver configuration.

3.3.2 Description of Data Acquisition

The PMT analog signals are read out by QIEs, standing for charge (Q), integration (I), and encode (E) [43]. Each such QIE has 6 differential inputs, which allow to digitize 6 PMTs simultaneously. Differential inputs are used to subtract any externally induced noise in signal cables. Each QIE has also 1 fiber-optic output, which transfers the digitized information to HTRs (HCAL Trigger and Readout) [44]. The output of 8 QIEs is further bundled up into 1 fiber and digitized information is carried to a HTR. Therefore, each HTR receives input from 48 PMTs: 24 PMTs for each tower's HAD channel and 24 PMTs for each tower's EM channel.

A large dynamic range in PMT signal processing is achieved using a multi-

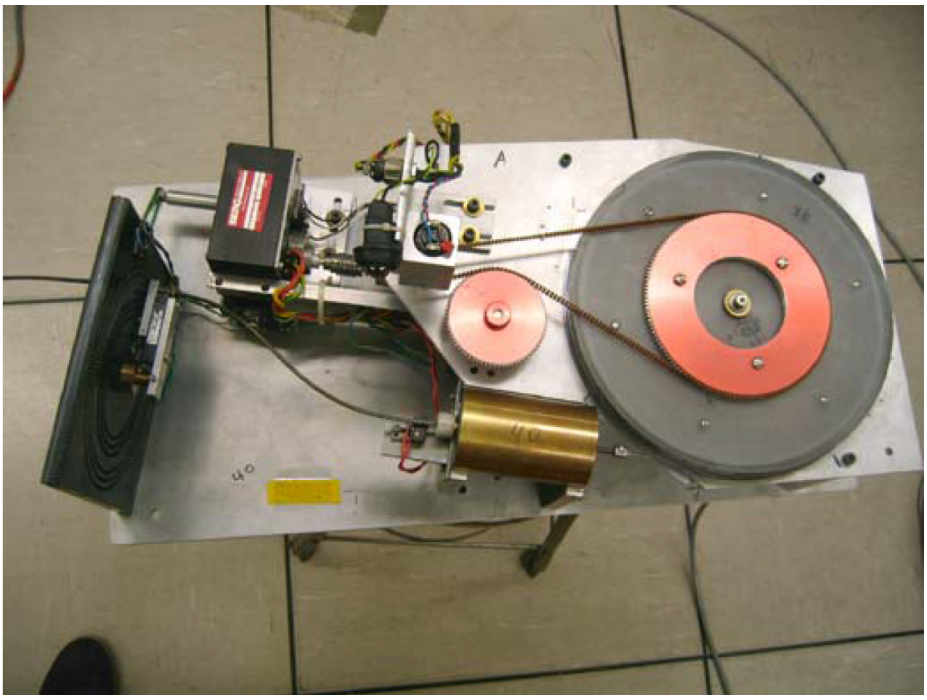


Figure 3.3: Source Driver System.

range technique. The input current is simultaneously integrated on all four ranges, and comparators are used to isolate the lowest range that is not at full scale. The selected voltage representing the integrated charge is then put through an on-chip piecewise linear Flash ADC (Analogue to Digital Converter), with bins weighted according to the time slice (TS) firmware used. For 1 TS, the first 15 bins are weighted by 1, the following 7 bins get a weight 2, the next 4 bins weighted by 3, then 3 bins are weighted by 4, the last 3 bins get a weight 5, providing a range from 0 to 63 ADC counts, with the last bin (overflow bin) containing all charge above 63 ADC counts. For 2 TS, 10 bins are weighted 1, 6 bins weighted 2, 5 bins weighted 4, 5 bins weighted 8, 3 bins weighted 16, 2 bins weighted 32, and 1 bin weighted 62, providing

a range from 0 to 194 ADC counts, with the overflow bin containing all charge above 194 ADC counts. Operations are time multiplexed and pipelined to allow signals to settle and to make the reset interval the same as the integration interval. Clocking is provided at the frequency of 40 MHz, with a latency of 100 ns, as the pipeline is four clock cycles deep. Each QIE contains a set of four capacitors, and only one capacitor is acquiring charge during a given clock cycle (25 ns). The output is a 5-bit mantissa representing the voltage on the particular capacitor, and a 2-bit exponent indicating the range represented by a coded address of the capacitor.

The firmware implemented during radioactive source data acquisition uses a given TS histogramming mode and operating voltage (OV). In the histogramming mode, the absorber's response charge is represented across 32 histogram bins, with the final bin, Bin 32, being the overflow bin containing all data with charge beyond the particular range. For 1 TS firmware, each TS was recorded in the histogram. For 2 TS firmware, 2 time slices (25 ns each) are summed and the total was recorded in the histogram. To accommodate the 2 TS maximum outputs, the expanded QIE bin range described above was used with respect to 1TS.

Each capacitor from a QIE set fills a separate histogram at the sampling rate, i.e. latent cycle, and once 6.5535×10^4 samples are collected the histogram is read out by the DAQ (DAQ is a generic acronym used to signify the full chain of data acquisition), two HTR fibers from each HTR half at a time. Data is saved into event format. Each event containing the histograms read out by all powered HTR halves - a given PMT's output is stored in every fourth event. This means that each event

represents 6.55 ms of source data. For 1 TS, each sample was collected over 25 ns, and for 2 TS each sample was collected over 50 ns.

3.4 Sourcing Campaigns

A ^{60}Co radioactive source, RP4118 with a measured activity of 83.46 ± 0.10 MBq on 14 September 2012, was used to produce the signal readout during the three sourcing campaigns: HF- in October 2013 with the old PMTs still installed; HF+ in April 2014 with new PMTs installed; and HF- in July 2014 with new PMTs installed. First, sourcing with the old PMTs was performed to extract the source energy deposition in terms of HF calibration during Run 1. Second, new PMTs were installed and commissioned. Finally, the sourcing was performed with the newly installed hardware and previously extracted source energy deposition was applied to compute the HF gains. Several OVs, which are different from the Voltage that HF is being planned to be operated at during Run II (OV2), were used depending on the histogramming mode (1 TS or 2 TS) as it was important on one hand to take into account the limitations of the QIE range. On the other hand, we have to have sufficient PMT gain not to lose events under the pedestal peak. All HF gains were adjusted for the difference in OVs.

3.4.1 2013 HF- Campaign

Quadrants 1 and 4, also referred to as the near side of HF- containing wedges 1-9, were exposed to the radioactive source on 23 - 27 October 2013, during which the activity of the source was 72.11 ± 0.15 MBq. The original PMTs were used during

this campaign in order to obtain the energy deposition within each tower from the provided radioactive source, as well as to provide useful data for future studies on radiation damage of the HF calorimeter fibers precipitating from Run I of the LHC. Data from both quadrants were recorded with the 1 TS firmware at OV1 (same OV was used during Run I), in addition to some data recorded with 2 TS firmware at OV1+100. For this specific campaign the overflow bin must not be included in the analysis because certain entries are unrelated to the response charge but instead are indicators of firmware error codes.

Data recording was set to trigger while the radioactive source was within 500 mm of a source tube. The source driver inserted the radioactive source into a source tube at a speed of 10 mm/s, held the source at rest approximately 800 mm from the source tube's back end long enough to collect reasonable statistics while the source is in the middle of the given tower, and then continued to insert the source until it reached the back end of the tube. The source is then immediately retracted from the tube, but in the process the source driver system must spool the source wire by periodically alternating from retracting to extending the wire to prevent tangles. Overall, approximately 7-8 minutes of data was collected for each source tube, with the majority of the statistics being recorded while the radioactive source was held stationary in the middle of each tower.

To help better illustrate how the data-taking process went and what kind of objects the analysis is performed with, refer to Figure 3.4. The first graph shows the response of a single channel as the source is moving along the source tube inside the

tower. Note that source enters from the “Back” and moves to the “Front”. That is because this definition is provided with respect to the CMS Center. Second distribution shows examples of ”signal” and “background” histograms: charge distribution when source is moving inside the tower and another charge distribution when source is moving in some other tower sufficiently isolated from the recorded tower.

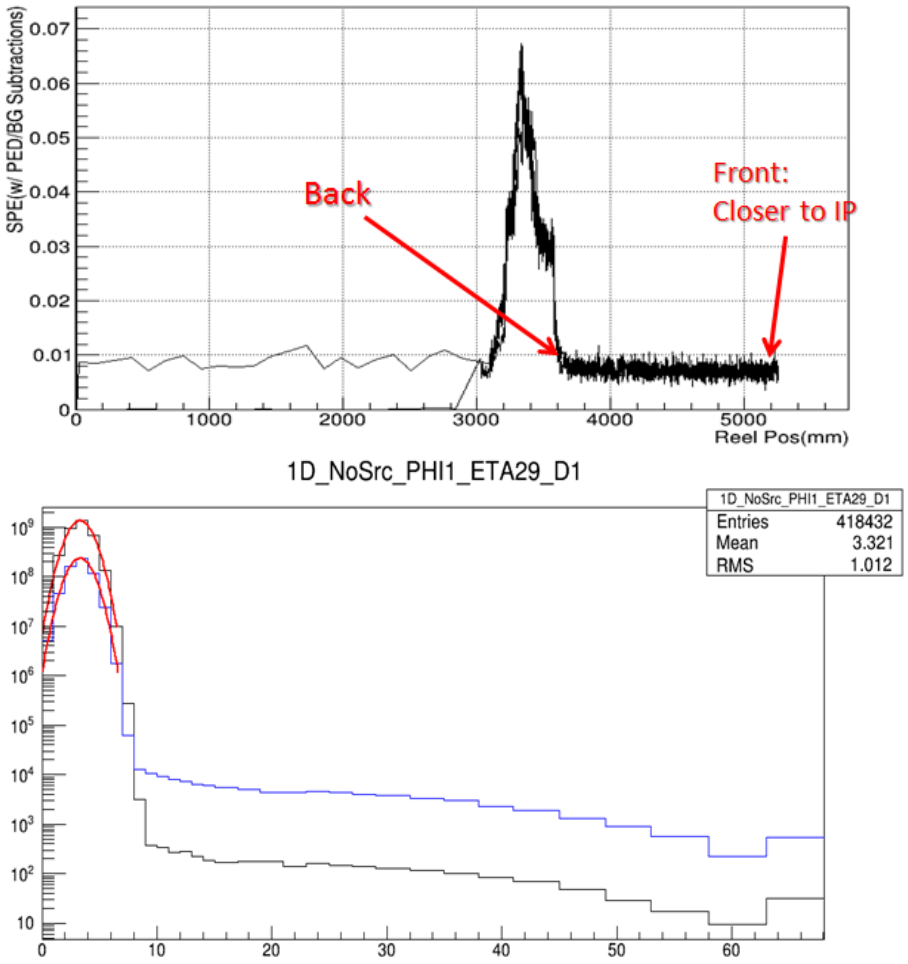


Figure 3.4: (a)Source Signal vs Position (b) 32-bin histogram - A basic object of the readout per a given channel.

3.4.2 2014 HF+ Campaign

All four quadrants of HF+ were exposed to the radioactive source on 22 - 27 April 2014, during which the activity of the source was 67.56 ± 0.18 MBq. At this stage of HCAL upgrades, the new PMTs were installed and functional. All four quadrants were sourced using the 1 TS firmware at OV1. In addition quadrants 1 and 4, also referred to as the near side of HF+ containing wedges 1-9, were sourced with 2 TS firmware at OV1+100. In the histogramming mode, the absorber's response charge is represented across 32 bins, but for this campaign the overflow bin could be included in the analysis because previous issues with firmware error codes were rectified to exclude any nonphysical entries from the 32 bins. Also, during this campaign, the radioactive source was extended and retracted into and out of the source tubes without being held stationary at any point - this procedure differs from that used during the 2013 HF- campaign.

3.4.3 2014 HF- Campaign

All four quadrants of HF- were exposed to the radioactive source on 1 - 11 July 2014, during which the activity of the source was 65.81 ± 0.21 MBq. As with HF+, the new PMTs were installed and functional in HF- by this time. All four quadrants were sourced using the 1 TS and 2 TS firmware, separately, at OV1 and OV1+100, respectively. For this campaign, the overflow bin could be included in the analysis because previous issues with firmware error codes were rectified to exclude any nonphysical entries from the 32 bins. Also, during this campaign, the radioactive

source was extended and retracted into and out of the source tubes without being held stationary at any point - this procedure differs from that used during the 2013 HF- campaign.

3.5 Analysis Procedure

The main objective of the HF sourcing analysis is to transfer the HF energy calibration used in Run I to the new hardware to be used during Run II. More specifically, it is to compute calibration coefficients (so called "HF Gains", not to be confused with PMT gains), validate them, and upload the results to the offline database for further use. All the data files collected during these campaigns are stored on disk at CERN in the EOS storage system.

The general content of the data files is reduced to include pertinent information for calibration, such as the indicies of each channel powered during the campaign, a two-dimensional array containing each channel's output signal and capacitor ID (CapID), the name of the source tube being sourced during any given event, the position of the radioactive source with respect to the source driver, etc. The reduced data file is still substantially large given that many wedges were connected to power during each campaign, each providing output signals for its 24 towers at two channels per tower (EM or H).

The output (EM and H channels) from the tower containing the radioactive source are considered signal data, while the background data is collected for each tower while the radioactive source is sufficiently far away. To guarantee that the

radioactive source is sufficiently distant from a given tower measuring background data, the following requirements for each tower's η index (IEta) for a wedge half with a given ϕ index (IPhi) were considered:

$$\begin{aligned} &\text{Source in IEta} = 29\text{: Record Background for IEta} \geq 34, \\ &\text{Source in IEta} = 39\text{: Record Background for IEta} < 34. \end{aligned} \tag{3.1}$$

Therefore, for a given channel within a given tower and wedge, signal and background data are collected separately based on the location of the radioactive source described above. Provided that histograms for each channel are being read out by the DAQ at a rate of every four events, and each event contains all four capacitors' histograms for each QIE, loop over all events recorded for a given channel and sum together all four capacitors' histograms. To eliminate extraneous data resulting from energy deposition in the optical fiber bundles where the radioactive source has yet to fully penetrate the absorber, restrict the loop over events when source is positioned strictly inside of the calorimeter: [start + 300 mm, end - 300 mm], where start and end are the distances from the source driver at which the given source tube begins and ends within the absorber, respectively. The reason to also exclude data when the radioactive source nears the Tube End position is due to extraneous data resulting from some optical fibers extending beyond the length of the absorber. This tail end cut also helps in unifying signal data definitions between the EM and H channels, given that the H fibers are shorter in length than the EM fibers.

In combining the histograms from all four capacitors within an event, one can then study the newly constructed total histogram as a function of reel position of the radioactive source within the absorber - this ability is crucial for radiation damage studies within the HF calorimeters. However, for the intent of calibrating the new PMTs installed for Run II, the choice is made to focus on comparing the total energy deposition recorded within a given tower with the expected energy deposition provided from the radioactive source. Therefore, we construct a 32 bin histogram, with similar ADC charge bin ranges as provided by the firmware, that stores the overall sum of all events' 4-capacitor histograms according to the following:

$$N_j = \sum_{i=l}^{m} \sum_{CapID=1}^{4} n_j \tag{3.2}$$

where N_j and n_j are the content of the j^{th} bin of their respective histograms, and i is the event with l and m restricted according to go only over events with source position within the tower as defined previously.

The mean ADC charge is defined separately for signal and background data of a particular tower by averaging over the bins of the total histogram from all events,

$$\langle Q \rangle = (\frac{1}{N} \sum_{j=1}^{31} N_j q_j) - P, \tag{3.3}$$

$$N = \sum_{j=1}^{31} N_j$$

where $\langle Q \rangle$ is the mean ADC charge for the total histogram, N is the total number of entries within the bin range [1,31], q_j is the central ADC charge for the j^{th} bin, and P is the Gaussian mean of the pedestal region. All sourcing histograms have a clear feature of a pedestal peak. That is due to the fact that there is no actual trigger during histogramming mode. Every TS data gets recorded. Therefore the number of events that fall under the pedestal is several orders of magnitude larger than those in the region beyond the pedestal, and the μ and σ of the pedestal are dominated by histogram's mean and width. The charge due to pedestal, P, is computed by fitting the histogram within the pedestal region [μ - 4σ, μ + 4σ] (if lower bound falls below 0, take 0 as the min value) with a Gaussian function and extracting the function's mean.

To determine the charge deposition for a given channel of a tower, $\langle Q \rangle_c$, the background mean ADC charge, $\langle Q \rangle_c^{(b)}$, must be subtracted from the signal yield, $\langle Q \rangle_c^{(s)}$, as follows:

$$\langle Q \rangle_c = \langle Q \rangle_c^{(s)} - \langle Q \rangle_c^{(b)}. \tag{3.4}$$

To account for energy leakage, due to the constraints on geometric containment of the radiated energy from the radioactive source within the tower, the geometric correction factors must be applied, G_c , listed in Table 3.1 as follows:

$$\langle Q \rangle_c^{Geom} = \frac{\langle Q \rangle_c}{G_c} \text{ADC/25ns}. \tag{3.5}$$

Up to this point, the analysis procedure was identical for both 2013 and 2014 Sourcing Campaigns. However, to proceed further with calibration (calculating the actual HF Gains), the source energy deposition had to be extracted from 2013 data.

3.5.1 Extracting Source Energy Deposition

The source energy deposition is the amount of energy deposited by the source in units of GeV/25 ns. The actual source signal is measured (computed) in ADC/25 ns. Therefore, a conversion factor GeV/ADC, which comes from Run I Conditions, is needed. To extract Run I HF Gains and QIE Slopes from the Conditions DataBase (CondDB), we used Run 203777. Since the 2013 sourcing was performed with the old PMTs, the Calibration Coefficients CC_c^{RunI} (GeV/ADC) used during Run I were used to convert the signal measured in ADC/25ns into GeV/25ns. It is important to point out that in calculation of energy deposition, response corrections are not included. That is because the Run I calibration for PMT gains should be preserved, when extracting the energy, but not the HF response, which is affected by other corrections accumulated in response corrections.

In Figure 3.5, correlation plots between the inverted Calibration Coefficients (ADC/GeV) from Run I and Source Signal (ADC/25ns) from 2013 Campaign are presented. Here, a similar strategy to Run I Calibration was used, when EM and HAD channels were calibrated separately. Except for some outlier channels, we observe a good correlation between Signals collected (in 2013) and the Run I Calibration Constants. Therefore, we can extract Source Energy Deposition:

$$\langle E \rangle_c = \langle Q \rangle_c^{Geom} \times CC_c^{RunI}. \tag{3.6}$$

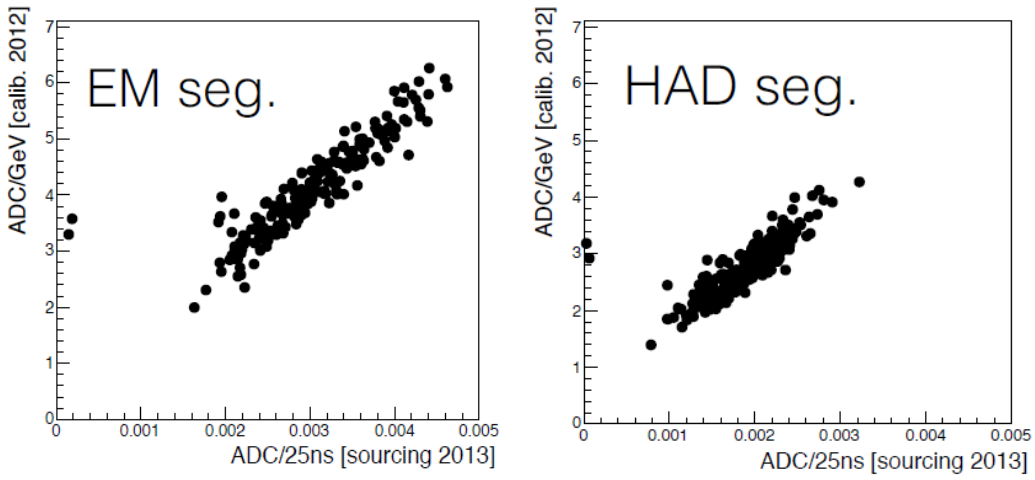


Figure 3.5: (a) Correlation plot between QIE response, in ADC/GeV, and 2013 HF-geometry corrected energy deposition, in ADC/25ns, per tower for the EM channel. (b) Similar correlation is observed for the H channel.

3.5.2 Calculating Calibration Coefficients

During Run II, HF will be operated at the Operational Voltage 2 (OV2). Since both 2014 sourcing campaigns were done at voltage settings, OV1 and OV1+100, different from the one to be used, the actual source signals should be converted to the desired voltage by using gains for the respective voltages. Another factor to account for is the firmware setting, either 1 TS or 2 TS. Taking these two arguments into

account, compute a source signal at OV2 in units of ADC/25ns:

$$\langle Q \rangle_c^{Geom,OV2} = \frac{\langle Q \rangle_c^{Geom}}{nTS} \times \frac{GAIN^{OV2}}{GAIN^{OV1,OV1+100}}. \tag{3.7}$$

Finally, compute the actual Calibration Coefficients (HF Gains) CC_c^{RunII}:

$$CC_c^{RunII} = \tau \times \frac{\langle E \rangle^{2013}}{\langle Q \rangle_c^{Geom,OV2}}. \tag{3.8}$$

where τ is the source radioactivity correction factor, which accounts for exponential source activity decrease. We take 2013 sourcing date as the starting point and compute the decrease for April and July 2014 with respect to that date.

3.6 Results and Discussion

3.6.1 2013 Results

As has already been pointed out, outlier channels from Figure 3.5 have been excluded. Moreover, to achieve a result for the average energy deposition per channel type, EM or H, from the radioactive source, we only considered data from towers with IEta < 35 to minimize the effects of radiation damage of HF fibers at higher η towers, shown in Figure 3.6.

Figure 3.7 shows the results from 2013 when only the near side of HF- was sourced, nine wedges, and only the towers below IEta $= 35$ are considered, eight towers per wedge. The average energy deposition extracted from 2013 sourcing campaign, for EM and H channels separately, is 744.6 $\pm$ 6.3 keV and 706.8 $\pm$ 7.7 keV per time slice,

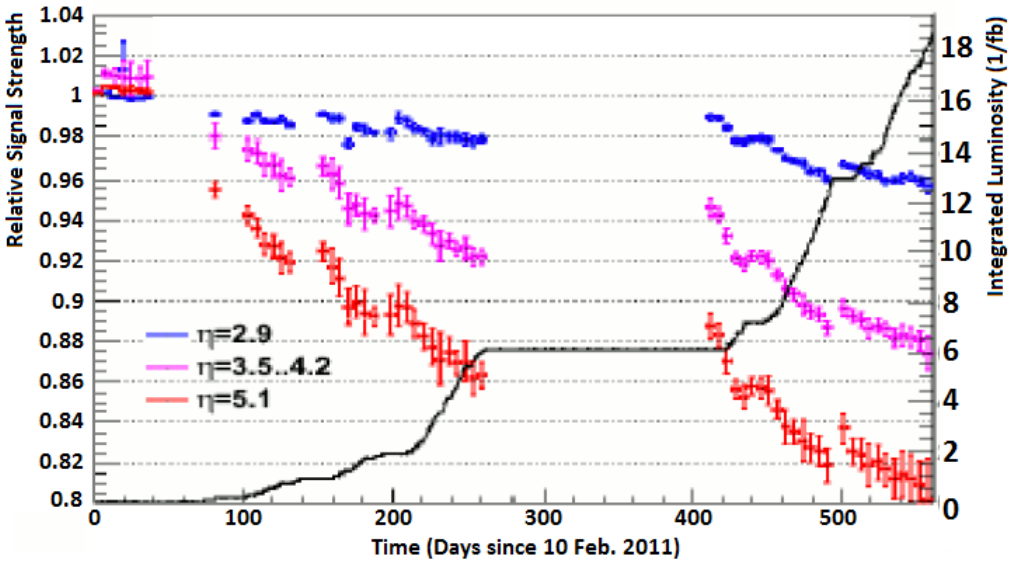


Figure 3.6: R7525 PMT relative signal strength with respect to 10 Feb. 2011 as a function of time since that date, for various η locations. The solid black curve represents the integrated luminosity within CMS over the same time period.

respectively. There is a 5% difference observed between the EM and HAD channels, while their respective precision is kept within 1 % each.

3.6.2 2014 Results

The source signals, $\langle Q \rangle_c^{Geom}$, for HF+ and HF- for 2014 data with new PMTs, corrected for geometry (geometry containment factor), firmware used (1 TS or 2 TS), Operating Voltage to be used (converting to OV2 from either OV1 or OV1+100), are calculated using Equation 3.7 and are presented in Figure 3.8.

The HF Gains, CC_c^{RunII} in units of GeV/ADC, for HF+ and HF- are computed using Equation 3.8 and are presented in Figure 3.9. In Equation 3.8, if c is a EM

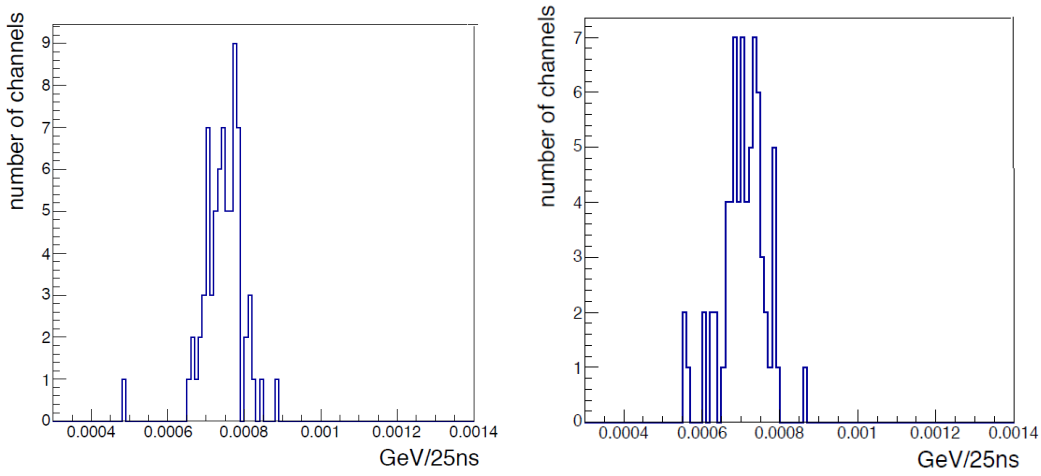


Figure 3.7: (a) EM Energy deposition for each tower below IEta = 35. (b) H Energy deposition for the same towers.

channel from a given tower, then use the EM result from 2013, and similar logic applies for the H channel counterparts.

To account for the differences in Operating Voltages for Sourcing vs Run II Physics campaign, the PMT Gain Ratios were applied as Conversion Factors, distributions of which can be found in Figure 3.10.

3.7 Systematics Evaluation

3.7.1 1 TS vs. 2 TS

The very first systematic study that was performed is to compare the results between two firmware configurations. All of HF+ towers and half of HF- were sourced twice, using either 1 TS (with Operating Voltage 1) or 2 TS (using Operating Voltage 2), which provides us a measure of consistency in computing the calibration coeffi-

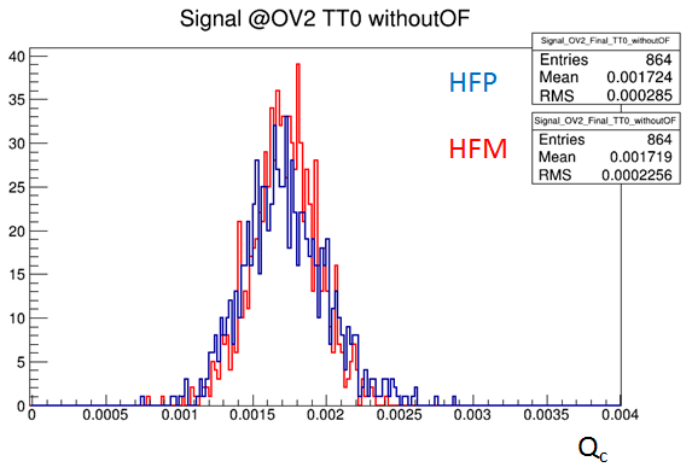


Figure 3.8: Actual Signal from the Source recorded by the PMT at OV2 (Operational Voltage 2).

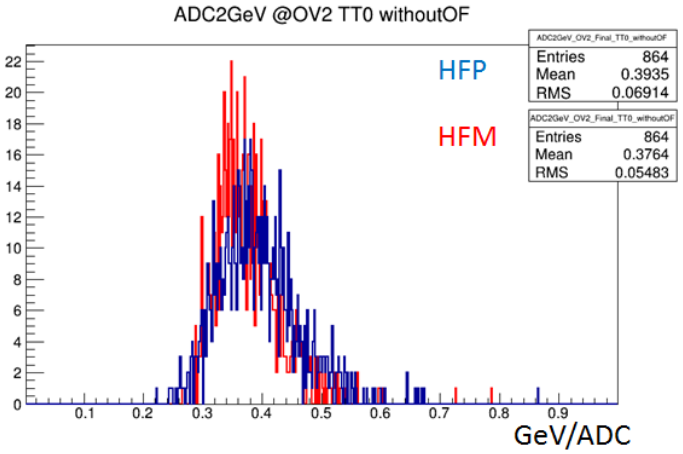


Figure 3.9: Distribution of Calibration Coefficients (HF Gains).

cient across different firmware versions and operating voltages. To compare these 2 modes of operation, we used CC_c^{RunII} computed for each sourcing configuration, after

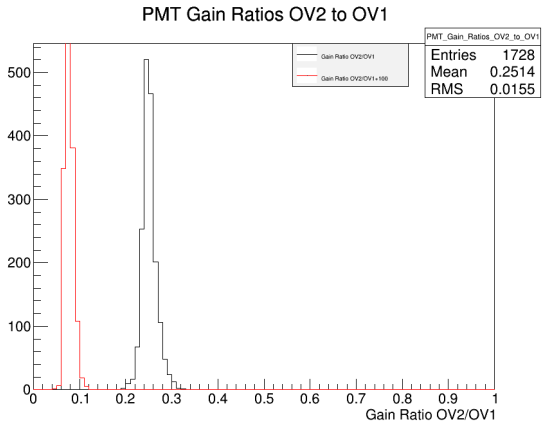


Figure 3.10: Black - Ratio of PMT Gains for OV2/OV1. Red - Ratio of PMT Gains for OV2/OV1+100.

applying all the required corrections. The distributions of ratios (1 TS over 2 TS) is presented in the Figures 3.11. Comparing 1 TS OV1 and 2 TS OV1+100 results, the source signals computed need to be corrected for the 25 ns vs 50 ns integration window. After correcting between the firmware gains, the results agree to an order of 1%.

As observed in the Figures 3.11, 1 TS and 2 TS results match both for HF+ (1.4%) and HF- (0.3%) at the order of under 1.5%, which establishes solid independence of calibration coefficients from the firmware used.

3.7.2 Transversal Uniformity: Tubes A vs. Tubes B

Approximately a quarter of HF wedges contains a second sourcing tube, which differ in the groove type and as a consequence in the location within a wedge. By comparing the obtained calibration coefficients using sourcing data from both tubes,

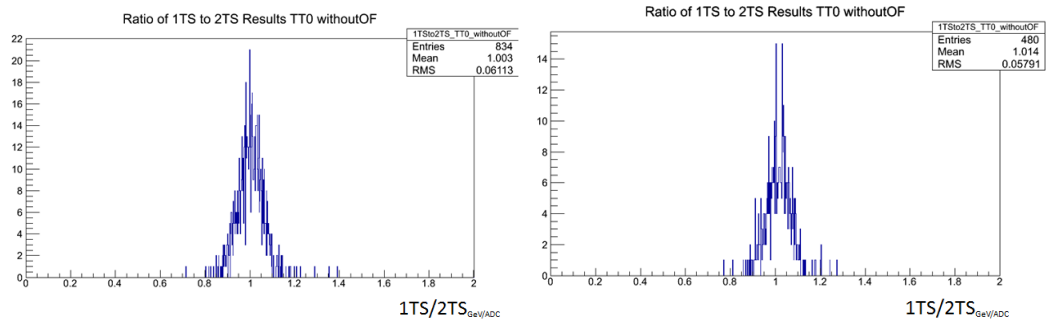


Figure 3.11: (a) Ratio of 1 TS/2 TS results for HF-. (b) Ratio of 1 TS/2 TS results for HF+. For both sides the compared quantity was CC_c^{RunII}.

it is possible to extract information on the transversal uniformity of the signal within a wedge. Again, as in the case of the 1 TS vs 2 TS study, the actual CC_c^{RunII} are to be compared. The ratios are presented in Figure 3.12. Transversal uniformity between A and B tubes within towers containing them show good agreement between results as well, differing by under 1%.

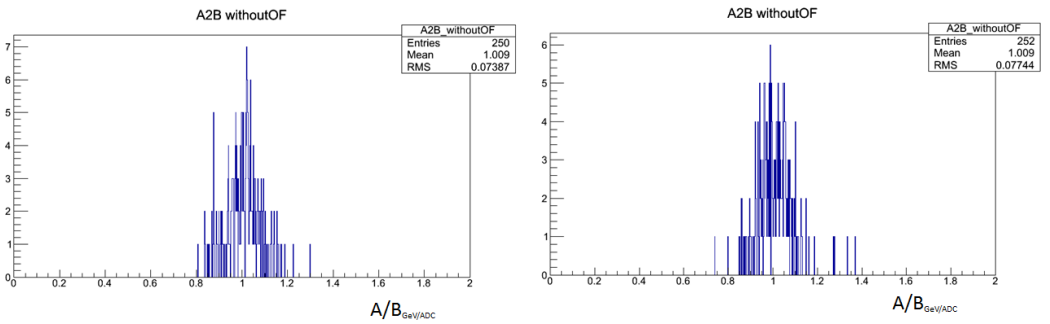


Figure 3.12: (a) Ratio of A/B results for HF-. (b) Ratio of A/B results for HF+. For both sides the compared quantity was CC_c^{RunII}.

3.7.3 Longitudinal Uniformity

As the radioactive source is moving along the source tube, it is possible to actually record the spatial information on the location of the source, which provides with another tool to estimate the uncertainty of our measurements. The general idea is that several regions along the source tube are defined and histograms for each region respectively are summed up and then charged are extracted and compared.

As it was described in Section 3.3, the tubes' start (tubeStart) and end (tubeEnd) positions are provided. In Table 3.2 definitions of the tubes' regions of interest are formally defined. "Front" and "Back" are defined so that "Front" is closer to IP and "Back" is further away. "Signal" is the region that has been used as the defining region for extracting the charge to be used in calibration coefficient calculation. And the "$\frac{2}{3}$Back" is an additional region defined to compare with the "Signal". It should be noted that "$\frac{2}{3}$Back" and "Signal" have overlapping and non-overlapping regions. Even though the overlapping region is the dominant one, using these two regions it is possible to estimate how much the choice of the region influences the actual charge computed.

From Figure 3.13, observe that the signal in "Front" region is about 94 - 95% of the signal in the "Back" region. But it is necessary to be more careful here when attributing this to systematics, as this mean ratio is consistent with the light attenuation in non-damaged fibers. Therefore, this 5 - 6% cannot be fully attributed to the systematic uncertainty. Considering the signal in "$\frac{2}{3}$Back" with respect to the "Signal" region, a difference of 2% on average is observed, which is the the contribution

Table 3.2: Source tube regions defined to provide a measure of uncertainty on the charge deposited in various regions along the source tube.

Region Name	Start	End
Front (Depth 1 or EM)	Tube End - 400	Tube End-100
Front (Depth 2 or H)	Tube End - 700	Tube End-400
Back	Tube Start+100	Tube Start + 400
$\frac{2}{3}$ Back	Tube Start	Tube Start + $\frac{2}{3}$ (Tube End - Tube Start)
Signal (EM and H)	Tube Start + 300	Tube End-300

to the systematic uncertainty due to the longitudinal non-uniformity of the signal.

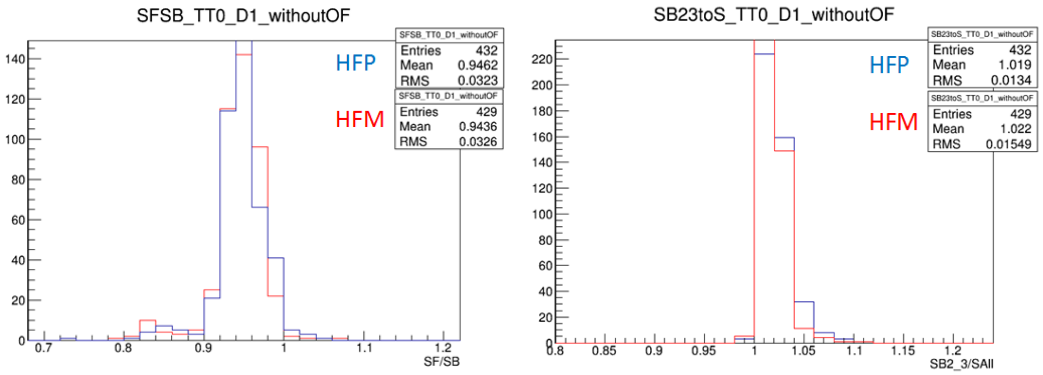


Figure 3.13: (a) Ratio of the charge extracted from the "Front" region to the charge computed in the region "Back". (b) Ratio of the charge computed within "$\frac{2}{3}$Back" region to the "Signal" region.

3.7.4 Cross-Check

An additional systematic uncertainty is included for the methodology described in this note for measuring the absorber response to the source energy. Independent analyses were performed on the same data, and the results were compared and agree within an order of 1%, as can be deduced from Figure 3.14.

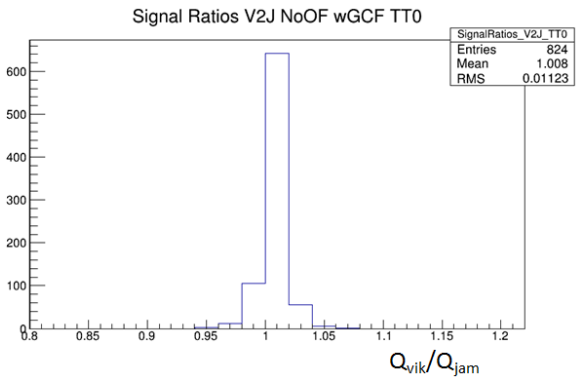


Figure 3.14: Cross-checking the results for HFM. On average the results agree within 1%.

3.8 Conclusions

In this chapter the results of the Hadron Forward calorimeter sourcing calibration procedure after new PMTs were installed on HF were presented. Sourcing data was collected prior to installing the new PMTs in HF- in order to obtain the average source energy deposition and extrapolate over both HF calorimeters. Calibration coefficients (HF Gains) were computed by averaging the background-subtracted signal

yield over the majority of the absorber depth (beyond 30 cm of each end of the source tube) for a given tower of each wedge. Systematic uncertainties have been computed. The total uncertainty is the sum in quadratures and established to be under 10%, which agrees with the expected precision for our calibration procedure.

CHAPTER 4
SIMULATIONS OF MODERN CALORIMETER SYSTEMS

4.1 Introduction

4.1.1 Simulation Objective

Any simulation is an approximation of a certain process. It is a model that one builds to try to understand something under investigation. It can come in various kinds with different names: virtual reality, avia simulator, computer game or be it any physics simulation. In the end, they are all trying to help model a process of interest. Therefore, the primary objective of any simulation is to model a certain process as a close approximation of reality.

Physics Simulations are not the exceptions - it is how Experimental High Energy Physics Experiments function. One starts out by having an objective to measure some quantity or to verify a hypothesis. Typically that means that a model already exists to provide with signatures and features to look for it. Then, a simulation of a system is built to measure physical quantities of interest and then to compare experimental results with simulated predictions. This final step is the most important part - if the simulation (model) has not been possible, one would not have something to compare results with and therefore no way to draw conclusions.

In what follows, two standalone calorimeter systems are examined: the High Granularity Calorimeter (HGC) [45] and the Shashlik + Hadron Endcap system. Both detectors were considered as potential candidates for the CMS Phase 2 Up-

grade. What matters most for the simulation is that these two systems present different choice of technology, in particular for the electromagnetic calorimeter part. Intrinsically, the Shashlik calorimeter has better resolution due to the scintillation process being very efficient, however the disadvantage is the introduction of the light to electric signal conversion part, which very often means that longitudinal segmentation is not possible. In turn, the electromagnetic part of the HGC, although it has a less efficient silicon active material, allows to put the readout electronics directly on the calorimeter, which results in the longitudinal segmentation and profiling of the showers.

4.1.2 Simulation Tools

For the purpose of simulating High Energy Physics (HEP) calorimeter systems physicists employ GEANT4 [46, 47], a software toolkit which provides an interface for building actual simulation, carrying it out and collecting the results. To be more precise, every simulation has the following features:

- **Materials**. All the materials that are to be used in the simulation must be defined. All the properties of all of the chemical elements must be also defined. It is important to point out that if, for instance, one has a material that acts as a scintillator in a calorimeter, optical characteristics of such a material must be provided separately and can be optimized.
- **Geometry**. Using various geometrical primitives (cube, cone, etc...) a geometrical specification of the simulation has to be provided. Technically speaking,

Materials + **Geometry** define the physical layout of a calorimeter.

- **Physics**. One of the most important parts of a simulation is the ability to change the physics processes that are being used and see how that affects calorimeter's response. For instance, if there is a scintillation material that is expected to output light upon incoming radiation, then by turning off Scintillation Physics Process, the output of the calorimeter will be suppressed. In GEANT4, this is typically achieved via **Physics Lists** - that contain the specification of the most important physics processes.
- **Readout**. The whole idea of performing a simulation is to yield some output that you can then analyze further. For that, there are **Sensitive Detectors**, which get attached to **Geometry Volumes** and get triggered for every single step of any particle within that volume. What gets stored is up to the user to specify.
- **Primary Generators**. In order to start a simulation, one needs a trigger - a way to artificially inject some physics objects to interact with the rest of the virtual realm. Within the GEANT4 context, this trigger is called **Primary Generator**. It can come in various forms: from simple particle guns to generating complicated decay processes.
- **Simulation Engine**. Finally, GEANT4 provides an engine to carry out the simulation itself: track all the particles for each step, trigger various transitions for readout, apply the physics processes, etc...

4.2 High Granularity Calorimeter

The High Granularity Calorimeter [45] (HGC) model that is built and analyzed in this chapter, although implementing the proper geometry, borrows all the basics from the CALICE Silicon-Tungsten [48] (SiW) system. Moreover, both HGC and CALICE SiW terms are used interchangeably, because although HGC has both electromagnetic and hadronic parts, for the sake of discussion the hadronic part is omitted. Figure 4.1 shows a custom GEANT4 built full CMS scale High Granularity Calorimeter as the CMS Endcap system. The HGC has been one of the potential candidates for the CMS Phase 2 upgrade.

4.2.1 Physical Layout

HGC consists of three parts: **Electromagnetic** (EM) calorimeter, **Front Hadron** (FH) and **Backing Hadron** (BH) Calorimeters. Both EM and FH have similar design principles: alternating layers of absorber and active materials mixed in together with a layer of electronics readout. Full geometry specification for the EM part can be summarized as follows:

- 25 Radiation Length (X_0) Device
- In total 30 layers, where each layer is (absorber (W), active (Si), readout (G10/PCB)).
- First 10 layers have 0.5 X_0 per layer
- Second 10 layers have 0.8 X_0 per layer
- Third 10 layers have 1.2 X_0 per layer

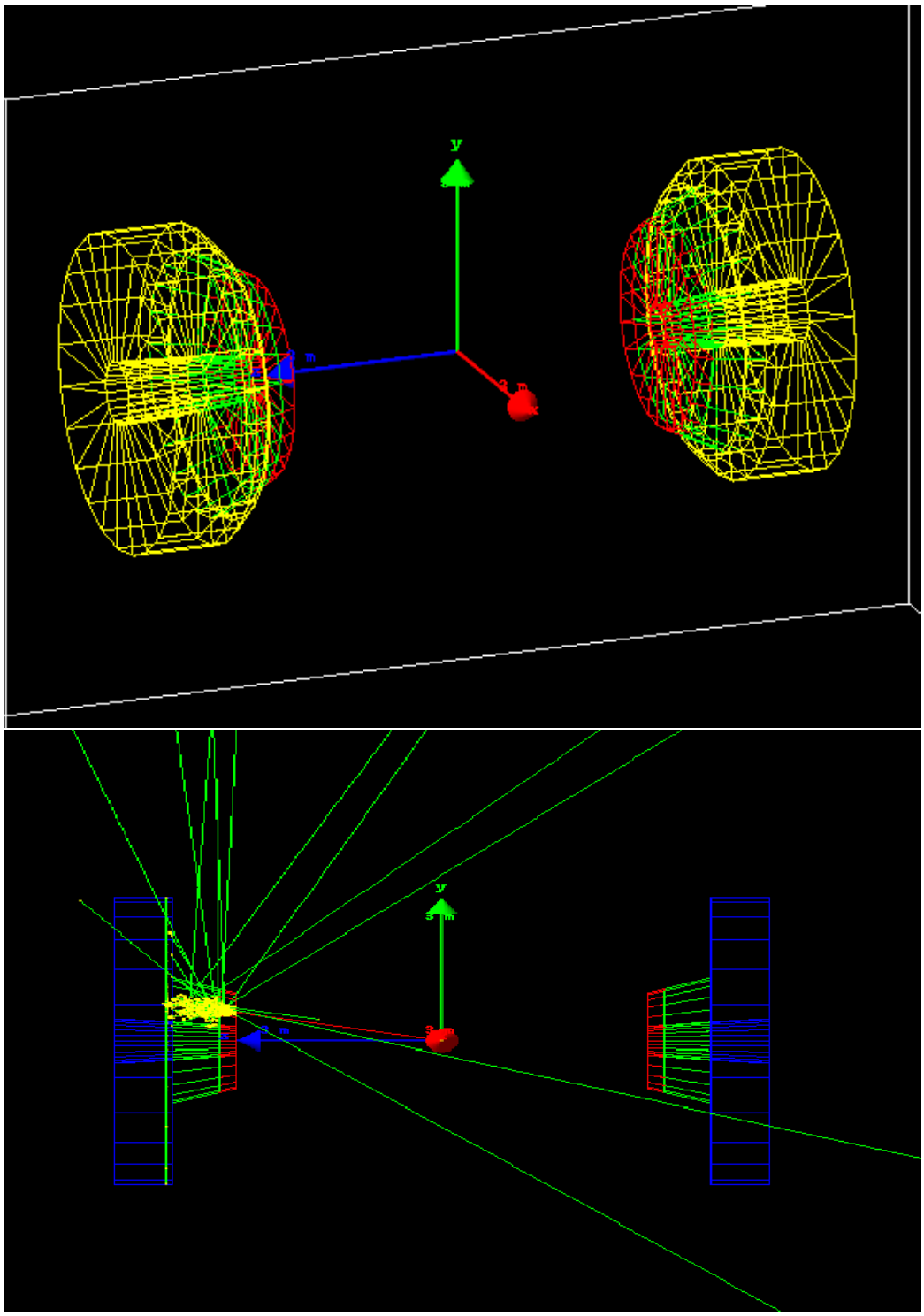


Figure 4.1: Full CMS Scale High Granularity Calorimeter (top). Example of a Particle Gun response (bottom).

- XY-plane is subdivided into pads, with an area of each channel of 0.9 cm^2 for the first 20 layers and 1.8 cm^2 for the last 10.

Full specification for the FH part can be summarized as follows:

- 4 Interaction length (λ) Device
- 12 layers of (absorber (W/Brass), active (Si), readout (G10/PCB)), each layer is 0.33 λ
- XY-plane is also subdivided into small pads, with each covering an area of 1.8 cm^2

4.2.2 Detector Readout

The operating principle of SiW calorimeter is based on electron-hole pairs generated by a charged track traversing the active material. For the purpose of simulation, assume a constant number of such holes per 1 µm step (80 holes per 1 µm), which is a valid assumption that has been previously used by the CALICE SiW Electromagnetic Calorimeter [48]. Equation 4.1 summarizes the computation of the response for a single pad per event.

$$R(cell) = \sum_{n=1}^{steps} \frac{80 \times \Delta x}{1 \mu m} \tag{4.1}$$

4.2.3 Simulation

The goal of the modeling is to establish the performance characteristics of the calorimeter system: response linearity and energy resolution. For that, an electron

Particle Gun with varying energies is used. Each event corresponds to shooting a single e^- of certain energy, performing all the tracking for all of the shower particles and recording all of the results as discussed in Section 4.2.2. It is important to point out that each primary particle enters the detector at an angle of normal incidence for the purpose of precise control of the simulation environment. Figure 4.2 provides an example of a shower distribution for a single event.

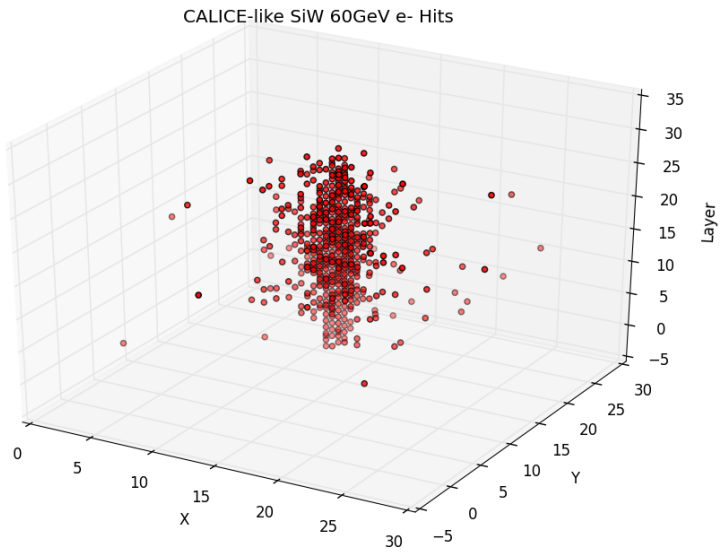


Figure 4.2: An example of a shower distribution within HGC for an incident 60 GeV e^-.

4.2.4 Analysis

The objective of the readout analysis is to establish calorimeter performance characteristics: linearity of response and energy resolution. Not all systems are linear and intrinsically SiW calorimeters are not linear in response. The reason for desirability of this characteristic is the ability to easily calibrate the device. In other words, given a linear system, the conversion of the raw response to energy is trivially established. In summary, the following steps are to be followed to establish the performance characteristics:

- Compute the total response. Use Equation 4.2a
- Compute the energy due to the calorimeter response. Use Equation 4.2a.
- Overlay particle gun energy with total response as in Figure 4.3. Perform a linear fit and extract the calibration coefficient.
- Using the derived calibration coefficient for each generated event, compute the energy due to the response. Figure 4.4 shows the reconstructed energy distributions.
- Using Equation 4.2b, compute and overlay resolution with the incident energy. Perform the fit of resolution vs energy using Equation 4.2c.

$$E = CC \times RawResponse. = CC \times \sum_{i=1}^{n_{layers}} w_i \times R_i \tag{4.2a}$$

$$Resolution = \frac{\sigma_E}{\mu_E} \tag{4.2b}$$

$$f(x) = \sqrt{\frac{\alpha^2}{x} + C^2} \tag{4.2c}$$

where in Equation 4.2c, α is the term that represents the level of intrinsic statistical

fluctuations of the calorimeter and C, constant or asymptotic term, sets the limits of the calorimeter at higher energies.

4.2.5 Results

Following the procedure described just above, first obtain the results for the linearity of HGC in Figure 4.3. The yield per 1 GeV, about 4.4×10^6, is the extracted Calibration Coefficient (CC). As can be observed - the electromagnetic response is highly linear!

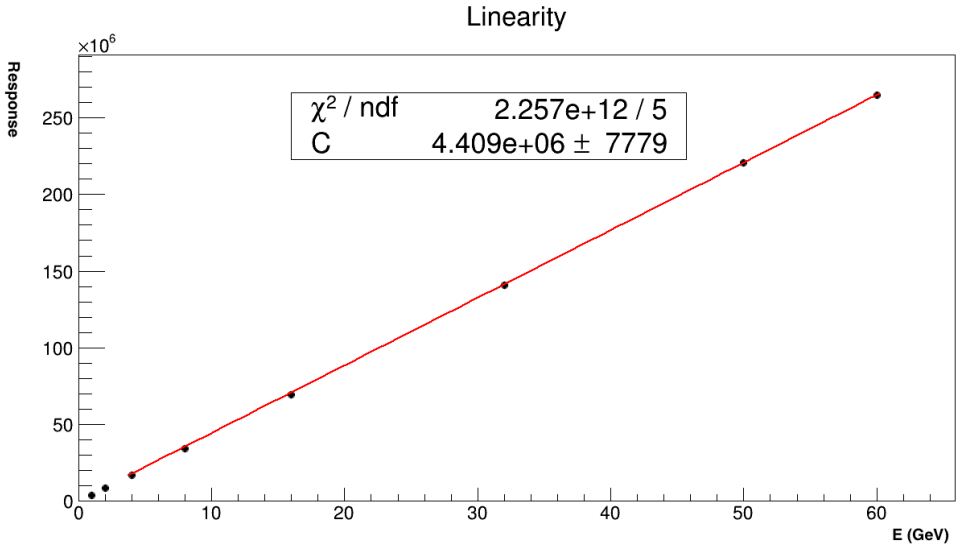


Figure 4.3: The linearity graph of the Electromagnetic component of the HGC. Shows the dependence of the total calorimeter response as a function of the energy of the incident e^-.

Furthermore, applying the obtained the Calibration Coefficient for the purpose

of the energy reconstruction according to the Equation 4.2 results in the energy distributions shown in Figure 4.4. Electron particle gun with eight different energies have been used. A Gaussian fit is performed in order to extract the mean and the width of the reconstructed energy distributions.

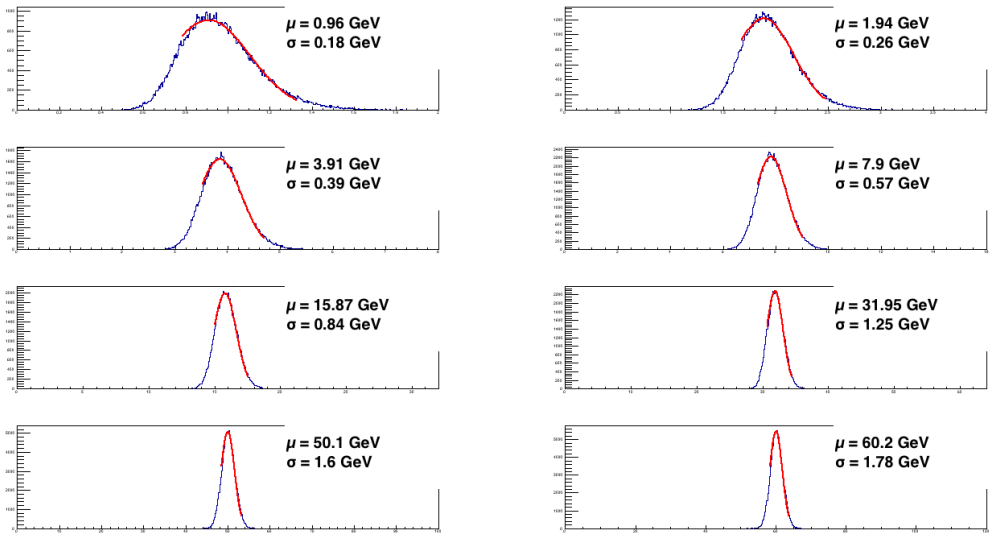


Figure 4.4: The HGC Reconstructed Energy Distributions. Electron particle gun used with eight different energies: 1 GeV, 2 GeV, 4 GeV, 8 GeV, 16 GeV, 32 GeV, 50 GeV, 60 GeV.

And, finally, the energy resolution results are shown in Figure 4.5. The asymptotic energy resolution of the HGC electromagnetic component is at 1%.

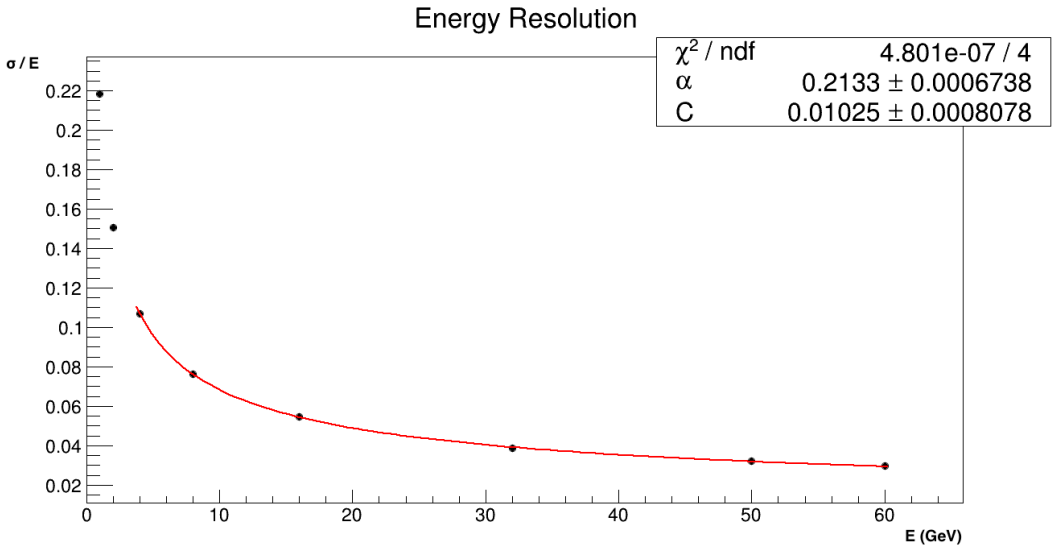


Figure 4.5: The HGC Energy Resolution curve. Stochastic component α (0.21) determines the level of statistical fluctuations and C (1%) shows the behavior of the system at high energies.

4.3 Shashlik + Hadron Endcap

The second candidate for the CMS Phase 2 upgrade was the Shashlik + Hadron Endcap. This system actually features two separate calorimeters, one for the electromagnetic component and one for the hadronic. Both rely on the scintillation process in order to measure the energy of the showers. Figure 4.6 shows a custom built full CMS scale "Shashlik + Hadron Endcap" detector.

4.3.1 Physical Layout

Shashlik is an Electromagnetic Calorimeter System with the name implicitly reflecting its structure and the choice of readout technology. The physical layout is

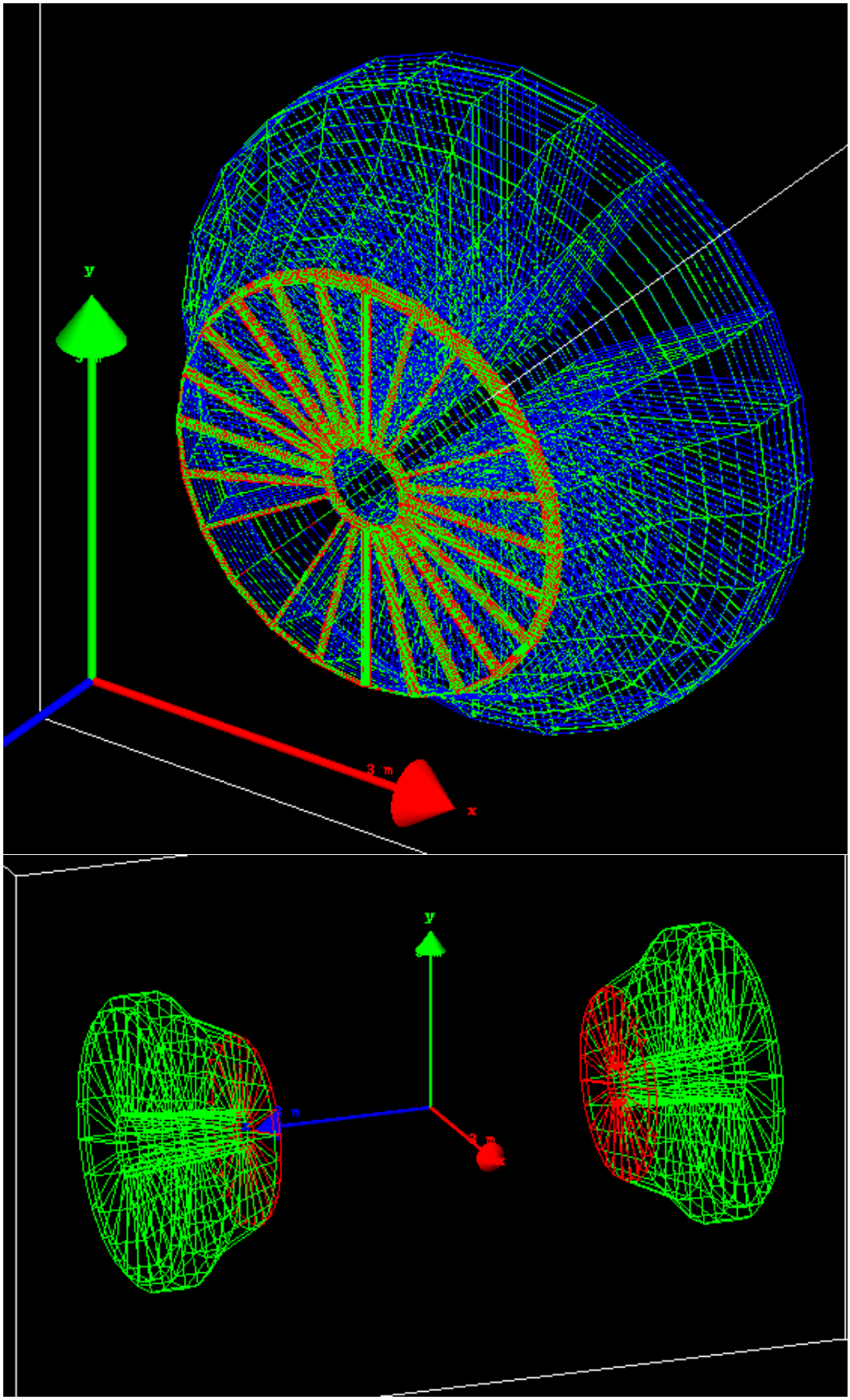


Figure 4.6: Full CMS Scale "Shashlik + HE" System (from different angles).

similar to HGC discussed in Section 4.2: alternating layers of absorber and active materials, however there is no electronics unit sitting on the calorimeter itself. Instead, wave-length shifting fibers go through the entire length of the calorimeter and are responsible for light capture and transmission to the electronics unit for conversion into an analog signal. The full specification of the Shashlik Calorimeter can be summarized as follows:

- 25-30 X_0 device.
- No Longitudinal Segmentation.
- Alternating layers of absorber (tungsten W) and active (LYSO) materials.
- Active Material used is LYSO crystal scintillator.

Hadron Endcap (HE) is an active CMS Endcap Calorimeter that is part of the HCAL subsystem. Its design principles are similar with respect to Shashlik, however hadron calorimeter has a slightly different choice of readout technology [49]. The Calorimeter specification can be summarized as follows:

- 10λ device.
- Alternating layers of absorber (Brass) and active (SCSN-81 plastic scintillator) materials [49].
- Partial longitudinal segmentation.

4.3.2 Detector Readout

For the purpose of simplification, in the simulation the assumption is made to have a perfect light capture and transmission for both Shashlik and Endcap. Also no

modeling of the wave-length shifting fibers or any kind of photodetectors is done. In other words, this study looks for upper limits of our calorimeter performance. The metric of the system's response is defined to be the number of generated photons within the scintillator. The scintillation mechanism is responsible for light generation and G4Scintillation [50]; the physics process of GEANT4 provides these capabilities via optical photons. For the purpose of modeling, GEANT4 defines two different types of photons: regular photons, that obey the laws of quantum physics, and optical photons, that follow the laws of geometrical optics. Optical photons do not participate in the conservation of energy laws and do not deposit any energy into the scintillator (this fact allows for optimizations).

4.3.3 Parametrized Detector Readout

Typical light yields for a scintillator are in thousands of optical photons per MeV of deposited energy into the scintillation material. The exact number is material-dependent and varies substantially. Given that the input particle has energies in the GeV range, the number of optical photons that gets is goes well above 1 million. Tracing all of these photons is a complicated and time-consuming task for GEANT4's engine. Moreover, given that optical photons can not deposit energy into the material, one can simply count and kill them right after they have been generated. Therefore, for the purpose of optimizing the time it takes to generate a single event, the parametrization of the response of a scintillator is performed. This procedure allows to substantially speed up the simulation without degrading the performance.

4.3.4 Analysis and Results

The analysis procedure is identical to the one described for HGC in Section 4.2 with two main differences. First, the readout metric here is defined to be simply the total number of generated optical photons. No layer by layer weighting is applied for neither Shashlik nor HE. And second, the HE is calibrated separately, by shooting pions with a different set of energies. Figure 4.7 shows the results of computing the linearity for both Shashlik and HE. It is clear, especially comparing to HGC, that Shashlik exhibits better linearity properties. The yield per 1 GeV is about 5.3×10^6 for Shashlik and 7.0×10^4 for Hadron Endcap.

Figures 4.8, 4.9 show the reconstructed energy distributions for Shashlik and HE, respectively. Shashlik and HE are calibrated separately as they constitute isolated parts of the system. Note this is different with respect to HGC, where calibration of EM and FH parts was done together. The energy reconstruction is obtained by applying the Calibration Coefficient (CC) that converts the number of optical photons generated to the actual units of energy.

Results for the energy resolution are provided in Figure 4.10. The differences between the parametrized response and the use of optical photons are negligible.

4.4 Conclusions

Two standalone simulations of the potential candidates for the CMS Phase 2 upgrade have been built and examined in this chapter: the High Granularity Calorimeter and the "Shashlik+HE" system. The GEANT4 toolkit has been utilized

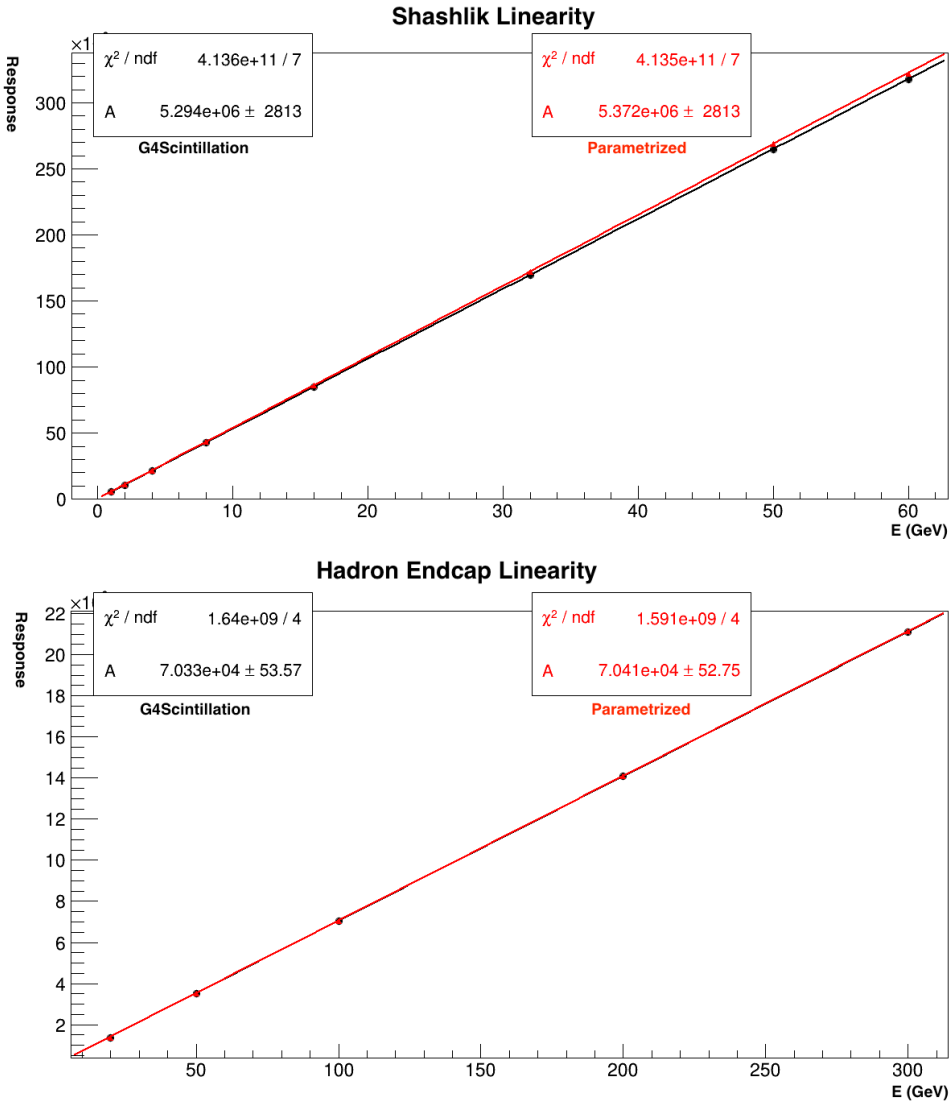


Figure 4.7: The linearity graph of the Shashlik electromagnetic calorimeter (Top) and Hadron Endcap (Bottom). G4Scintillation (Black) and Parametrized (Red). Shows the dependence of the response of the calorimeter as a function of the energy of the incident particle: electron (Shashlik) and pion (Hadron Endcap).

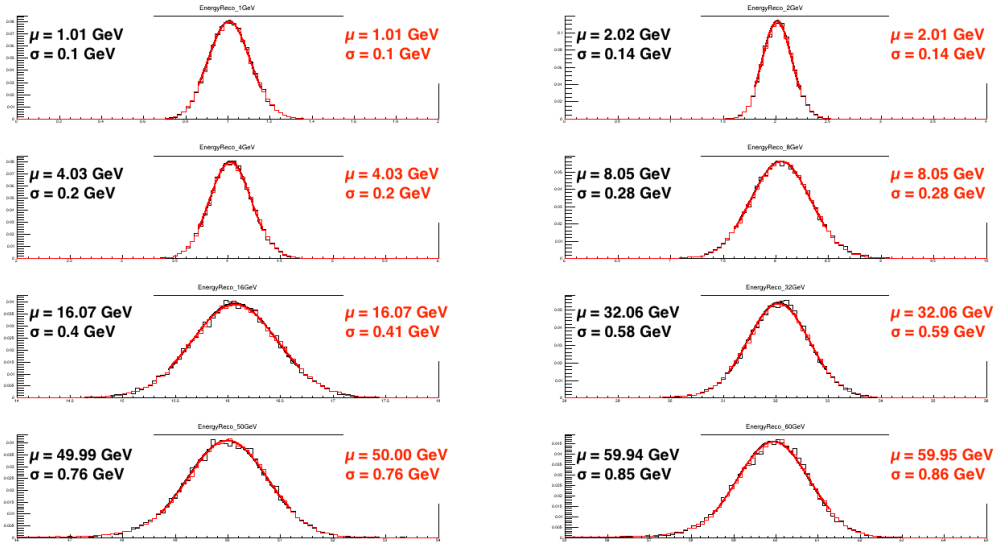


Figure 4.8: Reconstructed Energy Distributions for Shashlik. G4Scintillation (Black) and Parametrized (Red). Electron particle gun used with eight different energies: 1 GeV, 2 GeV, 4 GeV, 8 GeV, 16 GeV, 32 GeV, 50 GeV and 60 GeV.

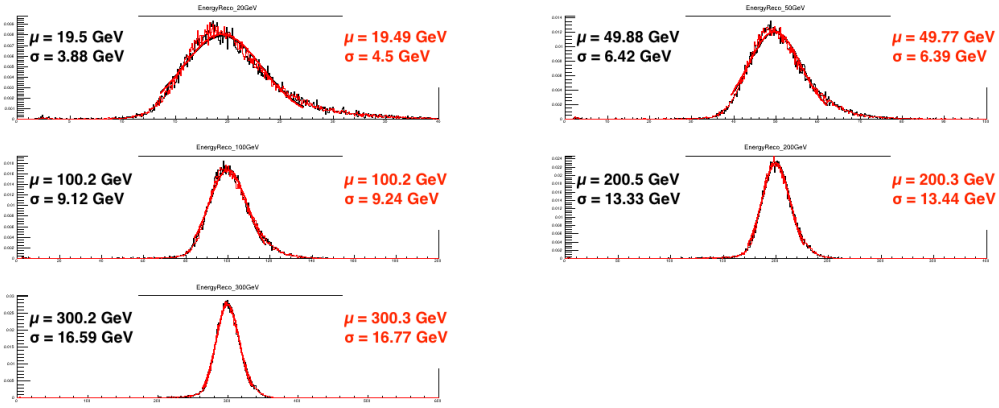


Figure 4.9: Reconstructed Energy Distributions for Hadron Endcap. G4Scintillation (Black) and Parametrized (Red). Pion particle gun used with five different energies: 20 GeV, 50 GeV, 100 GeV, 200 GeV, 300 GeV.

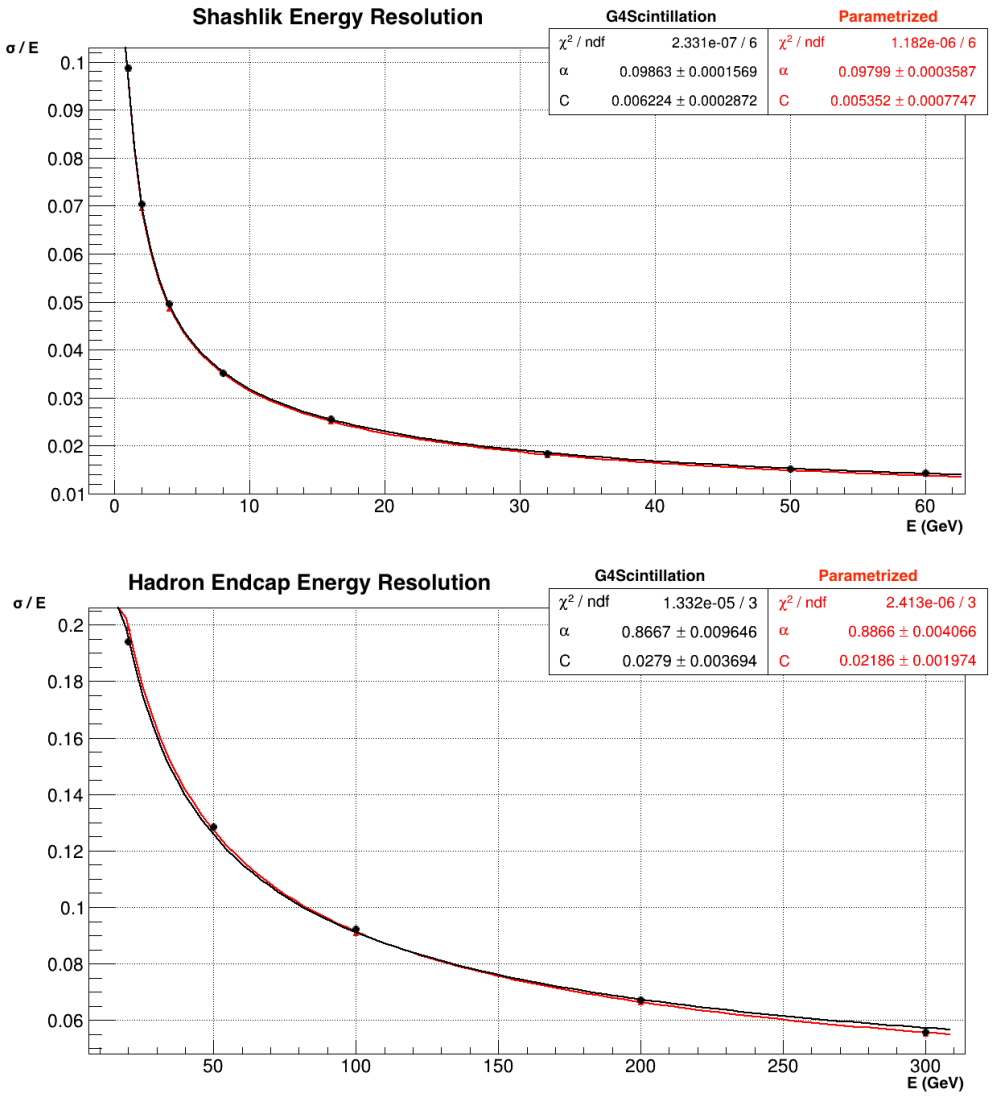


Figure 4.10: Energy Resolution for Shashlik (Top) and for Hadron Endcap (Bottom). G4Scintillation (Black) and Parametrized (Red). For the Shashlik system, the stochastic component α (≈ 0.01) determines the level of statistical fluctuations and C (0.6%) shows the behavior of the system at high energies. Similarly, for the Hadron Endcap, the stochastic component α is ≈ 0.87 and C is $\approx 0.03\%$.

for the purpose of geometry, physics and readout simulation. The baseline performance of both systems has been established. It was observed that Shashlik part has better energy resolution characteristics than the electromagnetic component of the HGC, with asymptotic resolution of 0.6% vs 1%. The statistical fluctuations also show a factor of 2 improvement for the Shashlik over the HGC.

Standalone Simulations of Calorimeters are ideal playgrounds for optimizing the parameters of future systems. It avoids unnecessary overhead, which is always introduced once you try to scale things up and use very precise simulations as are used for the CMS detector. The time to physics is substantially reduced and allows to accelerate the development of reconstruction techniques. A very interesting future use case is the Machine Learning (ML) Benchmarking. In comparison, MNIST Digit Classification, [51], has become the de facto benchmarking for the field of Computer Vision and Artificial Intelligence. Energy Regression of a standalone calorimeter system could be a similar standard for the field of Calorimetry.

CHAPTER 5
FINAL REMARKS

In order to conclude the topics examined in this work, I would like to go back to the introduction, restate and explain the fact how the topics presented cover the full phase space of the CMS experiment research program. Generally speaking, this field can be divided into three separate subareas: detector construction (development), operations (DevOps [52]) and physics analytics. Although, the separation is very clearly defined, the more one dives into a particular area, the more it becomes obvious that they are interconnected and complement each other. It is also interesting to note that recently the HEP experiments more and more follow approaches similar to the software development cycles [53].

To understand the connection, consider walking through the process of construction of some imaginary experiment. The very first thing required is to build the detectors themselves. Typically, this process goes in parallel with the construction of simulation models of such systems. That is a very important point, because one needs not to just construct a system, but to validate it against the simulated model and to be able to understand the performance characteristics upon which to judge the differences of proposed systems. That is where simulations of calorimeters come into play. Chapter 4 discusses details of building the geometry of detectors, performing the actual simulation and analyzing the response of the calorimeters to deduce basic performance characteristics. Although, simulations were discussed at the very end of this work, it is actually the very first step (phase) on a way to build a working

detector.

Next, imagine one has built a calorimeter or other device which has to measure some physical quantities (energy, position, time, etc.). The phrasing in the previous sentence is crucial, because the process of measuring a physical quantity always carries an implicit intermediate step - some sort of calibration procedure to map the response of a system to the actual physical units. In case of calorimeters, typically, the actual response is in digital form (or optical, which later gets converted to digital for easier manipulation). In reality, this gets even more complicated, because systems could be located in places without easy access (just like the CMS detector is located 100 m underground). Results and procedure of the calibration of the Hadron Forward calorimeter presented in the third chapter constitute a prime example for this topic. Moreover, calibration procedure (different means but the same purpose) is applicable not just to calorimeters, but to all types of particle detectors built at the CMS experiment. The reason calibration is part of the operations and not the construction phase of a system is because detector is physically constructed only once, but calibration will be performed multiple times. For instance, upon replacing a readout unit, one has to recalibrate this particular readout channel, or upon some upgrade, again, the energy scale has to be reestablished.

For the last phase, imagine a system has been built, simulated and calibrated. Is that the end of the story?! Of course not! Every system must carry a purpose for which it was constructed. I would like to provide an analogy. Humans built planes not just because they fly fast and look pretty, but because they allow for much

faster transportation, which, in turn, accelerates the growth, exchange of goods and simplification of trade among countries. The CMS experiment has been built and optimized targeting the discovery or exclusion of the Standard Model Higgs Boson. Therefore, ultimately, what matters most is the analysis of data collected by the CMS detector from proton-proton collisions, because only that will shed some light on the actual fundamental physics. Chapter 2 presents the results of the search for the Standard Model Higgs Boson in the dimuon final state with the CMS detector. This is the last phase of the research program in a sense that it achieves the purpose (although only upper limits were put on Higgs Boson production for this decay channel) for which the detector has been built.

REFERENCES

[1] Peskin M.E. and Schroder D.V. *An Introduction to Quantum Field Theory*. Addison-Wesley Publishing Company, 1996.

[2] D. Griffiths. *Introduction to Elementary Particles*. Wiley VCH, 2008.

[3] Merk M. Lectures on the Electroweak part of the Standard Model. https://www.nikhef.nl/ wouterh/teaching/PP1/LectureNotes.pdf.

[4] CMS Collaboration. The CMS Physics: Technical Design Report Volume 1: Detector Performance and Software. 2006. There is an error on cover due to a technical problem for some items.

[5] The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004, 2008.

[6] L. Evans and P. Bryant. The LHC Machine. *Journal of Instrumentation*, 3(08):S08001, 2008.

[7] The ATLAS Collaboration. The ATLAS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08003, 2008.

[8] The LHCb Collaboration. The LHCb experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08005, 2008.

[9] The ALICE Collaboration. The ALICE experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08002, 2008.

[10] *The CMS Tracker Project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1998.

[11] *The CMS Electromagnetic Calorimeter Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[12] *The CMS Hadron Calorimeter Project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[13] *The CMS Muon project : Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1998.

[14] *The CMS Technical Design Report for the Level-1 Trigger Upgrade.* Technical Design Report CMS. CERN, Geneva, 2013.

[15] *CMS The TriDAS Project : Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger.* Technical Design Report CMS. CERN, Geneva, 2002.

[16] S. Chatrchyan *et al.* Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30 – 61, 2012.

[17] G. Aad *et al.* Combined Measurement of the Higgs Boson Mass in *pp* Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments. *Phys. Rev. Lett.*, 114:191803, 2015.

[18] V. Khachatryan *et al.* Search for a standard model-like Higgs boson in the $\mu^+\mu$ and e^+e decay channels at the LHC. *Phys. Lett.*, B744:184–207, 2015.

[19] P. Nason. A New method for combining NLO QCD with shower Monte Carlo algorithms. *JHEP*, 11:040, 2004.

[20] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer. MadGraph 5 : Going Beyond. *JHEP*, 06:128, 2011.

[21] Johan Alwall. MadGraph5 amc at NLO web page. http://amcatnlo.web.cern.ch/amcatnlo/.

[22] T. Sjostrand, S. Mrenna, and P. Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[23] Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV. Technical Report CMS-PAS-PFT-10-002, 2010.

[24] F. Beaudette. The CMS Particle Flow Algorithm. In *Proceedings, International Conference on Calorimetry for the High Energy Frontier (CHEF 2013): Paris, France, April 22-25, 2013*, pages 295–304, 2013.

[25] A. Sirunyan *et al.* Particle-flow reconstruction and global event description with the CMS detector. 2017.

[26] S. Chatrchyan *et al.* Performance of CMS muon reconstruction in *pp* collision events at $\sqrt{s} = 7$ TeV. *JINST*, 7:P10002, 2012.

[27] The CMS Muon POG. The CMS Muon Physics Object Group. https://twiki.cern.ch/twiki/bin/viewauth/CMS/MuonPOG.

[28] The CMS Muon POG. The CMS Muon POG Recommendations. https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideMuonIdRun2.

[29] The CMS Collaboration. The CMS Rochester Corrections. https://twiki.cern.ch/twiki/bin/viewauth/CMS/RochcorMuon.

[30] The CMS Collaboration. The CMS Kalman corrections. https://twiki.cern.ch/twiki/bin/view/CMS/MuonScaleResolKalman.

[31] M. Cacciari, G. P. Salam, and G. Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008.

[32] S. Chatrchyan *et al.* Identification of b-quark jets with the CMS experiment. *JINST*, 8:P04013, 2013.

[33] R. Brun and F. Rademakers. ROOT - an Object Oriented Data Analysis Framework. *Nuclear Instruments and Methods in Physics Research*, 1996.

[34] D. de Florian *et al.* Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016.

[35] R Gavin, Y. Li, F. Petriello, and S. Quackenbush. FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order. *Comput. Phys. Commun.*, 182:2388–2403, 2011.

[36] The CMS Collaboration. The CMS RooMultiPdf. https://twiki.cern.ch/twiki/bin/view/CMS/HiggsWG/SWGuideNonStandardCombineUses.

[37] CMS Collaboration. Updated measurements of the Higgs boson at 125 GeV in the two photon decay channel. CMS Physics Analysis Summary CMS-PAS-JME-10-003, 2013.

[38] The CMS Higgs Combination Group. The CMS Higgs Combination Package. https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideHiggsAnalysisCombinedLimit.

[39] E. Gross O. Vitells G. Cowan, K. Cranmer. Asymptotic formulae for likelihood-based tests of new physics. 2013.

[40] CMS Collaboration. Technical proposal. *CERN/LHCC*, 94-39, 1994.

[41] CMS Collaboration. The Hadron Calorimeter Project Technical Design Report. *CERN/LHCC*, 97-31, 1997.

[42] N. Akchurin and R. Wigmans. Quartz fibers as active elements in detectors for particle physics. *Rev. Sci. Instrum.*, 74:2955, 2002.

[43] T. Shaw *et al.* Front End Electronics for the CMS Hadron Calorimeter. Fermilab Conference Note 2002/278, 2002.

[44] T. Grassi. HTR board Reference Guide. CMS Document, 2010.

[45] A. M. Magnan. HGCAL: a High-Granularity Calorimeter for the endcaps of CMS at HL-LHC. *JINST*, 12(01):C01042, 2017.

[46] S. Agostinelli *et al.* Geant4 - a simulation tookit. *Nuclear Instruments and Methods in Physics Research*, 506, 2003.

[47] J. Allison *et al.* Geant4 developments and applications. *IEEE Transactions on Nuclear Science*, 53, 2006.

[48] C. Adloff *et al.* Response of the CALICE Si-W Electromagnetic Calorimeter Physics Prototype to Electrons. *J. Phys. Conf. Ser.*, 160:012065, 2009.

[49] G. Baiatian *et al.* Design, performance, and calibration of CMS Hadron Endcap calorimeters. 2008.

[50] Geant4. Geant4 User Guide. http://geant4.web.cern.ch/geant4/UserDocumentation/UsersGuides/InstallationGuide/html/index.html.

[51] Y. LeCun. The MNIST Database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

[52] Wikipedia. Development and Operations. https://en.wikipedia.org/wiki/DevOps.

[53] Wikipedia. Systems Development Life Cycle. https://en.wikipedia.org/wiki/Systems_development_life_cycle.