Optimization of High-Energy Photon Identification at the LHC

Ryan S. Kim
Department of Physics
University of Notre Dame

Advisors

Professor Colin Jessop (University of Notre Dame)
Professor Riccardo Paramatti (University of Rome Sapienza)

In partial fulfillment of the requirements for the degree of
Bachelor of Science in Physics - Honors Track
May 2019

Acknowledgements

First and foremost, I would like to thank Professor Colin Jessop for all of his patience, encouragement, and guidance over the last three years in being my research advisor. I have learned so much about high-energy physics under his teaching and mentorship. I would like to also thank Professor Riccardo Paramatti for allowing me to work with him and for all of his help throughout my time in Rome, and the entire Rome CMS group for welcoming me warmly into their department. I am indebted to Dr. Livia Soffi, Dr. Francesco Michelli, and Kevin McDermott for supporting me in taking on this project and collaborating with them. I would also like to thank the entire CMS Collaboration for the opportunity for me to join this amazing international group of people. I would like to thank the rest of the ND CMS group, especially Professor Mitchell Wayne, Professor Kevin Lannon, Professor Michael Hildreth, and Daniel Ruggiero for helping me get started on the CMS experiment. I would like to thank the entire ND Physics Department for the kindness and support I have received throughout my time here. I am indebted to ND College of Science, Glynn Family Honors Program, ND Rome Global Gateway, and ND high-energy physics group for providing me with funding for research at various times of my undergraduate research career. Finally, I would like to thank my loving family for their continued support of me and my studies. I have grown so much as a person under their love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Standard Model of particle physics is the theory of the observed fundamental particles and their interactions. It has withstood experimental scrutiny over the years, including the crowning discovery of the elusive Higgs boson particle in 2012. Despite its successes, the Standard Model falls short of explaining some significant questions. First, it does not include gravity, one of the four fundamental forces. Further, it does not reveal the identity of dark matter, which makes up five times the amount of visible matter in the universe. With hopes of solving mysteries as these, particle physicists around the world attempt to improve our current understanding the Standard Model and to find physics beyond the Standard Model [1].

One experimental method of studying these particles is with particle accelerators such as the Large Hadron Collider (LHC) located at CERN in Geneva, Switzerland. The LHC is a large circular accelerator with a 17-mile circumference that accelerates beams of protons to nearly the speed of light and collides them at one of the four detectors located throughout the ring. These detectors, including the general-purpose Compact Muon Solenoid (CMS) detector, capture the particles that ensue from the collisions of protons. The CMS Collaboration, of which

the University of Notre Dame is a member, designed, built, and operates the detector and analyzes data from it. The University of Notre Dame group helped design and build the Electromagnetic and Hadronic Calorimeters and continues to have a leading role in the operation and upgrade of these detectors.

Various subsections of the detector (described in depth in Chapter 2) read out data which is then analyzed by physicists to reconstruct particles and physical processes in the detector. One of these Standard Model particles that is detected by the CMS detector is a photon. Low-energy photons (50 MeV - 70 GeV) are common in the detector, and they were instrumental in the discovery of the Higgs boson, as it was the $H \rightarrow \gamma\gamma$ decay channel that was used in the discovery. A highly-accurate identification method (ID) for low-energy photons was essential to achieve this feat [2].

However, this ID for low-energy photons is not appropriate for all photons, as high-energy photons (>200 GeV) can look different from low-energy photons in the detector. Though a high-energy photon ID has previously been developed for a specific kind of physical analysis, that ID is not appropriate for more general usage in multiple analyses. There are good motivations to develop a highly-accurate and more general high-energy photon ID, as high-energy photons could be key to new physics beyond the Standard Model if observed with the right physical processes.

As an example, in some theoretical models of dark matter and graviton production, there are a variety of final states that consist of a single high-energy photon and additional missing energy E_T^{miss}. All three Feynman diagrams shown in Figure 1.1 represents high-energy monophoton states with new physics involved. Just as the Higgs boson was discovered with a highly-accurate ID for low-energy photons, the searches for these theorized particles can only succeed with a highly-accurate ID for high-energy photons [3]. The goal of this thesis is

to develop an improved and more general ID for high-energy photons that can be utilized in multiple analyses moving forward, including these searches for new physics beyond the Standard Model.

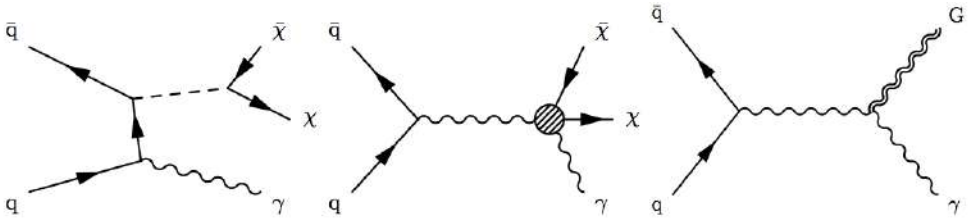

Figure 1.1: Theorized high-energy monophoton states with potential new physics beyond the Standard Model including dark matter models (left, middle) and a graviton model (right), where q = quark, γ = photon, χ = theorized dark matter particle, and G = theorized graviton particle [3].

To develop the ID, Monte Carlo simulations involving high-energy photons will be used where the identity of the particles are known exactly. A selection-based approach will be used for the ID, meaning that several variables characteristic of high-energy photons will be taken and given a range of selection criteria. The optimal set of selection criteria will demonstrate the ability to pick out high-energy photons from everything else in the simulations. Once satisfactory performance is demonstrated with simulations, the ID can be applied to data.

The paper is organized as follows: Chapter 2 presents a brief overview of the LHC and the CMS detector, Chapter 3 presents some important concepts involved in developing a particle ID, Chapter 4 presents variable distributions of the Monte Carlo simulations used and the process involved in the optimization of this ID, Chapter 5 presents the final sets of selection criteria and the performance of the ID, and Chapter 6 presents the conclusion of the paper.

Chapter 2

The Large Hadron Collider and the CMS Experiment

2.1 The Large Hadron Collider

The Large Hadron Collider [4] is an underground proton-proton accelerator with proton bunch crossings every 25 ns. Its designed center of mass energy is $\sqrt{s} = 14$ TeV, the highest energy accelerator in the world currently.

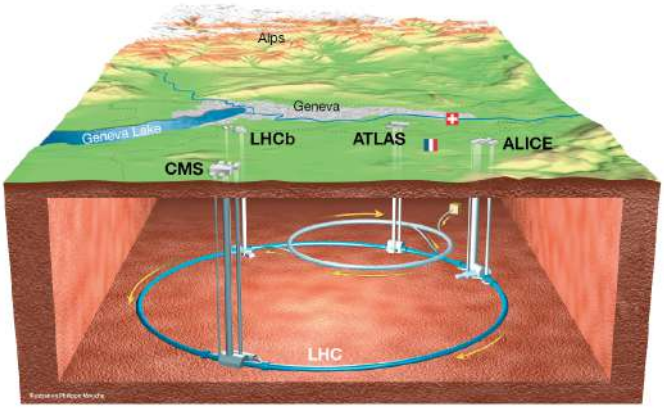

Figure 2.1: An overview of the experiments located at the LHC [5].

The LHC uses radiofrequency acceleration to accelerate the beams of protons. With appropriate waveguides and resonant cavities in place, a longitudinal E field is created that accelerates the protons from 450 GeV to 6.5 TeV. Many superconducting magnets (chilled by liquid Helium at 1.9K) are used to steer the beams of protons in the circular shape. The two beams move in opposite directions, and quadrupole magnets are utilized to tighten the beams to ensure they cross paths inside each of the four detectors located throughout the LHC ring. These magnets are located at both ends of the detectors, consisting of four magnetic poles arranged symmetrically around the beams to squeeze them vertically and horizontally.

2.2 The CMS Detector

The CMS detector is one of the two general-purpose detectors located at the LHC, designed to discover the Higgs boson and to measure existing particles and phenomena with high precision. Figure 2.2 shows a detailed diagram of the detector. Ideally, the energy, momentum, and identity of all particles produced become known through the various subsections of the detector.

2.2.1 Coordinate Conventions

First, the coordinate systems utilized by the CMS detector is shown in Figure 2.3.

With the origin centered inside the detector at the collision point, the x-axis points towards the center of the LHC ring, the y-axis points vertically upwards, and the z-axis points in the direction of the beam. The polar angle θ is measured from the z-axis on the z-y plane, and the azimuthal angle ϕ is measured from the

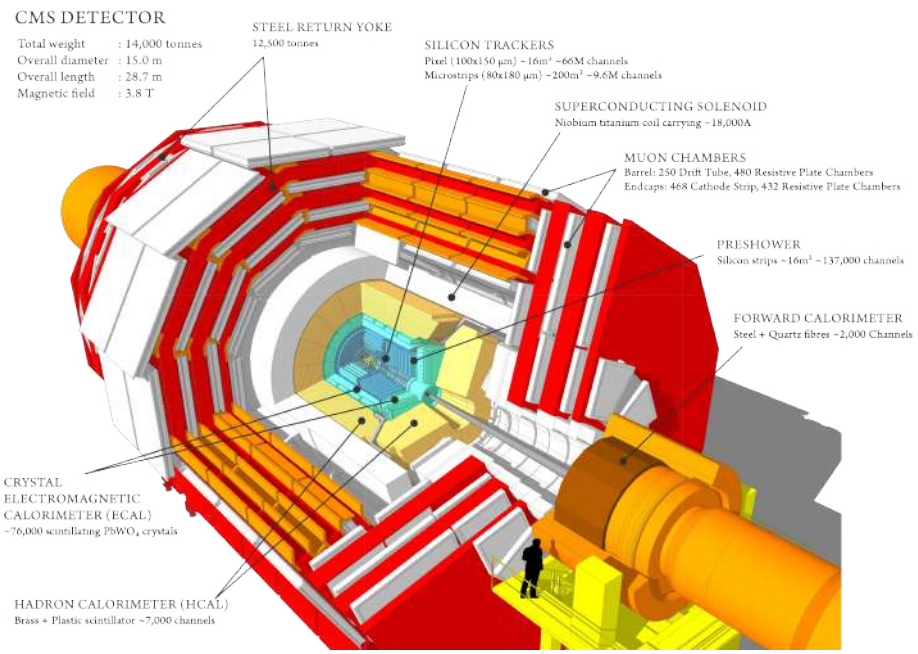

Figure 2.2: A diagram of the CMS detector.

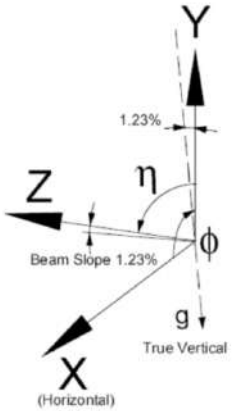

Figure 2.3: Coordinate system used in CMS

x-axis on the x-y plane. Pseudorapidity η is defined as

$$\eta = -\log \tan (\theta/2). \tag{2.1}$$

$\eta = 0$ corresponds with the direction perpendicular to the beam, and $\eta = \infty$ corresponds with the direction parallel to it.

2.2.2 Tracking System

Addressing the detector subsections from the inside out, the tracker [6] is the sub-detector located nearest to the interaction points. It measures the momentum and vertices of particles with high precision. Charged particles are bent in a magnetic field and its momentum is calculated by

$$p = qrB. \tag{2.2}$$

By measuring the trajectory r of the particle with its known charge q and magnetic field B, momentum p can be easily calculated.

The tracker utilizes silicon technology, good for its radiation hardness and high granularity. When charged particles hit the silicon atoms, electrons from the atom gets released, creating electron-hole pairs. With a electric current applied, the electron and the hole drift to negative and positive surfaces to be read as a pulse of charge. There is a silicon pixel detector for the inner part and a silicon strip detector for the outer part of the tracking system. The inner pixel detector is made of 65 million 150×150 μm silicon pixels that measure the exact position of the particle in both the x and y dimensions. The outer silicon strips are thinner but cover more area overall, and recorded hits in the silicon strips are used to create tracks of particles that calculate the momenta of the particles.

2.2.3 Electromagnetic Calorimeter

The Electromagnetic Calorimeter (ECAL) encloses the tracker. It is designed to measure energies of electrons and photons with high precision. It is a homogenous calorimeter made of almost 76,000 lead tungstate ($PbWO_4$) scintillating crystals that are divided into a barrel ($| \eta | < 1.479$) and two endcaps ($1.48 < | \eta | < 3.0$). As a result of the high-density (8.28 g/cm^3) crystals, the calorimeter is fast, has fine granularity, and is radiation-resistant.

When electrons, positrons, and photons hit the ECAL, they produce electromagnetic showers. These arise because the electron or positron can bremstrahlung, a process in which it radiates a photon and loses energy. The photon then pair produces an electron and a positron, also losing energy in the process. At a critical energy of at least 20 MeV, these processes occur one after another, creating a big cascade of electromagnetic activity called an electromagnetic shower, as shown in Figure 2.4.

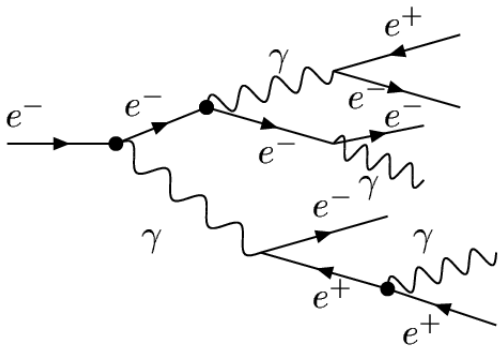

Figure 2.4: Diagram of an electromagnetic shower, γ = photon, e^- = electron, and e^+ = positron [7].

The lateral profile of the electromagnetic shower is a Gaussian with a width of the Moliere radius of the $PbWO_4$, which is 2.2 cm. About 94% of the energy from a particle is contained within the 3×3 array of crystals centered around

the crystal with the most energetic deposit, and about 97% of the energy is contained within the 5×5 array of crystals. Further, a "cluster" of ECAL crystals is a group of crystals, defined in the following method: starting from the local energy maximum, adjacent crystals are added to the cluster unless there is a rise in energy or there is no energy deposit at all. These clusters can then be clustered in a similar method to form "superclusters," with the main difference of the superclusters being extended in the ϕ direction. They are clustered this way because the electromagnetic showers in the ECAL often spread in ϕ due to the presence of a strong magnetic field in the detector.

2.2.4 Hadronic Calorimeter

Surrounding the ECAL is the hadron calorimeter (HCAL). It is designed to measure the energy and direction of jets, which are high-energy showers of mesons and baryons that arise from quarks and gluons. Charged hadrons do not shower electromagnetically so they pass the ECAL without interacting, then they interact with the brass in the HCAL. The resulting ionizing particles are then detected by plastic scintillator tiles that scintillate upon interaction. These tiles are read out by wavelength-shifting fibers that carry the light to the connected readout system that is based on multi-channel hybrid photodiodes. The HCAL is also divided into a barrel and two endcaps.

2.2.5 Magnet

The magnet of the CMS detector surrounds the tracker and the calorimeters. It is a powerful superconducting solenoid with mass 500 t and magnetic field strength of 3.8 Tesla. Its magnetic flux is returned by an iron yoke.

2.2.6 Muon Detector

The muon system [8] is the outermost subsection of the CMS detector, designed to identify tracks of high-penetrating muons and measure their momentum. Muons are the most penetrating particle in the detector, so the muon system is located in the outermost part of the detector. Most other particles are stopped before they reach the muon system.

It relies on three types of gas detectors: drift tubes (DT), cathode strip chambers (CSC), and resistive plate chambers (RPC). The layers of the muon detector are interwoven with layers of the return yoke of the magnet.

2.2.7 Triggers

The trigger system is what begins the physics event selection process, as it makes decisions every 25 ns on whether or not events should be retained for further consideration. The system has an input rate of 10^9 interactions per second, but this must be reduced by a factor of 10^7 down to 100 Hz, the maximum rate at which data can be recorded to disk. CMS utilizes a two-level trigger system for this. The Level-1 Trigger consists of custom electronics as a part of detector hardware and reduces the data rate to nearly 100 kHz. It holds the data in a "latency buffer" while it decides if an event is interesting, and if it is interesting, then it passes it on to the High Level Trigger. The High Level Trigger is a subset of the processing farm and further reduces the data rate to a maximum of 100 Hz.

2.2.8 Other Important Concepts

Aside from all of the detector subsections, there are a few quantities and concepts relating to the detector that are important to note. One of them is the transverse

momentum, the momentum of particles measured in the direction transverse to the beam, defined as

$$p_T = p \sin \theta. \tag{2.3}$$

For photons that travel at the speed of light, $E = p$. Because of this, what was introduced in the Introduction as high-energy photons are really high-p_T photons. "p_T" will be used instead of "energy" from this point on.

Another important concept is pile-up. In each bunch crossing, there are multiple proton-proton collisions that overlap in the detector, known as pile-up. The amount of pile-up can be measured with the median energy density ρ, defined as

$$\rho = median\left[\frac{p_{T\,j}}{A_j}\right], \tag{2.4}$$

where j iterates over all of the jets reconstructed in the event, and A_j is the jet area.

Chapter 3

Particle Identification Concepts

In order to begin understanding the physics of the data coming from the detector, particles in the detector must be reconstructed and identified. This chapter focuses on a few important concepts involved in developing an identification method (ID) for particles, specifically for high-p_T photons.

3.1 Existing Photon Identification in CMS

Photons in the detector can have a large range of p_T, from 50 MeV to 2 TeV. An ID developed for low-p_T photons applied to high-p_T photons does not perform well due to the differences in the distributions of photon identification variables (described in Section 4.4). For this reason, the identification of photons is divided to a "standard" photon ID (roughly between 50 MeV and 70 GeV) [9] and a high-p_T photon ID (roughly between 70 GeV and 2000 GeV). Though a high-p_T ID has previously been developed for a specific kind of analysis [3], that ID is not appropriate for more general usage in multiple analyses. For this reason, we develop an improved and more general high-p_T photon ID that can be utilized in multiple analyses.

3.2 Signal vs Background Events

"Signal" events are simply defined to be the particles that we wish to identify. In our case, they are high-p_T photons from final states of physical processes within the CMS detector. The "background" events are events that mimic the signal photons but are not actual signal photons. Examples of a common background event are π_0's that arise in hadronic jets. The π_0 can decay into two lower-p_T photons, and because the particles are traveling at relativistic speeds, the two resulting daughter photons travel in the same direction as the π_0 (Lorentz boosted) and are very closely spaced. Because the two photons are so closely spaced, they create what is called a "merged cluster," leaving an energy deposit in the ECAL that is similar to that of a single high-p_T photon. The goal of the high-p_T photon ID is to differentiate the signal events from these background events as efficiently as possible.

3.3 Efficiency & Purity

In any given set of events, there is a number of signal events S and a number of background events B. After applying a selection cut on these events, there is a new number of signal events S_{cut} and a new number of background events B_{cut}. The signal efficiency associated with this selection is $\frac{S_{cut}}{S}$ and the background efficiency $\frac{B_{cut}}{B}$. The signal purity associated with this selection is $\frac{S_{cut}}{S_{cut}+B_{cut}}$.

3.4 Selection-Based Identification

A selection-based identification method is pursued for the optimization of this ID. The selection-based method takes several variables characteristic of the signal events and applies to them a range of selection criteria that best separate the

signal events from the background events. An optimized set of selection criteria provides the lowest possible background efficiency level for all signal efficiency levels. If the selection ranges on the variables are narrow, there will be a lower signal efficiency with a higher signal purity. If the ranges are wide, there will be a higher efficiency with a lower purity. There is a balance to be struck between the efficiency and the purity depending on the analyses that will use the ID. Since the needs of future analyses are not known, several working points are created — 70% signal efficiency (called tight ID), 80% signal efficiency (called medium ID), and 90% signal efficiency (called loose ID) — that future analyses can choose from.

3.5 Monte Carlo Simulations

Monte Carlo (MC) simulations [10] are computer simulations that replicate the CMS detector and the particle interactions that take place within it. They have been tuned against data from the detector to give a reliable representation of it. These simulations are essential in developing highly-accurate particle IDs. With data from the detector, the identities of the particles are not exactly known and the signal events cannot be differentiated from the background events. However, with MC simulations, the identities of the particles are exactly known and the signal events can be differentiated from the background events. Using this information, the particle ID can be precisely developed and its performance calculated.

Chapter 4

Optimization of the Identification

4.1 Simulation Samples Used

There are two different simulation samples that are used for the optimization of this ID, γ + jets and monophotons.

The γ + jets sample simulates a physical process that involves one photon and a hadronic jet of quarks or gluons in the final state of a physical interaction, produced with the best current understanding of quantum chromodynamics and the Standard Model. Because the π_0 background comes from these hadronic jets, there is ample background activity with this γ + jets sample.

The monophotons sample simulates a theorized physical process that involves one photon and a graviton in the final state, as in Figure 1.1. A graviton is a hypothetical elementary particle that mediates the force of gravity. It has not been discovered yet, and if it were to be discovered using the CMS detector, it would only be known through a large E_T^{miss} rather than through direct detection [3]. Because of this, the final state of the theorized graviton process would have a single high-p_T photon, simulated with the monophotons MC sample. The monophotons sample consists of mainly signal events with a small number of

background events, while the γ + jets sample provides higher statistics on both the signal and background events. Because of the advantage of higher statistics, the γ + jets sample will be used to carry out the ID optimization. Once the ID is optimized, its performance will be checked utilizing the monophotons sample to ensure that the performance of the ID is consistent with both the monophotons sample and the γ + jets sample.

4.2 Requirements for Signal and Background

In order for a reconstructed photon (with $p_T > 70$ GeV) in the MC samples to be considered a signal high-p_T photon, it must pass some requirements with respect to the associated generated photon in the MC simulation. The reconstructed photon must 1) have spatial difference $\Delta R < 0.1$ from the generated photon, 2) be in the same bunch crossing as the generated photon, and 3) pass the conversion-safe electron veto, where the spatial ΔR is defined to be

$$\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}. \tag{4.1}$$

To understand the conversion-safe electron veto, converted photons must be understood first. Photons traveling in the tracker have substantial probabilities of converting into electron-positron pairs, and if converted, the e^-e^+ track pairs and their vertices can be reconstructed. The conversion-safe electron veto requires that all charged-particle tracks in the inner layer of the pixel detector are matched to a reconstructed conversion vertex. A charged-particle track that is not matched to a reconstructed conversion vertex is most likely an electron, which is the particle we want to veto. We apply this conversion-safe electron veto to avoid misidentifying electrons as photons.

Background events are then objects (with $p_T > 70$ GeV) reconstructed to

be photons that either 1) have $\Delta R > 0.1$ from the generated photon, 2) are not in the same bunch crossing as the generated photons, or 3) do not pass the conversion-safe electron veto.

4.3 Basic Variables

With the signal and background events defined, we can now compare distributions of p_T, η, and ρ between signal and background events in the γ + jets MC sample. For these three variables, the distributions of the signal and background events should be ideally identical, as there is no particular reason for any differences in the number of signal or background events to be present at specific p_T, η, or ρ ranges. In fact, the performance of the ID should be independent of these quantities, performing equally well across the entire range of each p_T, η, and ρ. This performance check is demonstrated in Section 5.3. Distributions of signal events (blue) and background events (red) from the γ + jets sample are normalized to 1 and overlaid (p_T in Figure 4.1, η in Figure 4.2, and ρ in Figure 4.3), separated by the ECAL endcap and barrel.

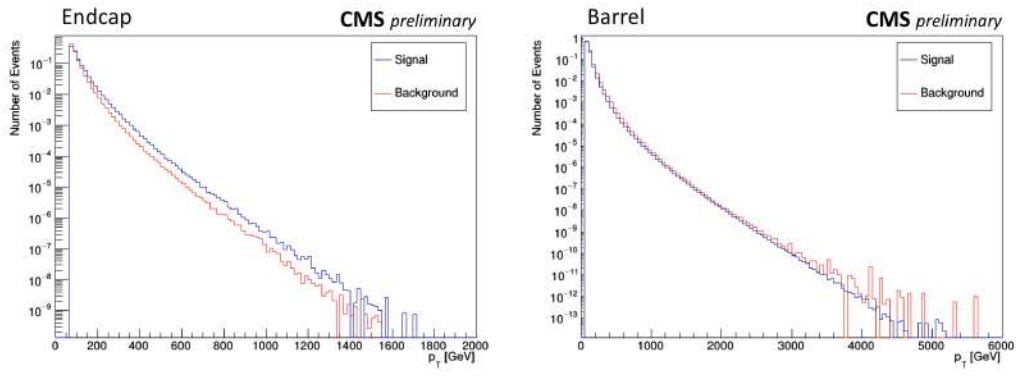

Figure 4.1: p_T distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

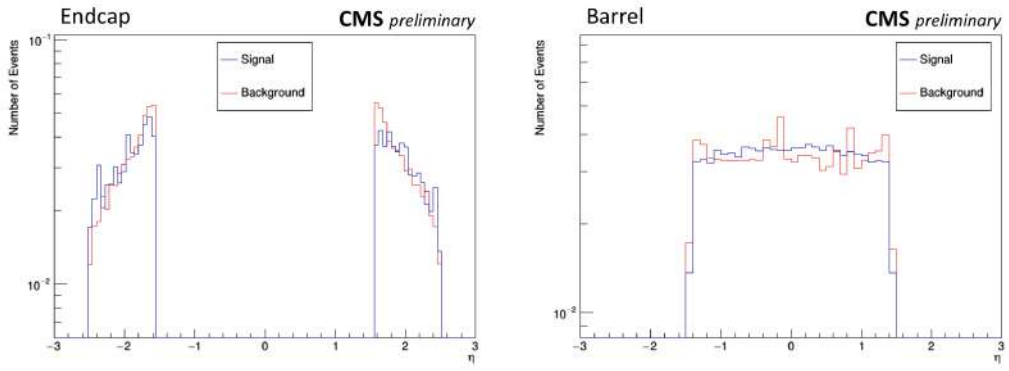

Figure 4.2: η distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

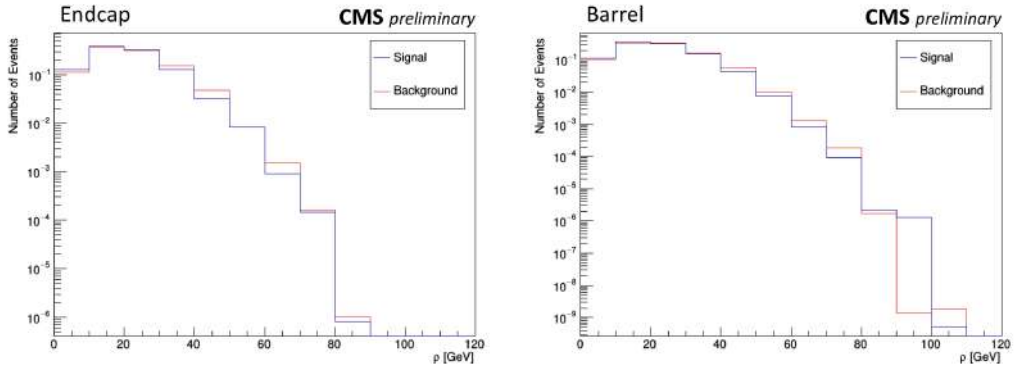

Figure 4.3: ρ distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

4.4 Photon Identification Variables

The variables that are utilized in the ID optimization are shower shape variables and isolation variables. These are variables that have proven to be most useful in separating the signal and the background events for previous photon IDs such as the standard ID. Descriptions of these variables and their distributions in the γ + jets MC sample follow.

4.4.1 Shower Shape Variables

As the name suggests, shower shape variables describe the shape of the electromagnetic shower that is associated with a photon. $\sigma_{\eta\eta}$ is the lateral extension of the shower, which is the variance in the η direction. Hadronic jets have a wider shape in the η direction, making them have larger values of $\sigma_{\eta\eta}$. The distributions of the signal and background events for $\sigma_{\eta\eta}$ are shown in Figure 4.4.

R_9 is $\frac{E_{3\times3}}{E_{SC}}$, where $E_{3\times3}$ is the energy sum of the 3×3 crystals centered around the most energy crystal and E_{SC} is the energy sum of the Super Cluster. A converted photon has a lower R_9 value since the energy is more spread out and not concentrated in the 3×3 crystals around the most energetic crystal. An unconverted photon, however, has a higher R_9 value since the energy is not spread out and concentrated in the 3×3 crystals around the most energetic crystal. The distributions of the signal and background events for R_9 are shown in Figure 4.5.

Lastly, H/E is the ratio of the hadronic energy in the HCAL towers behind the Super Cluster to the total energy of the ECAL Super Cluster. Both charged and neutral hadrons can interact with both the ECAL and the HCAL, meaning they can deposit some energy in both. However, photons interact only with the ECAL and deposit very little energy in the HCAL, so H/E for background events should be higher than signal photons. The distributions of the signal and background

events for H/E are shown in Figure 4.6.

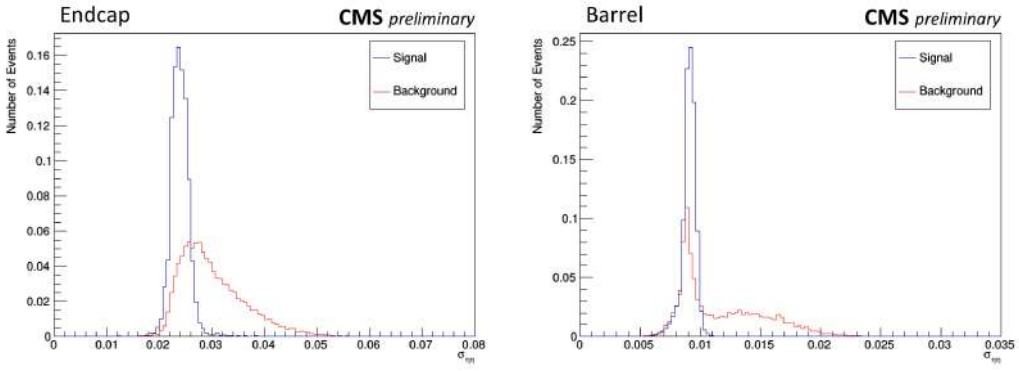

Figure 4.4: $\sigma_{\eta\eta}$ distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

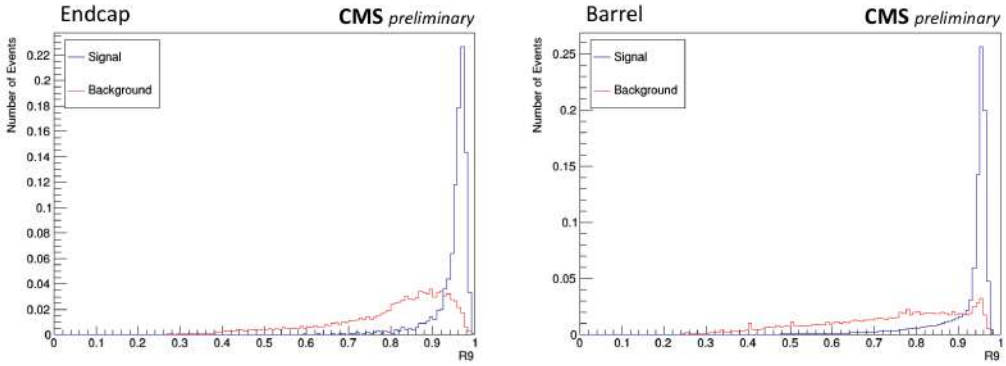

Figure 4.5: R_9 distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

4.4.2 Isolation Variables

Another type of variables utilized in the photon ID are isolation variables, which measure the amount of energy near the reconstructed photon. Background events are often produced in association with other particles which lead to higher values of isolation, but signal photons are uncorrelated with other events which lead

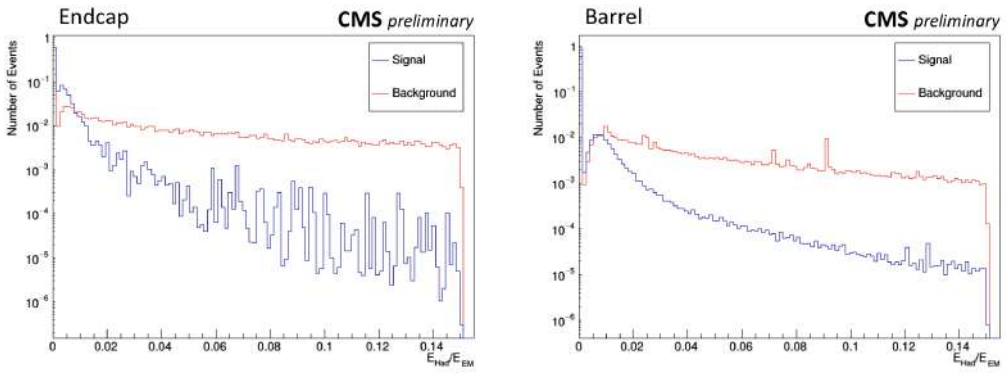

Figure 4.6: H/E distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

to lower values of isolation. Isolation variables are calculated utilizing particle-flow event reconstruction, the algorithm for which combines information from all subsections of the detectors and reconstructs four-momenta of all particles in the event. There are three isolation variables: Charged Hadron Isolation (I_π), Neutral Hadron Isolation (I_n), and Photon Isolation (I_γ). I_π is calculated by taking the sum of the p_T of tracks originating from the primary vertex and within a cone of size $\Delta R < 0.3$ centered around the line joining the primary vertex and the ECAL Super Cluster. I_n is calculated by taking the sum of the p_T of HCAL energy deposits in crystals located in a cone of size $\Delta R < 0.3$ centered on the shower axis. I_γ is calculated in the same way, just with ECAL energy deposits rather than the HCAL energy deposits. The distributions of I_π, I_n, and I_γ are shown in Figure 4.7, 4.8, and 4.9.

4.5 Isolation Corrections

The isolation cones defined for each of the isolation variables above get contaminated by the energy from pile-up. Because of this, the isolation variables are

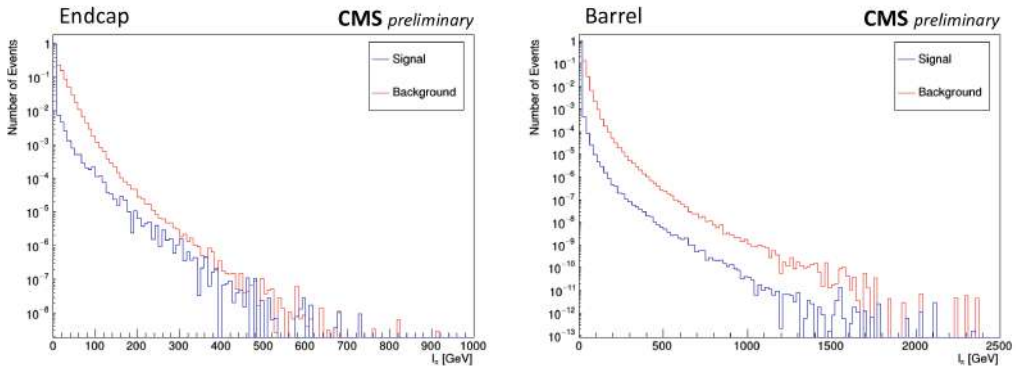

Figure 4.7: I_π distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

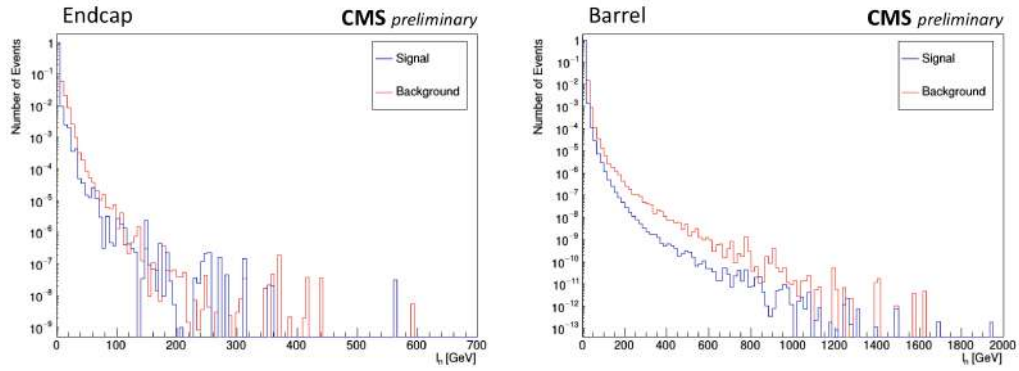

Figure 4.8: I_n distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

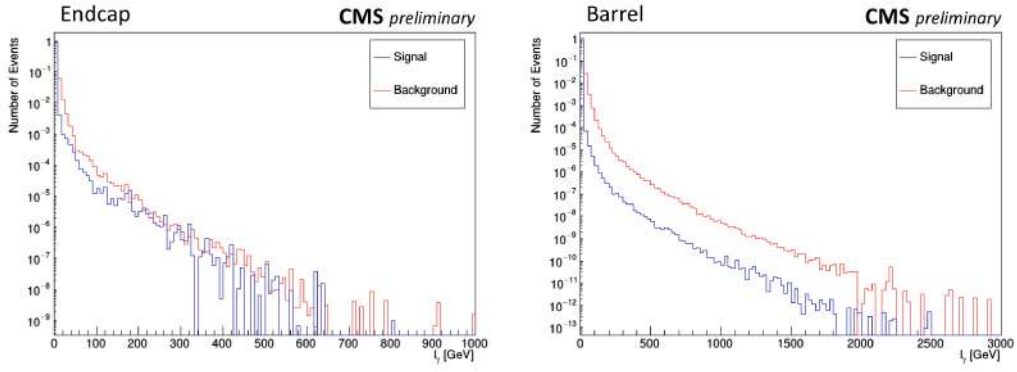

Figure 4.9: I_γ distributions in the γ + jets MC sample for signal vs background events, both normalized to 1, separated by endcap and barrel.

strongly dependent on ρ, and this dependence must be corrected. As examples, 90% quantile plots of I_n and I_γ vs ρ for the barrel are shown in Figure 4.10. Without a correction, a single cut on a isolation variable would lead to varying efficiency levels depending on the corresponding ρ values. The corrected isolation variables I_{Corr} take the form

$$I_{Corr} = I - \rho \times A_{Eff}, \tag{4.2}$$

where effective area A_{Eff} corresponds to an effective isolation cone. Empirically, the A_{Eff} is the slope of the linear fit of the 90% quantile plots of isolation vs ρ.

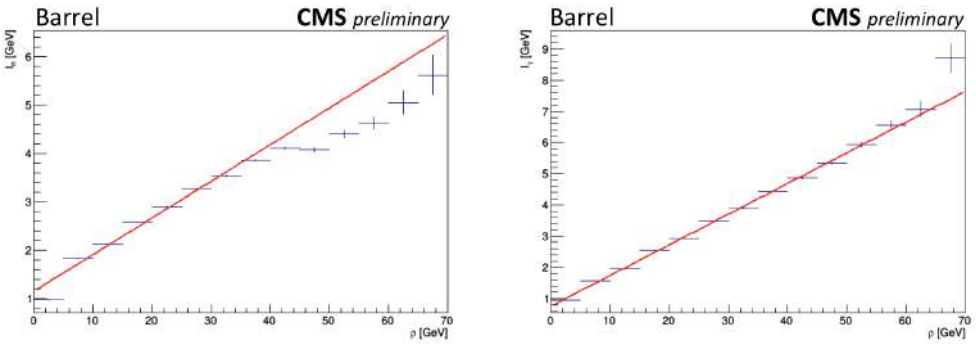

Figure 4.10: 90% quantile plots of I_n and I_γ vs ρ for the barrel.

Once the ρ dependence is corrected, an additional dependence of the isolation variables on p_T must be corrected through a similar procedure called p_T-scaling. The 90% quantile plot of I_{Corr} vs p_T can reveal no dependence, linear dependence, or a quadratic dependence. Having no dependence means that a correction is not required, but a linear or quadratic dependence means that an appropriate fit and a correction are required, utilizing Equation 4.3 or Equation 4.4 depending on the dependence.

$$I_{DCorr} = I_{Corr} - P1 \times p_T \tag{4.3}$$

$$I_{DCorr} = I_{Corr} - (P1 \times p_T) - (P2 \times p_T^2) \tag{4.4}$$

Figure 4.11 shows 90% quantile plots of corrected I_n and I_γ vs p_T in the barrel. Corrected I_n shows quadratic behavior and requires correction using Equation 4.4, while corrected I_γ shows linear behavior and requires correction using Equation 4.3. Once corrected, the double-corrected isolation variables are then no longer ρ and p_T dependent, and they are prepared to be utilized for the optimization process. Table 4.1 summarizes the A_{Eff} and p_T-scaling values for I_π, I_n, and I_γ.

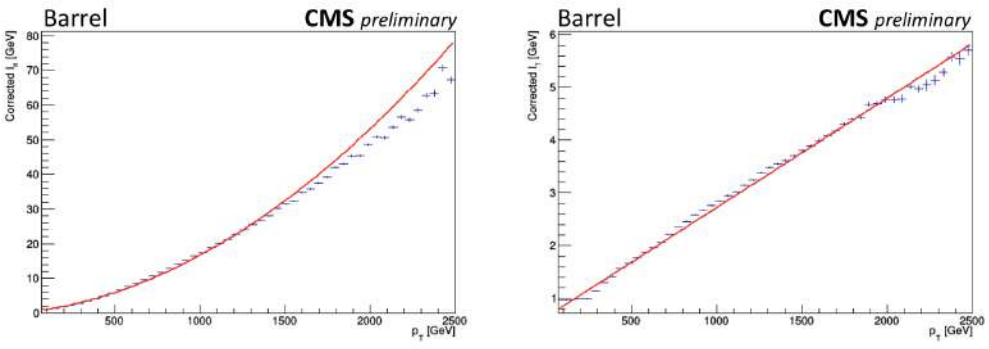

Figure 4.11: 90% quantile plots of corrected I_n and I_γ vs p_T for the barrel.

		I_π	I_n	I_γ
Endcap	A_{Eff}	0.008	0.046	0.090
	p_T-scaling P1	0	0.009	0
	p_T-scaling P2	0	1.5E-06	0
Barrel	A_{Eff}	0.005	0.076	0.098
	p_T-scaling P1	0	0.007	0.002
	p_T-scaling P2	0	1.0E-04	0

Table 4.1: Summary of A_{Eff} and p_T-scaling values for endcap and barrel.

4.6 Optimization Method

Once isolation corrections are carried out, the variables in the MC sample are ready for optimization of the ID. A selection-based method with shower shape variables and isolation variables is utilized. Not all six variables are necessarily helpful in the optimization, so different combination sets of variables are used during the optimization process, *e.g.* just shower shape variables, just isolation variables, some combination of both, etc.

For each combination set of variables, an optimization for the selection criteria is carried out. This optimization process consists of maximizing the background rejection at a given signal efficiency, and scans over the full range of signal efficiency. The background rejection is maximized by applying various selection ranges to the variables at hand and finding the selection ranges that lead to the highest background rejection level at each signal efficiency level.

This means that for each combination set of variables, there is an optimized set of selection criteria. The performance of the optimized selection criteria (associated with each combination set of variables) must be compared to one another to determine the best one, as explained further in the next chapter.

The optimization of the ID is carried out separately for endcap and barrel.

Chapter 5

Performance of the Identification

After running several iterations of the optimization, the best set of selection criteria is determined. The primary measure of performance of the selection critera is the area under the background rejection vs signal efficiency curve. The greater the area under the curve, the better the performance of the ID. The background rejection vs signal efficiency curve also provides an easy means to compare the performance of the newly developed ID to the performance of the standard photon ID applied to high-p_T photons. The three working points of the standard ID applied to γ + jets should fall under curve of the high-p_T photon ID.

Once the ID with the best performance is selected, there are two final steps to ensure the reliability of the ID. First, it should be ensured that the signal efficiency is not dependent on p_T, η, and ρ, *i.e.* signal efficiency levels should be flat when plotted against these variables. Second, the ID should be applied to the monophotons MC sample and checked that there is not a significant reduction in performance. Once these two inspections are completed, MC scale factors can be derived to check for differences in the ID performance when applied to real data from the detector.

5.1 Final Selection Criteria

After running the optimization, the sets of selection criteria listed in Table 5.1 and in Table 5.2 prove to be the best set of selection criteria for the endcap and for the barrel at loose, medium, and tight working points.

	Loose	Medium	Tight
Signal Efficiency	90%	80%	71%
Background Rejection	89%	92%	94%
HoE $<$	0.067	0.019	0.101
R9 $>$	0.940	0.955	0.960
ρ-corrected I_π $<$	3.59	18.35	2.86

Table 5.1: Final set of selection criteria for the endcap.

	Loose	Medium	Tight
Signal Efficiency	90%	80%	69%
Background Rejection	74%	81%	85%
$\sigma_{\eta\eta}$ $<$	0.011	0.028	0.012
HoE $<$	0.060	0.045	0.066
R9 $>$	0.938	0.944	0.947

Table 5.2: Final set of selection criteria for the barrel.

5.2 Performance Criterion

The criterion for determining the best performing ID for the endcap and barrel is the area under the background rejection vs signal efficiency curve for each ID. The three working points in Table 5.1 and Table 5.2 are simply three points from the background rejection vs signal efficiency curves, and the curves are shown in Figure 5.1.

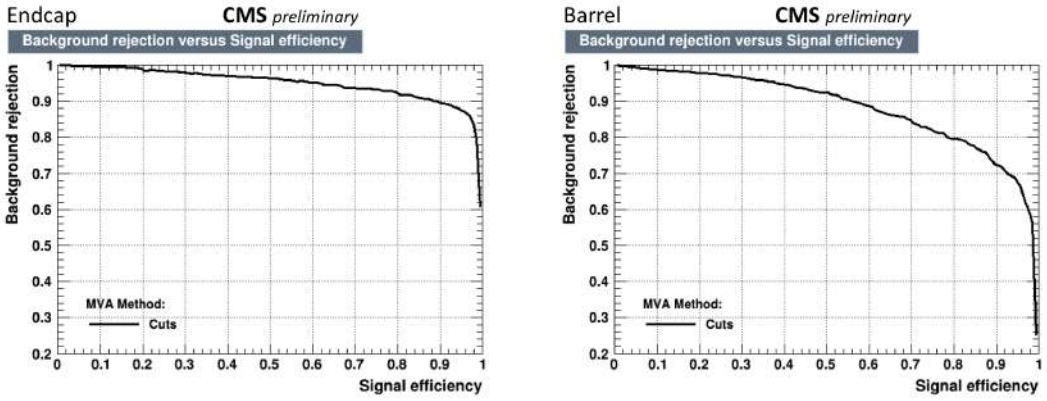

Figure 5.1: Background rejection vs signal efficiency curves for endcap (area under the curve 0.949) and barrel (area under the curve 0.882).

5.2.1 Comparison with Standard Photon Identification

To ensure that the newly developed set of selection criteria does a better job selecting high-p_T photons than the standard photon ID does, we apply the selection criteria of the three working points of the standard photon ID to the γ + jets MC sample and plot the resulting efficiency levels in the same plot as the background rejection vs signal efficiency curves, as shown in Figure 5.2.

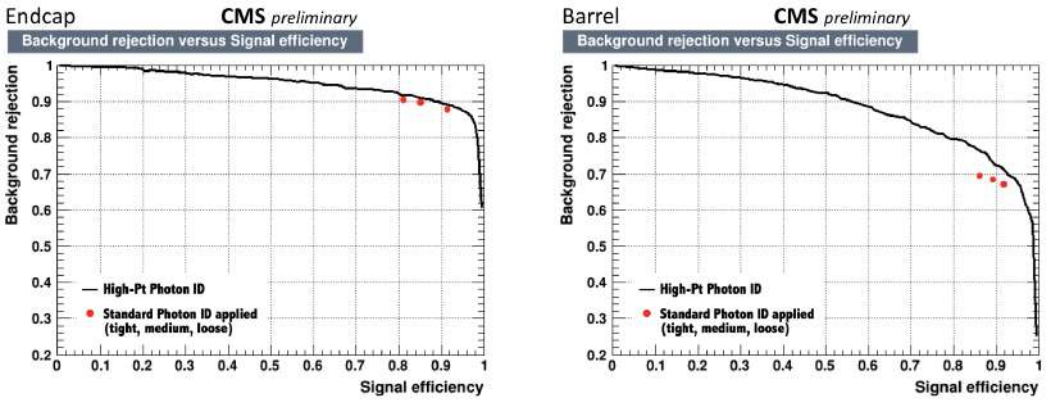

Figure 5.2: Working points of the standard photon ID applied to the γ + jets sample compared to high-p_T photon ID curve.

In both cases, the standard photon ID does not deliver the 70%, 80%, and 90%

signal efficiency levels that it should. Further, the three working points from the standard photon ID fall to the left of the background rejection vs signal efficiency curve of the high-p_T photon ID, meaning that the high-p_T photon ID performs better than the standard photon ID.

5.3 Efficiency Plots

Signal efficiency for all three working points should be constant at all ranges of p_T, η, and ρ, as the performance of the ID should not depend on these quantities. Figure 5.3 and Figure 5.4 show the signal and background efficiency plotted against these three variables for endcap and for barrel.

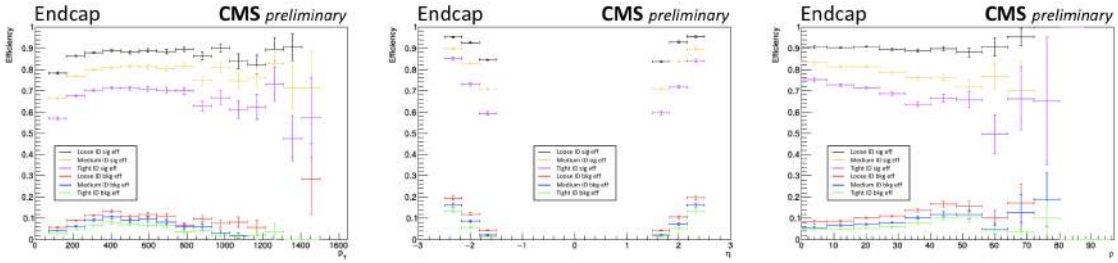

Figure 5.3: Signal and Background Efficiency (tight, medium, loose) vs p_T, η, and ρ for endcap.

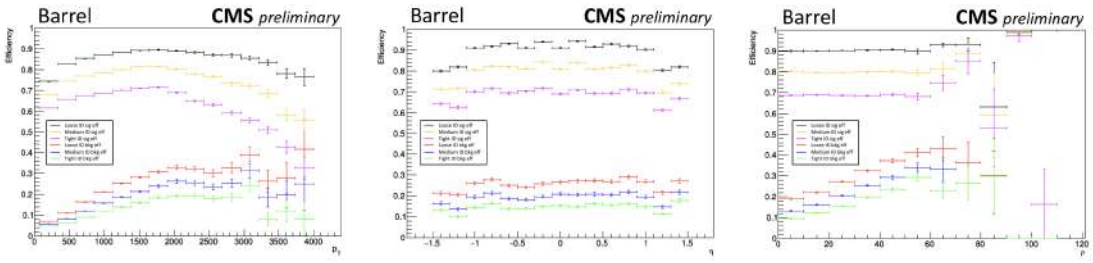

Figure 5.4: Signal and Background Efficiency (tight, medium, loose) vs p_T, η, and ρ for barrel.

The efficiency is flat for all of these plots except for the efficiency vs η in the

endcap. The reason for the this dependency is not clear, and further investigation should be conducted. However, the effect is not large enough to be a point of concern.

5.4 Performance on the Monophotons Sample

The high-p_T photon ID is developed using the γ + jets MC sample, but the performance of the ID should be checked with the monophotons MC sample as well. Ideally, the ID should result in the same levels of signal efficiency as it does in the γ + jets sample. Applying the ID to the monophotons sample, the signal efficiency levels for the three working points are as follows.

Endcap: 51%, 59%, and 77%
Barrel: 62%, 70%, and 78%

These fall short of 70%, 80%, and 90%, but especially given the lower statistics of the monophotons sample, the effect is once again not large enough to be a point of concern.

5.5 Calculating Simulation Scale Factors

The ID is developed with MC simulations, but the end goal of the ID is to apply it to data and have it perform equally well. To know how the efficiency levels of data will compare with the efficiency levels of MC samples, the final step of developing the ID is to calculate the data-to-simulation scale factors. These scale factors are useful in correcting any mismodelling in the simulations. The method of doing this is called the tag and probe method.

The tag and probe method uses a pure sample of leptons from Z boson decay in the data. Decays such as $Z \to e^+e^-$ are well-understood and have well-established narrow resonance. This method is used assuming that at the reconstruction level, electrons and photons are almost the same electromagnetic object in the ECAL. In this method, an electron from the Z boson decay is required to pass a set of very tight HLT requirements. This is called the "tag" candidate, and it must be matched to the triggering electron. The other electron is the "probe" candidate, and it is tested as to whether it passes the selection criteria developed by this ID. Passing probes and failing probes are used to compute the efficiency levels, and the scale factors are ratios of the data efficiency to the simulation efficiency of the ID. Figure 5.5 shows the scale factors and uncertainties, binned in p_T and η.

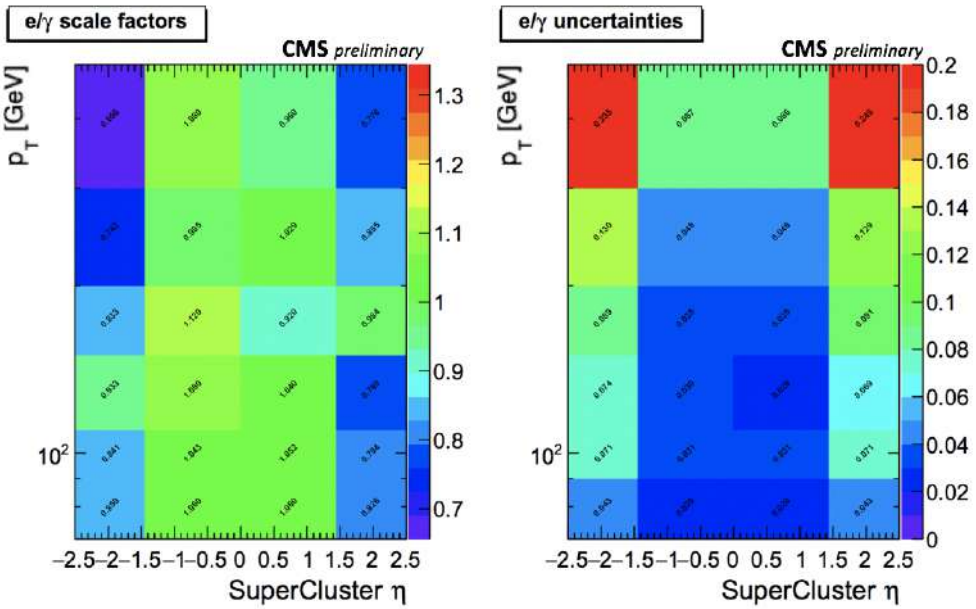

Figure 5.5: MC scale factors and uncertainties for the high-p_T photon ID.

The scale factors should be ideally within a few percentages of 1, and this is what we see in the barrel. In the endcap, however, the scale factors are low,

meaning that the ID results in higher efficiency levels in simulation than in data. This is possibly due to the enhanced effects of radiation damage in the endcap which are not precisely modelled in the simulation. However, further investigation will be conducted.

Chapter 6

Conclusions

The objective of this thesis was to optimize a high-p_T photon ID for the CMS detector located at the LHC. There have previously been high-p_T photon IDs developed for specific physical analyses, but they are not appropriate for more general usage for multiple analyses. So the goal of this project was to develop an improved and more general high-p_T photon ID that can be used in multiple analyses. This ID is motivated by the potential of new physics beyond the Standard Model that have final states which include high-p_T monophotons.

To do this, MC samples of γ + jets and monophotons were utilized. Because of the higher statistics in γ + jets, this sample was used to develop the ID. Shower shape variables and isolation variables were observed, with isolation corrections applied to correct for their ρ and p_T dependence. With these corrections in place, the ID was optimized, where different combinations of variables and selection criteria were tested to produce the best background rejection levels at every signal efficiency level.

The final sets of selection criteria of this newly developed high-p_T photon ID are provided in Section 5.1, developed separately for endcap and barrel of the ECAL. The performance of the selection criteria are checked using the background rejection vs signal efficiency curves, and it is shown that they perform better than

the standard photon ID applied to high-p_T photons. Efficiency levels for all three working points are relatively flat at all ranges of p_T, η, and ρ as they should be. The ID is applied to the monophotons sample to check performance on it, and though signal efficiency levels are not as quite good as they are with the γ + jets sample, the difference is not big enough to be a point of concern. Finally, data-to-simulation scale factors were derived using the tag and probe method to correct for any mismodelling that may be present in the MC samples. A highly-accurate and improved high-p_T photon ID has successfully been developed and is ready to be utilized in physical analyses moving forward.

References

[1] Steven Weinberg, "The making of the standard model," *Eur.Phys.J., C34:513*, 2004.

[2] CMS Collaboration, "Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc," *Phys. Lett. B 716 (2012) 30*, 2012.

[3] CMS Collaboration, "Search for new physics in the monophoton final state in proton-proton collisions at $\sqrt{}$(s) = 13 tev," *J. High Energ. Phys. (2017) 2017: 73.*, 2017.

[4] Thomas Sven Pettersson and P Lefvre, "The large hadron collider: conceptual design. technical report," *CERN-AC-95-05*, 1995.

[5] AC Team, "The four main lhc experiments," June 1999.

[6] Christian Piasecki, "Development of the cms tracker and reconstruction of secondary vertices of b and c hadrons," 2006.

[7] "Electromagnetic shower," *Robert Craig Group*, 2006.

[8] "The cms muon project: Technical design report," *Technical Design Report*, 1997.

[9] CMS Collaboration, "Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{}$(s) = 8 TeV," *2015 JINST 10 P08010*, 2015.

[10] S Banerjee, "Cms simulation software," *Journal of Physics: Conference Series, 396(2):022003*, 2012.