ETP-KA/2024-01

Enhancing the Phase Space for the Analysis of Inclusive $H \rightarrow b\bar{b}$ Production Through Trigger-Level Analysis at the CMS Experiment

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN
(Dr. rer. nat)

von der Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)
vorgelegte

DISSERTATION

von M.Sc. Adelina Lintuluoto
Tag der mündlichen Prüfung: 19/04/2024

Referent: Prof. Dr. Günter Quast
Korreferent: Prof. Paris Sphicas

Contents

Chapter 1

Introduction

Particle physics is a branch of science that examines the natural world at its most fundamental level. By studying the properties of the smallest building blocks in nature (known as *elementary particles*) and their interactions, particle physicists continue to advance our understanding of the universe.

In the early 1970s, our understanding of the fundamentals of particle physics was formulated into a unified quantum field theory known as the *Standard Model* (SM) of particle physics [1, 2]. Since its creation, the SM has become an established particle theory, and has been used to precisely predict the outcome of several foundational experiments. For example, elementary particles including gluons, the $W^{\pm}$ bosons, the Z boson, and the top quark were predicted by the SM prior to their experimental discovery. More recently, the SM successfully predicted the experimental discovery of the Higgs boson. This was significant, as according to the SM, several elementary particles acquire their masses through interactions with a field referred to as the Higgs field, which manifests itself as the Higgs boson.

Despite its universal acceptance, the SM is limited both theoretically and experimentally. From a theoretical perspective, the SM lacks an explanation as to why elementary particles such as leptons and quarks exist in precisely three generations, with similar properties but different masses. Furthermore, the SM does not include a theory of gravity. Experimentally, cosmological observations suggest that the SM is only able to explain about 16% of the total matter in the universe [3], with the rest being referred to as *dark matter*. In addition to missing a dark matter particle, the SM cannot explain the expansion of our universe associated with dark energy (which

accounts for approximately 68% of the universe [3]). Moreover, while understanding of the Higgs boson has advanced in the years since its discovery, current knowledge remains incomplete. The SM in its current state cannot, therefore, be considered a complete theory.

Today, particle physicists around the world and at the European Organisation for Nuclear Research (CERN) seek to address these limitations and complete our understanding of the universe by refining the current iteration of the SM. CERN is the largest particle physics laboratory in the world, and is host to the most powerful particle accelerator ever built (the *Large Hadron Collider*, or LHC [4]). The LHC accelerates beams of charged particles to speeds approaching the speed of light, which collide at the interaction points of four main experiments. One of the four experiments, the *Compact Muon Solenoid* (CMS), operates a general-purpose detector composed of layers of tracking detectors, calorimeters and muon detectors. Protons and ions accelerated by the LHC collide at the centre of the CMS detector, creating new particles that traverse the layers of the detector. Information from each layer is combined to reconstruct the collision event. The reconstruction is subsequently analysed to identify the particles involved and study the SM.

To observe all particles produced by the collisions, and thereby fully examine the SM, it is necessary for the CMS experiment to investigate all accessible regions of phase space. However, due to bandwidth limitations, only 0.005% (or 2 000) of the 40 million collisions that occur every second in the LHC can be stored and analysed. This is because the CMS detector reduces the read-out rate with a multi-stage trigger system by selecting events based on potential physics interests. First, the *Level 1 trigger* (L1T) hardware system reduces the read-out rate from 40 MHz to 100 kHz [5]. Then, the *High-Level trigger* (HLT) system, composed of a large computer farm, further reduces this rate to 2 kHz [6]. In this standard trigger strategy, only events that pass the HLT selection are available for further analysis. This limited capacity is a significant limitation of the CMS experiment, as valuable information that could lead to improvements of the SM might reside in regions of phase space excluded by the selection.

However, a technique known as *data scouting* can increase read-out rates by several kHz [7]. This allows a greater proportion of phase space to be studied. Unlike the standard trigger strategy (which stores the full event information in the form of raw detector signals), data scouting stores the trigger-level reconstruction of physics objects that have been identified

from the signals of the detector. The event size of trigger-level objects are much smaller, allowing data scouting to increase the read-out rate while having a negligible impact on the data acquisition system.

The technique of data scouting increases the rate of events passing the HLT by applying a less stringent selection process. This allows physicists to detect, analyse and explore events that would not previously have been allowed to pass the HLT. This provides opportunities for analyses outside the boundaries of the standard trigger strategy, and consideration of previously unexplored regions of phase space. The data scouting strategy is particularly advantageous for searches involving jets. As the majority of proton-proton collisions result in relatively uninteresting low-energy jet production from quark and gluon interactions, the standard trigger strategy must adhere to stringent energy and momentum thresholds to prevent overwhelming the data acquisition protocols. In contrast, the scouting strategy offers a notable reduction in these thresholds, providing a more flexible approach and allowing searches for interesting but rare physics processes involving low-energy jets.

While data scouting enables a wider range of analyses, the approach is limited by an inability to store the full event information. Without the raw detector outputs, it is not possible to reconstruct an event after collection, when a better understanding of the detector conditions allows for a more precise reconstruction. It is therefore imperative to assess the quality of the data scouting objects, and address potential discrepancies with dedicated correction studies.

This thesis addresses two important aspects of leveraging data scouting jets in high-energy physics research, and comprises two parts. First, in Part I, a discussion of the data scouting strategy used at the CMS experiment in 2022 and 2023 is provided. This includes a brief review of both the LHC (Chapter 2) and the CMS detector (Chapter 3), followed by an overview of the techniques used for collision event reconstruction at the CMS experiment (Chapter 4). An introduction to the CMS trigger system as well as details of the data scouting technique is provided in Chapters 5 and 6, respectively. In order to validate the data scouting technique, statistical analyses are performed on collision data recorded with the data scouting strategy in 2022. The goal of these studies is to assess the performance of data scouting jets with respect to jets part of the standard trigger strategy, and the results are presented in Chapter 7.

Part II focuses on the Higgs boson, and features an assessment of the vi-

ability of searching for Higgs bosons produced with high transverse momentum decaying to bottom quark-antiquark pairs using data scouting jets. The theoretical underpinnings and statistical methods used for this purpose are presented in Chapters 8 and 9, respectively. This is followed by a brief review of existing Higgs boson measurements (Chapter 10). An analysis conducted with data collected by the standard trigger strategy in 2016–2018 is then discussed in detail in Chapter 11. This study leads an investigation into the potential integration of scouting jets into the analysis. The potential of extending the search to a lower energy scale using collision data recorded through the data scouting technique in 2022–2023 is presented in Chapter 12. Finally, a conclusion and discussion of future prospects of Part I and II are given in Chapter 13.

Part I

Experimental methods

Chapter 2

Large Hadron Collider

The Large Hadron Collider (LHC) is a particle accelerator located near Geneva, spanning the border between Switzerland and France. The accelerator is housed in a circular tunnel approximately 100 meters beneath ground level, which has a circumference of approximately 27 km [4] . The tunnel was previously occupied by the Large Electron-Positron (LEP) collider — a particle accelerator that operated from 1989 to 2000 [8]. The LHC's depth provides adequate shielding against external radiation that might bias the detector measurements, while also absorbing the ionising radiation produced by the collider.

The LHC circulates and accelerates two particle beams in opposite directions. The beams collide at interaction points which are surrounded by detectors that select and store (referred to as *record* in the following text) the collision events. The particles comprising the beams determine the type of collisions that occur. These include proton-proton, lead ion-lead ion, xenon ion-xenon ion, lead ion-xenon ion and lead ion-proton collisions. Due to the focus of this thesis, the following discussion describes the proton-proton collisions. The LHC beams during proton-proton collisions are not a continuous stream of protons, but instead consist of thousands of *bunches*; each bunch comprising of around 10^{11} protons. The distance between the bunches leads to a bunch-bunch collision (or *bunch crossing*) taking place every 25 ns [4].

The primary functions of the LHC and its two general-purpose experiments, Compact Muon Solenoid (CMS) [9] and A Toroidal LHC ApparatuS (ATLAS) [10], are to study the mechanism of electroweak symme-

try breaking and search for physics beyond the Standard Model (SM). The LHC was built to advance current understanding of both areas across a range of energies (due to both the uncertainty surrounding the energy scale at which each may appear and the importance of studying their behaviour over several energy levels). This necessitates the use of hadron-hadron collisions. This is because hadrons, such as protons, are composite particles whose constituents (so-called *partons*) share varying fractions of the hadron's total energy. In proton-proton collisions, the partons of the two protons interact, resulting in a large range of collision energies even when the beam energy is kept constant.

This chapter provides an overview of the LHC, with a focus on the concepts utilised in this thesis. A more detailed description of the accelerator can be found in Ref. [4].

2.1 Centre-of-mass energy

The *centre-of-mass energy* is the energy available to produce new particles in collisions, and is determined by

$$\sqrt{s} = (p_1 + p_2)^2, \tag{2.1}$$

where p_1 and p_2 are the four-momenta of the two colliding particles.

When two protons from each beam collide head-on, the centre-of-mass energy is simply the sum of the two beam energies. The LHC is designed to achieve high centre-of-mass energies — up to 14 TeV with proton-proton collisions [4]. This is significantly higher than any previous particle accelerator and facilitates particle interactions at energy scales that were previously unattainable. For comparison, LEP achieved $\sqrt{s} = 210$ GeV, while Tevatron, the second most powerful particle accelerator to have existed, achieved $\sqrt{s} = 2$ TeV [11].

During the operational lifetime of the LHC, the centre-of-mass energy has gradually increased towards the 14 TeV target. After low energy tests in 2008–2009, the LHC started to operate with a $\sqrt{s} = 7$ TeV in 2010–2011, increasing to 8 TeV in 2012 (Run–1). During 2015–2018 (Run–2) and from 2022–present (Run–3), the $\sqrt{s}$ has been 13 and 13.6 TeV respectively.

2.2 Luminosity

Luminosity, in the context of particle accelerators, relates to the rate per area at which particles collide. Higher luminosity is desirable for many physics analyses as it increases the probability of rare physics processes occurring. In the context of the Higgs boson discovery, the production of Higgs bosons has increased with higher luminosity, allowing scientists at the CMS and ATLAS collaborations to study their properties with greater precision. There are two concepts of luminosity often referred to by LHC experiments, *instantaneous* and *integrated* luminosity, which are discussed in the following text.

2.2.1 Instantaneous luminosity

Instantaneous luminosity is a measure of the number of particle collisions that occur per unit area per unit time, and is often expressed in units of $\text{cm}^{-2}\text{s}^{-1}$. The instantaneous luminosity characterises a particle accelerator, and is computed as

$$L = \frac{N_1 N_2 n_b f \gamma}{2\pi\epsilon\beta^*}, \tag{2.2}$$

where N_1 and N_2 denote the number of particles in each colliding bunch, n_b the number of bunches, f the bunch revolution frequency, γ the relativistic factor, ϵ the normalised beam emittance and β^* relates to the beam size at the collision point.

As colliding protons are converted to energy, and the instantaneous luminosity depends on the number of protons in each colliding bunch, the instantaneous luminosity decreases with each beam revolution if machine optics are kept constant. However, a technique referred to as *levelling by* β^* can maintain a near constant luminosity. As the instantaneous luminosity reaches a lower tolerance limit, a β^* levelling step is automatically initiated to maintain current levels (illustrated in Fig. 2.1). Levelling by β^* is performed until machine optics are optimised to their limits, at which point the luminosity begins to decrease at a constant pace [12, 13].

The LHC was designed to operate at an instantaneous luminosity of 10^{34} $\text{cm}^{-2}\text{s}^{-1}$; two orders of magnitude higher than the LEP collider [11]. As

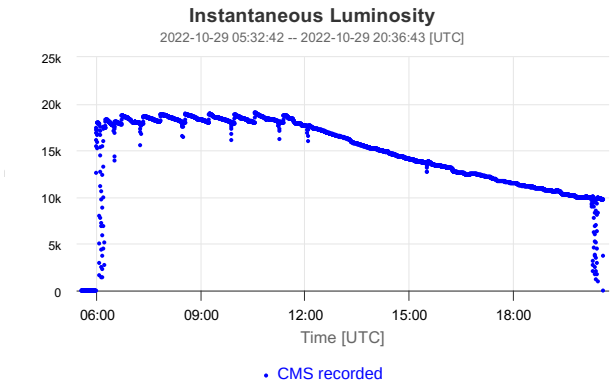


Figure 2.1: Instantaneous luminosity recorded by the CMS detector in units of 10^{30} cm^{-2}s^{-1} as a function of time, as displayed by the CMS Online Monitoring [14] tool. Between 06:00 and 12:00 UTC levelling by β^* is performed in order to maintain a near constant instantaneous luminosity. From 12:00 UTC, levelling stops and a constant decrease in luminosity is apparent. The plot displays LHC Fill 8321 at 29th of October 2022.

with the centre-of-mass energy, the instantaneous luminosity has gradually increased from Run–1 to Run–3. In 2016, the LHC reached its designed luminosity, a value it has since surpassed by a factor of more than two.

2.2.2 Integrated luminosity

Integrated luminosity is the cumulative measure of the total number of particle collisions that occur over a specific period. It is the integral of the instantaneous luminosity with respect to time, and is often expressed in units of inverse femtobarns (fb^{-1}) or inverse picobarns (pb^{-1}).

Integrated luminosity is crucial for assessing the overall performance of the CMS experiment. By comparing the integrated luminosity provided by the LHC with that recorded by the CMS detector (Fig. 2.2), a direct evaluation of the operations of the CMS detector is facilitated. In addition, integrated luminosity is an important input to many physics analyses. For example, it is used when comparing the number of predicted events from theory to those observed in collision data recorded by the CMS detector (referred to as *collision data* throughout the rest of this thesis).

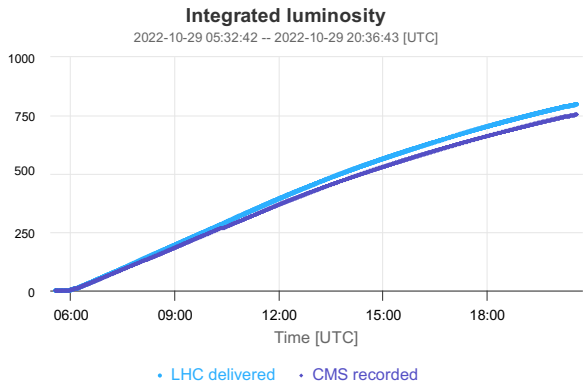


Figure 2.2: Integrated luminosity in units of pb^{-1} as a function of time, as displayed by the CMS Online Monitoring [14] tool. A comparison of the integrated luminosity delivered by the LHC (light blue) and recorded by the CMS detector (dark purple) is shown. The difference stems from inefficiencies in data collection by the CMS detector. The plot displays LHC Fill 8321 at 29th of October 2022.

2.2.3 High-luminosity LHC

To enhance the sensitivity of physics analyses currently constrained by their statistical uncertainty, a planned upgrade of the LHC aims to elevate the instantaneous luminosity. This will increase the number of collision events collected, thereby reducing the statistical uncertainty. Initiated in 2011, the upgrade named the *High-Luminosity Large Hadron Collider* (HL-LHC), sets out to achieve a five-fold increase in instantaneous luminosity and a tenfold increase in integrated luminosity compared to the LHC's nominal design values [15].

Elevating luminosity necessitates the reduction of the beam size at the collision point, coupled with either the reduction of bunch length and spacing, or a significant increase in both bunch length and number of protons. Either intervention poses unprecedented challenges to the accelerator infrastructure. This necessitates a dedicated research effort spanning over 10 years; the LHC experiments are projected to start data collection around 2025–2030 [15]. The undertaking involves a series of upgrades and enhancements, including the incorporation of advanced magnets, beam optics and cryogenic systems.

2.3 Pile-up

Due to the substantial number of protons involved in each collision, multiple proton-proton interactions can take place simultaneously within the CMS detector. These interactions can arise from the same or nearby bunch crossings. As a consequence, particles stemming from the *primary interaction* (the most energetic proton-proton interaction) are recorded along with particles originating from additional interactions. This phenomenon is referred to as *pile-up*, and is usually quantified in terms of mean number of interactions per bunch crossing.

Most interactions in a proton-proton collisions occur between low-energy quarks and gluons of the proton (so-called *QCD multijet production*). These interactions lack the energy to produce high-mass final state, resulting in the characteristic steeply falling spectra of both the QCD mass and momentum distributions. These low-energy interactions constitute the majority of the pile-up.

Due to the increase of instantaneous luminosity combined with the shortening of the bunch crossing time from 50 ns to 25 ns between Run–1 and Run–2 , the pile-up observed by the CMS detector has increased over the years. This is presented in Fig. 2.3. For comparison, the mean pile-up increased from 10 to 52 between 2011 and 2023.

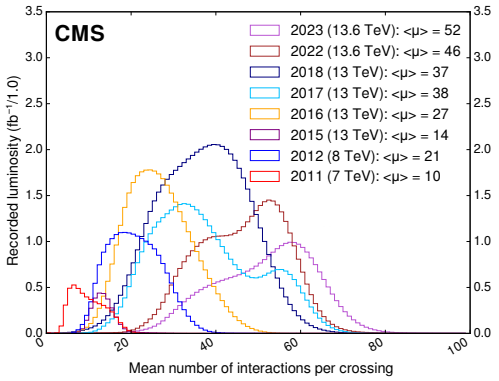


Figure 2.3: The pile-up distribution observed by the CMS detector, expressed as the mean number of interactions per bunch crossing, from 2011 to 2023. The mean of the distributions are listed in the legend, denoted as $\langle\mu\rangle$. The distributions are normalised to the integrated luminosity of the data recorded that year. Figure taken from Ref. [16].

Chapter 3

Compact Muon Solenoid

The Compact Muon Solenoid (CMS) experiment centres around a large, multi-purpose detector located at the LHC in Cessy, France. The name CMS is derived from three central features of the detector. While the CMS detector is the heaviest of the LHC detectors, weighing 14 000 tonnes, its dimensions of 22 $\times$ 15 m make it relatively *compact* when compared to the ATLAS detector (46 $\times$ 25 m). The majority of the weight stems from structural steel plates which are interleaved between *muon* chambers. These chambers enable precise measurement of the muons traversing the detector; a distinctive feature of the CMS experiment. Measurement of muon momentum is facilitated by a large superconducting *solenoid* magnet situated inside the detector. The magnet provides a strong magnetic field of 3.8 T within the detector, which bends the trajectories of charged particles for precise measurement of momentum and charge.

The detector is a near hermetic apparatus, consisting of several sub-detectors designed to identify charged and neutral hadrons, electrons, muons and photons [17–19]. Each sub-detector measures a particular property of particles traversing the detector, and when combined it is possible to identify these particles based on their unique signatures. The principle sub-detectors are a silicon pixel and strip tracking system, a lead tungstate crystal electromagnetic (EM) calorimeter, a brass and iron hadron calorimeter and muon chambers utilising gas-ionisation technology.

This chapter begins by providing an overview of the CMS detector, followed by an outline of the principle sub-detectors. A more detailed description of the detector can be found in Ref. [9].

3.1 Overview

The detector is situated 100 meters underground, around an *interaction point* where the beams of the LHC are configured to collide. The detector is designed to measure proton-proton collisions as well as heavy ion collisions. In the following text, the detector's configuration in 2022 and 2023 is described. Figure 3.1 illustrates a simplified cross-sectional view on the structure of the detector.

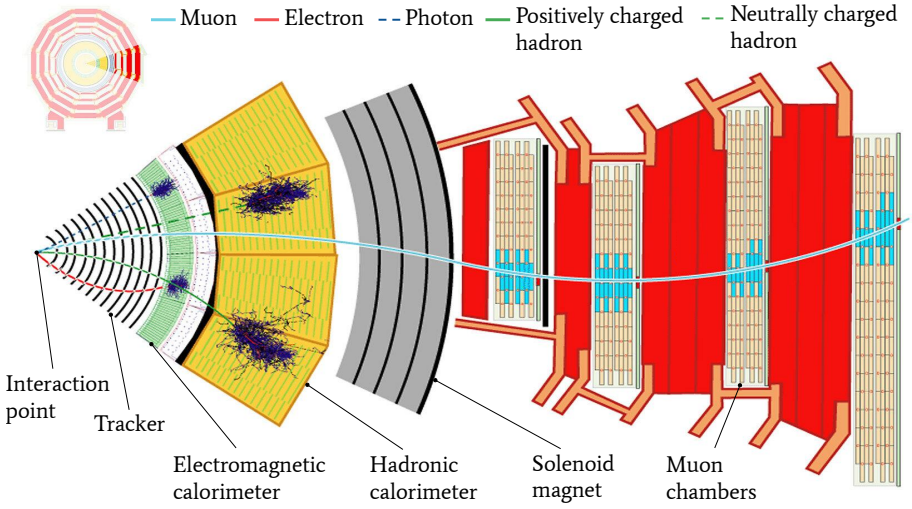


Figure 3.1: A simplified cross-sectional view of the CMS detector and its main sub-detectors. Example trajectories of the particles identified by the detector are displayed above the diagram. Illustration based on Ref. [20].

The CMS experiment employs a right-handed coordinate system, as depicted in Fig. 3.2. The origin of the system is located at the interaction point situated at the center of the detector. The x-axis points towards the center of the LHC, while the y-axis extends upwards, perpendicular to the LHC plane. The x- and y-axes span the *transverse* plane, where the azimuthal angle (ϕ) is defined. The z-axis aligns with the longitudinal axis of the CMS detector and points along the direction of the anticlockwise beam (as viewed from above). The polar angle θ is measured with respect to the positive z-axis. In practice, instead of θ, pseudorapidity (η) is commonly used as defined by

$$\eta = -\ln(\tan\frac{\theta}{2}). \tag{3.1}$$

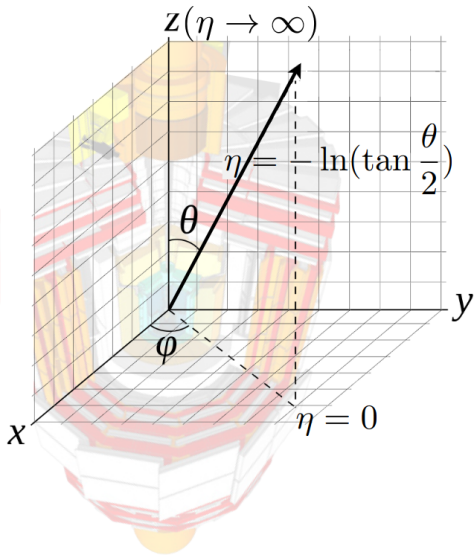


Figure 3.2: Illustration of the right-handed coordinate system used by the CMS experiment. Figure taken from Ref. [21].

Common physical attributes related to the coordinate system used by the CMS experiment are *transverse momentum* (p_{T}) and *rapidity* (y). p_{T} is the component of an object's momentum that is perpendicular to the z-axis. While high momenta in the direction of the beam-line do not necessarily indicate a high-energy collision, large momenta perpendicular to the colliding beams may be an indication of a high-mass particle decaying and emitting particles in opposite direction due to momentum conservation. Rapidity is used to express angles with respect to the axis of the colliding beams, and approximates to pseudorapidity at speeds approaching the speed of light. Rapidity is defined as

$$y = \frac{1}{2}\ln\left(\frac{E + p_{\mathrm{z}}}{E - p_{\mathrm{z}}}\right), \tag{3.2}$$

where E is the energy and p_z is the component of the momentum along the z-axis.

3.2 Tracking system

As newly created particles traverse the CMS detector, the first sub-detector encountered is the *tracking system* [22–24]. The tracking system is designed to measure the trajectory of charged particles, which facilitates vertex reconstruction in three spatial dimensions. The charge and momentum of a particle can then be deduced from the curvature of its trajectory. The tracking system is placed near to the interaction point, allowing the initial paths of charged particles to be captured before potentially destructive interactions with the detector material occur.

The tracking system is cylindrical and surrounds the interaction point. The part closest to the interaction point is equipped with a fine-granularity pixel detector, while the remainder of the system consists of larger strip modules (illustrated in Fig. 3.3).

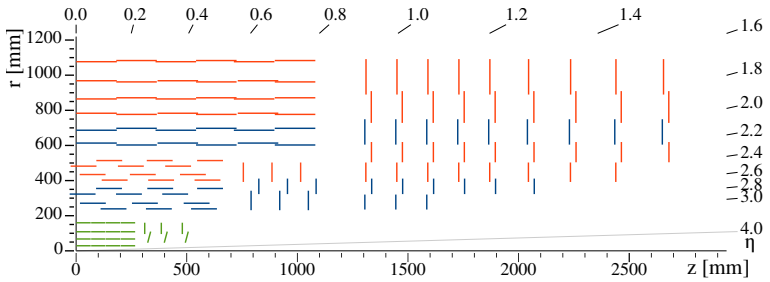


Figure 3.3: Illustration of one quarter of the tracking system; displaying the pixel detector (green) and strip modules (red and blue). Figure taken from Ref. [25].

At the LHC design luminosity, which produces an average of 20 pile-up interactions, around 1 000 charged particles traverse the tracker every 25 ns [26]. A detector technology featuring both high granularity and a fast response is therefore required. In addition, the intense particle flux due to its proximity to the interaction point causes severe radiation damage to the system, necessitating the use of a material capable of surviving such a harsh environment. The current solution to these requirements is a tracking system entirely based on silicon detector technology.

3.3 Electromagnetic calorimeter

The tracking system is surrounded by the *electromagnetic calorimeter* (ECAL) [27–29]. The ECAL is designed to measure the energy and position of electrons and photons produced in particle collisions. The ECAL is composed of a single layer of lead tungstate ($PbWO_4$) crystals, necessitated by its proximity to the interaction point. Lead tungstate crystals are capable of withstanding high levels of ionising radiation while maintaining an excellent energy resolution and a fast response time.

Lead tungstate is very dense, creating a high probability of interactions with electrons and photons. When an electron or photon collides with the atom of the the ECAL crystals, a collimated EM *shower* is created (comprising electrons, positrons, and photons). The shower particles deposit their energy in the lead tungstate crystals, causing the crystals to become excited. As the excited atoms return to their ground state they emit scintillation light. The light is then detected by photodetectors located behind the crystals. Figure 3.4 shows a lead tungstate crystal with the photodetector attached at one end. By measuring the intensity of the scintillation light, the ECAL can determine the energy of the initial electron or photon. The position of the deposited energy can be precisely reconstructed by analysing the distribution of light across the crystals. This type of calorimeter, which both absorbs and scintillates, is referred to as a *homogeneous calorimeter*.

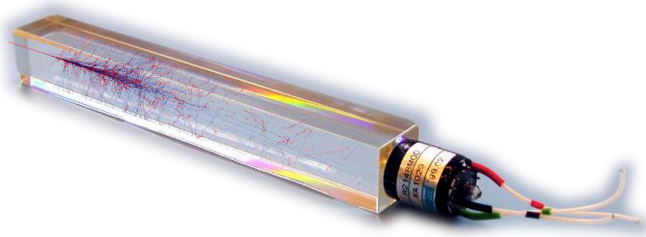


Figure 3.4: Photograph of an ECAL crystal and photodetector attached at the back. The illustration of an electromagnetic shower depositing its energy in the crystal has been overlaid. Figure taken from Ref. [30].

An additional detector referred to as the *pre-shower*, is placed in front of the ECAL in the endcap region on both sides of the CMS detector. Pre-shower detectors induce EM showers from highly-energetic particles before they

enter the lead tungstate crystals. The primary purpose of pre-shower detectors is to help improve the energy resolution and particle identification capabilities of the ECAL.

Although the primary purpose of the ECAL is to measure the energy of electrons and photons, hadrons also interact with the detector material and create showers within the ECAL. In order to distinguish between the EM and hadronic showers, collision data recorded by the ECAL is complemented by the adjacent detector, the *hadronic calorimeter* (HCAL).

3.4 Hadron calorimeter

The HCAL [31, 32] surrounds the ECAL. The primary role of the HCAL is to absorb and measure the energy of the large quantities of hadrons produced by proton-proton collisions. Unlike the ECAL (a homogeneous calorimeter), the HCAL is a *sampling calorimeter* with alternating layers of absorbing and scintillating materials. The majority of the HCAL volume is composed of brass absorber plates and plastic scintillator tiles, while the layers at large $|\eta|$ comprises of steel and quartz fiber. The volume comprises several cells, enabling the reconstruction of the spatial distribution of deposited energy.

Hadronic showers are created by particles interacting with the absorbing material within the HCAL. As a hadron moves through the absorbing material, it undergoes successive interactions with the atomic nuclei. The new particles produced during these initial interactions can themselves undergo further interactions, creating a cascade of particles (so-called shower).

When the particle showers reach the scintillating layer, light is created as a result of the excitation of the atoms of the detector material. As the excited electrons return to their lower energy levels, the atoms release energy in the form of photons, which are subsequently measured by photodetectors placed near the scintillating material. The portion of the shower's energy deposited in the absorbing layer cannot be measured directly. The total energy of the shower is therefore estimated indirectly using the information measured by the scintillating layer. It is also necessary to account for the energy deposited in the ECAL, as the hadronic shower is typically initiated in the ECAL as described in Section 3.3.

3.5 Solenoid magnet

A large *superconducting* solenoid magnet surrounds the main bulk of the HCAL. The solenoid magnet is a fundamental component of the CMS detector; bending the paths of particles based on their charge, thereby facilitating momentum measurement and particle identification.

The solenoid magnet generates a magnetic field of 3.8 T. Such a high magnetic field is achieved by employing superconducting technology. The solenoid magnet is constructed with niobium-titanium (NbTi) alloy wires, which are cooled to temperatures approaching absolute zero through the use of liquid helium. These low temperatures create a superconducting state within the magnet, allowing the magnet to conduct electricity without any resistance. This results is the generation of a strong and stable magnetic field.

3.6 Muon chambers

Muons play an essential role in many of the physics results of the CMS collaboration, including the discovery of the Higgs boson [33, 34]. In order to achieve precise muons measurements, the CMS experiment incorporates a dedicated muon detection system consisting of *muon chambers* [35, 36]. By combining information from the silicon tracker, the muon chambers and the magnetic field generated by the solenoid, the curvature of the muon tracks can be precisely determined and the muon's momentum calculated.

The muon chambers are placed furthest from the interaction point, beyond the reach of most SM particles. As the probability of EM interactions is influenced by the masses of the particles involved, lighter particles such as electrons and photons are absorbed by their interactions with the inner detector components. However, muons (which are about 200 times heavier than electrons [37]) experience fewer interactions, allowing them to pass through the inner detector without significant deflection or energy loss.

The detection principle of the muon systems relies on gaseous detectors. Such detectors commonly have lower resolution than solid state detectors (such as the silicon tracking detector) but can cover a larger area due to the comparably cheaper design. These types of detectors are filled with a gas which is ionised when a charged particle passes through it. There are four main types of muon chambers, each utilising different technologies to

cover different spatial regions within the detector.

Drift tube (DT) chambers are located at $|\eta| < 1.2$. The DT chambers utilise cylindrical drift tubes filled with an ionisable gas mixture. When a muon passes through the DT chambers, ionisation electrons are produced. These electrons drift towards a central wire, and measurement of the drift time allows the muon's position in the drift tube to be determined. *Cathode Strip Chambers* (CSC) are located at $0.9 < |\eta| < 2.4$ of the detector. CSC function in a similar way to DT chambers, but instead of a central wire contain cathode strips and anode wires for recording electron signals. Ionisation electrons are collected by the anode wires, inducing a charge on the cathode strips which provides the muon's position in the direction perpendicular to the anode wires. CSC have better spatial resolution than DT chambers, but worse time resolution. The *Resistive Plate Chambers* (RPC) are complementary detectors situated at the full range of $|\eta| < 2.4$. The RPC use gas-filled parallel plates constructed with a high-resistivity material which detects the ionisation caused by muons. This enables rapid triggering and improves the overall timing resolution. Finally, *Gas Electron Multiplier* (GEM) chambers are located adjacent to the CSC. The GEM chambers provided both a fast response and a good spatial resolution, complementing the CSC in a region of high particle flux.

The fast response time of the RPC and GEM chambers, combined with the tracking information from the CSC and DT chambers, play an important role in the selection of interesting events by the CMS trigger system. A reconstruction of the events are performed in order to facilitate the selection.

Chapter 4

Event reconstruction

Event reconstruction at the CMS experiment involves the interpretation of signals generated by particle collisions in the detector. For example, the trajectories of charged particles are reconstructed by the signals they leave in the layers of the silicon tracking system. The goal of the reconstruction process is to integrate the various signals produced by the sub-detectors to provide an accurate picture of what happened during the collision, including which particles traversed the detector. An overview of the trajectory reconstruction, together with other object reconstructions relevant to this thesis, is provided in the following text.

4.1 Trajectories of charged particles

The tracking algorithms employed by the CMS experiment [38] aim to precisely reconstruct the trajectories of charged particles. The reconstruction begins by converting the signals from the silicon tracking system into *hits* representing particle interactions with its various layers. These hits are then utilised in a three-step process involving (1) *track seed generation*, (2) *track finding* and (3) *track fitting*.

During track seed generation, potential track candidates are identified using subsets of hits. The subsequent track finding stage extends these candidates through the sub-detector layers, iteratively refining their parameters to best fit the observed hit positions. Finally, track fitting optimises the track parameters, such as position, direction, and momentum, by minimis-

ing the discrepancies between the predicted and measured hit positions.

4.2 Interaction vertex

The vertex algorithms employed by the CMS experiment [38] aim to identify the locations of *primary* and *secondary vertices* within a bunch crossing. The primary vertex is the location of the primary interaction, while secondary vertices appear from secondary interactions including pile-up and B-hadron decay. The first stage is to select high-quality tracks that are likely to be associated with the primary interaction. This involves the application of track quality criteria to filter out noise and badly reconstructed tracks.

Once the initial vertex seeds are found, a vertex fitting algorithm is employed. This algorithm iteratively refines the vertex positions and uncertainties by considering the selected tracks associated with each vertex candidate. After the fitting process, the vertices are ranked based on certain criteria relating to compatibility with the tracks, and the primary vertex is identified as that which is most likely to be associated with the primary interaction.

4.3 Particle flow algorithm

The *particle flow* (PF) algorithm [39] is foundational to many of the specific object reconstructions. It serves to identify and reconstruct individual stable charged and neutral hadrons, muon, electrons and photons from the signals left by particles traversing the detector (conceptualised in Fig. 4.1). By combining information from several sub-detectors within the CMS experiment, the precision and accuracy of the event reconstruction is optimised.

The PF algorithm reconstructs the trajectories of charged particles by grouping individual signals from the tracking system based on their spatial and kinematic compatibility. Once the trajectories of all charged particles have been reconstructed, vertices are created by analysing the intersection of multiple trajectories.

The PF algorithm combines the energy deposits from the ECAL and the HCAL to estimate the particles' energies. Trajectories from the tracking

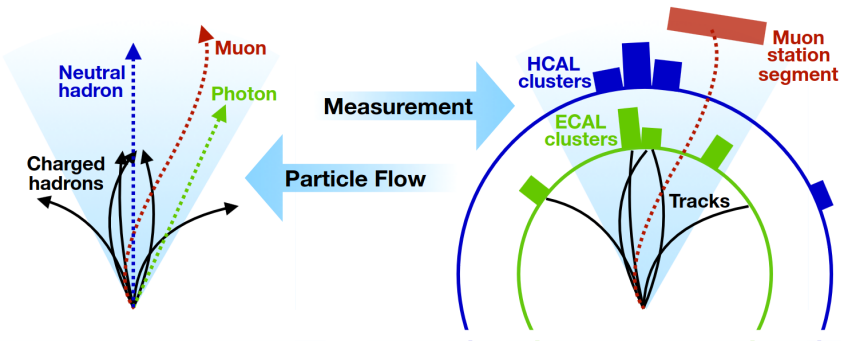


Figure 4.1: Illustration of the particle flow algorithm that combines information from all sub-detectors in order to obtain a global view of the event. Figure taken from Ref. [21].

system are extrapolated to the calorimeters in order to facilitate the identification of EM and hadronic particles. The muons are reconstructed by building tracks from the hits in the DT and CSC sub-detectors. A final muon track is constructed by matching muon system tracks to tracks reconstructed from hits in the silicon tracking system. The muon momentum is determined based on the curvature of this track. A more detailed description of the PF algorithm can be found in Ref. [39].

4.4 Jets

Due to colour confinement, quarks and gluons cannot exist freely. Instead, they form colour-neutral hadrons in a process referred to as hadronisation. Owing to colour confinement, detection of final-state quarks and gluons is not possible. Instead, event properties that have a close correspondence with their distributions are studied. These event properties are known as *jets*. An illustration of a *dijet* event (two highly-energetic jets created in nearly opposite directions) is displayed in Fig. 4.2.

The identification and reconstruction of jets is an important part of many physics analyses, including those described in this thesis. These analyses commonly focus on the jet with the highest p_{T} in the event, the so-called *leading jet*. The jet with the next highest p_{T} is referred to as the *sub-leading jet*.

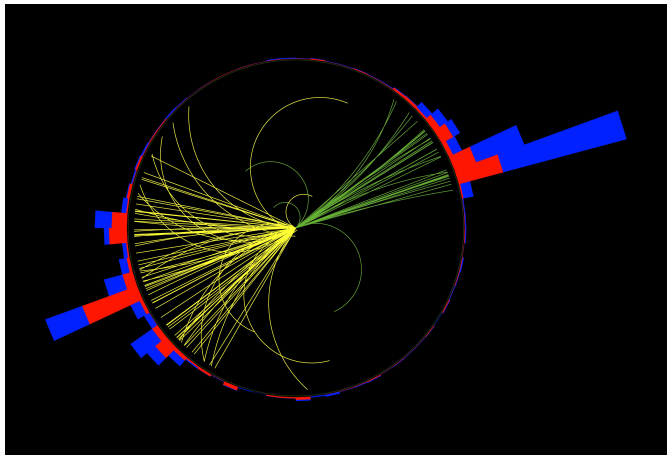


Figure 4.2: An event display of the creation of two jets. The trajectories of the particles constituting the jets are illustrated in yellow and green, while their energies deposited in the ECAL and HCAL are illustrated in red and blue, respectively. Figure taken from Ref. [40].

To create a jet, the final-state particles are clustered according to best estimates of the initial-state quark or gluon from which they originated. Jet creation functions to reduce the complexity of the final state by simplifying many hadrons to a simple object. The definition of a jet is dependent of the jet *clustering algorithm*. At the CMS experiment jets are clustered using the anti-k_T [41, 42] clustering algorithm where the distance parameter used is 0.4 for small-radius jets (AK4 jets) and 0.8 for large-radius jets (AK8 jets). Tuning of the distance parameter allows particle decays of different types to be efficiently captured by jets.

4.4.1 Jet distance parameter

A smaller distance parameter requires the particles to be closer to each other in order to be considered part of the same jet. Small distance parameters are particularly useful when aiming to capture intricate details of individual particles. Conversely, larger distance parameters allow particles to be farther apart and still be included in the same jet, and are beneficial for capturing the overall structure of the event.

When a heavy particle decays into lighter particles, the lighter particles receive a *boost* of energy, resulting in collimated decay products. The use of a

large distance parameter during jet reconstruction enhances the efficiency of reconstructing heavy particles with significant boosts as single jets. This is due to the inverse proportionality between the average angular distance between decay products and the p_{T} of the decaying particle [43], as expressed by

$$R_{\mathrm{qq}} \approx 2 \times \frac{\text{Mass of the decaying particle}}{p_{\mathrm{T}} \text{ of the decaying particle}}. \tag{4.1}$$

Consequently, at sufficiently large boosts (approximately $p_{\mathrm{T}} > 200\,\mathrm{GeV}$), the final state hadrons from the decay of particles (such as the Z, $W^{\pm}$ and Higgs bosons) merge into a jet that is more efficiently reconstructed with a larger distance parameter. Fig. 4.3 illustrates this efficiency, showcasing the reconstruction of W bosons into a single jet as a function of the W boson p_{T}. Notably, the efficiency increases with p_{T} when using large-radius distance parameters ($R = 0.8$), whereas the opposite trend is observed for small-radius distance parameters ($R = 0.5$). The choice of 0.5, instead of 0.4, for the small-radius distance parameter aligns with its widespread use in CMS publications at the time of the study. This does not affect the overall conclusion and the same trend is observed when comparing $R = 0.8$ and $R = 0.4$.

4.4.2 Jet type

The inputs to the clustering algorithm can be the four-momentum vectors of calorimeter energy deposits or of PF reconstructed particles, and result in a *calorimeter jet* or a *PF jet*, respectively.

Calorimeter jet

Calorimeter jets are reconstructed from energy deposits in the calorimeter towers. A calorimeter tower consists of one or more HCAL cells and the geometrically corresponding ECAL crystals. In this process, the contribution from each calorimeter tower is assigned a momentum; the absolute value and direction of which is given by the energy measured in the tower, and the coordinates of the tower. The jet energy is obtained from the sum of the tower energies, and the jet momentum by the vectorial sum of the tower momenta. The jet energies are then corrected to establish a relative

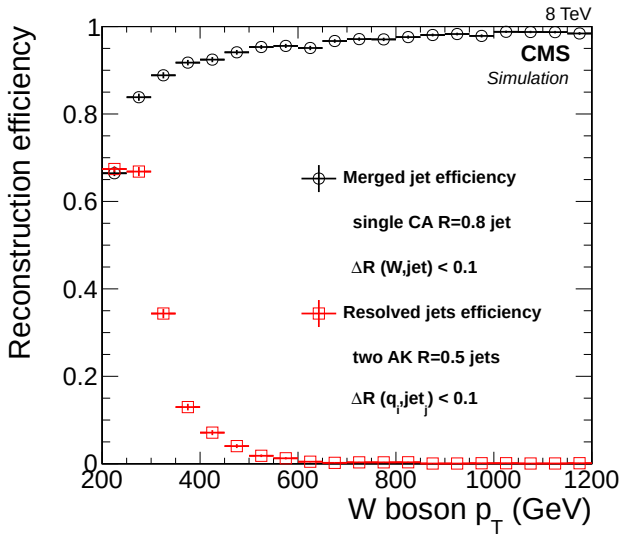


Figure 4.3: Efficiency of W boson reconstruction as a function of their p_{T} using the distance parameter 0.8 (black data points) and 0.5 (red data points). Figure taken from Ref. [43].

uniform response of the calorimeter in η and a calibrated absolute response in p_{T}.

PF jet

PF jets are reconstructed by clustering the four-momentum vectors of PF candidates. The jet momentum is determined as the vectorial sum of all the particle momenta in the jet. The clustering of PF jets operates iteratively in the following steps.

First, the algorithm starts by choosing a seed particle i at random. The distance between adjacent particle j and seed particle i is computed as

$$\Delta_{ij}^2 = \left(y_i - y_j\right)^2 + \left(\phi_i - \phi_j\right)^2, \tag{4.2}$$

where y is the rapidity and ϕ the azimuthal angle. Next, a variable reflecting the computed distance and the p_{T} is calculated as

$$d_{ij} = min\left(p_{T,i}^{-2}, p_{T,j}^{-2}\right)\frac{\Delta_{ij}^2}{R^2}. \tag{4.3}$$

The distance parameter R defines the width of the jet. Next, the clustering algorithm combines particles iteratively by always choosing the particle j that minimises d_{ij}, while requiring $d_{ij} < p_{T,i}^{-2}$. The iteration stops when all particles satisfying this relation have been clustered into the jet. Hence, the parameter R can also be regarded as a cut-off parameter. The algorithm then continues by choosing a different seed particle i for the next jet.

4.4.3 Pile-up mitigation

Pile-up can contribute additional tracks and calorimetric energy depositions to the reconstructed jet momentum. To mitigate this effect, the PF jets are often subject to pile-up mitigation methods such as the Charged-Hadron Subtraction (CHS) or the Pileup Per Particle Identification (PUPPI) [44, 45] algorithm.

In CHS, the CMS tracking system is used to identify charged particles originating from pile-up and they are subsequently discarded. PUPPI builds upon the CHS algorithm and applies a more rigorous selection to charged and neutral particles according to their probability of originating from the primary interaction. This is possible because neutral particles from the primary interaction are typically aligned with charged particles originating from the same interaction, while neutral particle from pile-up are more uniformly distributed in all directions.

4.5 Missing transverse energy

The presence of neutrinos in the event is accounted for by the reconstruction of *missing transverse energy* (MET). The MET is calculated by summing the transverse momentum of all particles detected by the CMS detector. If all particles in an event were detected and accounted for, the sum of their transverse momenta should be zero since the momentum is conserved in the transverse plane. However, if there are undetected particles like neutrinos, the net transverse momentum will not be zero. The MET is an essential component of many physics analyses, including searches for new particles

that may escape detection and manifest themselves as an imbalance in the event's total momentum.

4.6 Type of event reconstruction

Various event reconstructions exists for different scenarios and tasks. At the CMS experiment, the *offline reconstruction* is achieved when collision events are reconstructed hours (and sometimes even months or years) after data collection. Offline reconstruction requires the complete detector output, referred to as *raw* data. As current understanding of detector conditions is continuously improving, reconstructing after data collection often result in a better event reconstruction. As a result, the *standard analysis strategy* at the CMS experiment is based on the analysis of offline reconstructed physics objects.

In contrast, the *online reconstruction* takes place during data collection and is traditionally used by the experiment when selecting collision events to store for subsequent analysis. The online reconstruction is optimised for speed more than for high resolution or other similar performance metrics. To better understand the online reconstruction, it is necessary to discuss the CMS trigger system.

Chapter 5

Trigger system

The CMS trigger system is tasked with rapidly filtering the vast amounts of data produced by the detector in order to allow events of interest to be recorded for further analysis. Every second, approximately 40 million bunch crossings take place in the CMS detector. As the storage required per collision event is approximately 1 MB [5], the detector would have to process and save 40 TB per second if all collisions were to be recorded. Such a rate of recording is beyond existing detector capabilities. The trigger system therefore functions to reduce the rate by 99.925 –99.995%, allowing the CMS detector to record a more manageable number of events with such an event size. To achieve this reduction, every event is required to pass criteria based on the properties of the physics objects set by a two-tiered trigger system in order to be recorded by the CMS experiment.

This criteria determines the type of physics processes recorded by the CMS experiment. As the instantaneous luminosity increases with HL-LHC (Section 2.2.3), these constraints are only expected to gets stronger in order to discard the growing amount of QCD produced jets. While these are necessary to reduce the volume of collision data recorded, they limit the physics reach of the experiment. Once an event is discarded by the trigger system, it is permanently lost and the information it contains cannot be recovered. It is therefore highly important to design a reliable trigger system and to ensure its effective daily operation.

A CMS trigger system comprising two tiers was necessitated by the high rate of collisions. The first tier functions to achieve the majority of the reduction and must therefore be simple and robust in order to manage such

a high input. The second tier has a much smaller input and is thereby able to apply more sophisticated selection criteria to minimise the likelihood of discarding interesting physics events. The two tiers are the Level-1 trigger (L1T) [5] and the High-Level trigger (HLT) [6]. Each are now discussed in detail.

5.1 Level-1 trigger

The purpose of the L1T is to reduce the event rate to a level that the software-based HLT can process. When a particle created by a collision traverses the detector, it interacts with the various sub-detectors, creating signals that allow its passage to be recorded. The data from these signals are compressed and zero suppressed by the data acquisition (DAQ) system. As previously mentioned, current software implementations are not capable of processing a bandwidth of such an event rate and size. The L1T is therefore built from customised hardware processing units (so-called *Field-Programmable Gate Arrays* or FPGAs). These hardware units allow the L1T to operate at a latency of about 3.8 μs [5]. Within this time limit, the L1T creates an approximate reconstruction of the collision event (Section 5.1.1), which is then used to decide if the event is discarded or retained for further processing by the HLT (Section 5.1.2).

5.1.1 Event reconstruction

As discussed in Section 4.6, the event reconstruction performed by the CMS trigger system is referred to as the online reconstruction. The L1T does not have the capacity to create an online reconstruction of the full event information of 40 million events per second. To circumvent this limitation, the FPGAs receive input from only the CMS calorimeters and the muon chambers. To further facilitate low latency, the L1T receives this data in the form of *trigger primitives* (TPs, a reduced version of the full information). TPs are basic detector-level measurements that provide information about only specific physics objects. Calorimeter TPs include energy deposits, energy sums and patterns of energy distributions from ECAL and HCAL, while muon TPs include momentum measurements and patterns of hits in the muon detectors.

The information provided by the TPs is fed to the *calorimeter* and *muon*

triggers. The calorimeter trigger combines the ECAL and HCAL TPs into single *trigger towers* (TTs). L1T jets are then reconstructed out of the TTs as described in Section 4.4.2. The calorimeter and muon triggers process the TPs information and sends the results to the *global trigger*, that makes the final decision on whether to accept or reject the event. A more detailed presentation of the L1T architecture is provided in Ref. [5].

5.1.2 Event selection

The L1T selects events through the application of a *trigger menu*, which is a combination of several algorithms programmed in the FPGAs. These algorithms are referred to as *seeds*. The seeds are programmed to accept an event if it satisfies predefined criteria, such as a the p_{T} of a L1T jet exceeding a certain threshold. An event is accepted for further processing by the HLT if it passes at least one seed within the trigger menu, unless it is vetoed by the application of a *prescale*. A prescale determines the fraction of events selected from all the events accepted by a certain seed. A prescale of N means that for every N number of events accepted by a seed, 1 event is selected for further processing. The rate of events passing the seed can therefore be reduced by applying a prescale. An *unprescaled* seed has a prescale equal to 1 and therefore selects every event that passes the seed. In contrast, a prescale of 0 means that the seed is disabled.

A *prescale column* is a set of specific prescales applied to every seed within the trigger menu. Several different prescale columns are used with the same trigger menu. As the condition of the collisions change, the prescale column applied to the menu may be changed in order to select a different set of prescales for each seed. For example, as instantaneous luminosity decreases with time (Section 2.2), the target output rate can be maintained by changing to a column with smaller prescales. This is displayed in Fig. 5.1, where the L1T output rate decreases as a function of time and increases after changing a prescale column.

Each prescale column tunes the passing rate of each seed in order to produce a total L1T output rate of about 100 kHz. Studies of the L1T operation have concluded that approximately 100 kHz is the maximal rate before disrupting the normal functioning of the DAQ system [5]. Collision data is buffered locally while the L1T makes the decision whether to accept the event for further processing by the HLT. The buffer size is limited and starts to overflow if the decision takes too long, which in turn leads to an uncon-

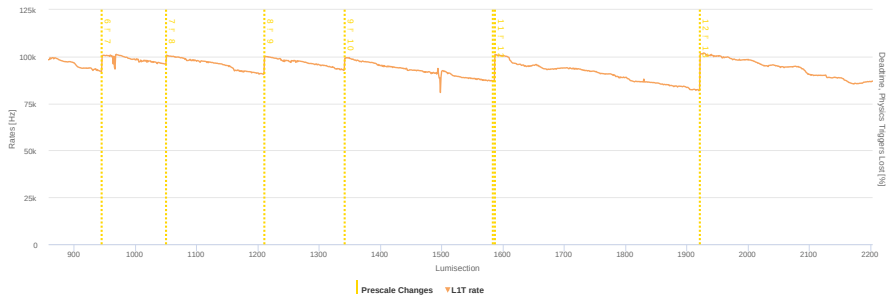


Figure 5.1: The effect of prescale changes on the L1T rate, displayed by the CMS Online Monitoring [14] tool. As instantaneous luminosity decreases with time, the total L1T rate also decreases (orange curve). Each change of prescale column (yellow line) increases the rate to the target rate of approximately 100 kHz. The plot is displayed as a function of lumisections (time interval lasting approximately 23 seconds) for LHC Fill 8321 Run 361303 at 29th of October 2022.

trolled loss of data known as *dead time*. Dead time occurs when the detector is busy processing previous data, and is therefore unable to record any new information. Dead time may be caused by the trigger rate being too high or due to technical issues such as hardware malfunctions or software errors. Figure 5.2 displays the total dead time of a typical run and its effect on the L1T output rate, which is around 2–3% for a typical run. By limiting the output rate, the CMS can manage the data volume and maintain its real-time decision-making capabilities whilst avoiding dead time.

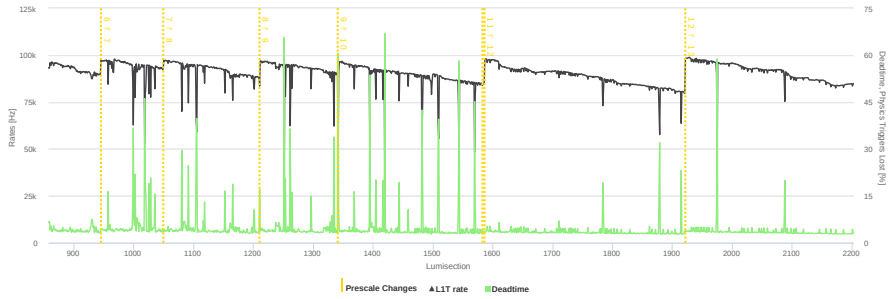


Figure 5.2: The dead time (green) together with the L1T output rate after taking dead time into account (black), displayed by the CMS Online Monitoring [14] tool. The plot is displayed as a function of lumisections for LHC Fill 8321 Run 361303 at 29th of October 2022.

5.1.3 Trigger efficiency

The seeds are designed to achieve maximal *trigger efficiency*, defined as the fraction of events accepted by the seed out of all of the events targeted. For example, the seed `L1_SingleJet180` targets event based on the existence of a L1T jet with p_{T} exceeding 180 GeV. However, as can be seen in Fig. 5.3, the seed is only fully efficient from approximately 300 GeV.

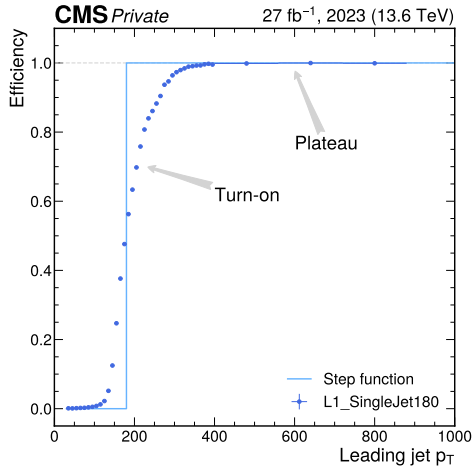


Figure 5.3: Trigger efficiency curve for the `L1_SingleJet180` seed as a function of leading jet p_{T} (dark blue points), and its comparison with a step function at 180 GeV (light blue line). Turn-on and plateau are annotated.

The threshold value of 180 GeV sits in the middle of the so-called *turn-on* of the efficiency curve. While some events with a jet with $p_{\mathrm{T}} > 180$ GeV are correctly identified and selected by the seed, not all are. Ideally, the seed would estimate the p_{T} of every jet correctly, providing an efficiency curve that is a step function; jets with p_{T} below 180 GeV are never selected while jets above the threshold are always selected. However, in reality the limited p_{T} resolution can occasionally cause jets with too low or too high p_{T} to be accepted or rejected, respectively. A desired property of the efficiency is to have a sharp turn-on, where the efficiency rapidly rises from zero to its maximal value. The sharper the turn-on, the better the seed is able to correctly estimate the jet p_{T}. Beyond a certain point the seed will always estimate the p_{T} to be above the threshold, at which point the trigger efficiency curve reaches its *plateau*.

There are a number of methods that can be used to estimate the trigger

efficiency. Two of the most commonly used approaches at the CMS experiment are the *reference* and the *tag-and-probe* methods. In the reference method, the efficiency of the trigger of interest (the so-called signal trigger) is computed with the help of a reference trigger. If the reference trigger is selected so that it fires independently of the signal trigger, the likelihood of both triggers firing is defined as

$$P(\text{S and R}) = P(\text{S}) \times P(\text{R}), \tag{5.1}$$

where S and R refer to the signal and reference trigger firing respectively.

Taking Eq. 5.1 into account, as the efficiency of an event passing a seed is defined by the fraction of events that passed the seed (N_{pass}) out of all events considered (N), the efficiency of a certain seed can be computed as

$$\epsilon(\text{S}) = \frac{N_{\text{pass}}^{\text{S and R}}/N}{N_{\text{pass}}^{\text{R}}/N}, \tag{5.2}$$

$$= \frac{N_{\text{pass}}^{\text{S and R}}}{N_{\text{pass}}^{\text{R}}}. \tag{5.3}$$

In contrast, the tag-and-probe method does not require a reference trigger. The method involves selecting a tag and a probe object, where the tag object fires the signal trigger independently of the probe. As the properties of two objects are independent of each other, the probability that one object causes the trigger to fire does not depend on the probability that the other object also fires the trigger. Eq. 5.1 can therefore be used to defined the trigger efficiency as

$$\epsilon(\text{probe}) = \frac{N_{\text{pass}}^{\text{probe and tag}}}{N_{\text{pass}}^{\text{tag}}}. \tag{5.4}$$

The two methods of computing the trigger efficiency are complementary, and can be used in parallel to achieve the same result. In practice, the reference method is often used when studying the efficiency of seeds selecting events based on hadronic activity. On the other hand, the tag-and-probe

method is preferred for processes that have distinct signatures such as the study of the efficiency of lepton based seeds.

Studying the trigger efficiency is an important part of the development and assessment of L1T seed. It is similarly used by the HLT to study the performance of the HLT triggers.

5.2 High-Level trigger

If the L1T accepts an event (Section 5.1.2), the full detector readout is sent for further processing by the HLT. The maximal input rate from the L1T is approximately 100 kHz, which is reduced by around 75–98% by the HLT. There are four primary reasons to further reduce the rate:

1. **The latency of the trigger decision**. In order to avoid a dead time (and risk of losing data) the HLT must swiftly process arriving events. The average processing time of the HLT can be up to 500 ms per event in Run–3 [46].
2. **The finite bandwidth of the DAQ system**. Restrictions on the data volume are imposed both by the size of the temporary data storage at the site of the CMS experiment, and by the bandwidth of the link (10 Gbps [47]) connecting the site and the CMS computing center 10 kilometres away at the main CERN site.
3. **The time pressure of reconstructing the recorded collision data**. In order to facilitate data quality monitoring and prompt calibration procedures, it is necessary to complete all data reconstruction within 48 hours of collection (referred to as *prompt* reconstruction).
4. **The finite space for permanent data storage**. Budgeting constraints, relating to the cost of purchasing tape and disk storage, limit permanent data storage capacity.

As a consequence of this event reduction, the HLT and therefore the CMS experiment nominally records events at a rate of approximately 2 kHz — around 0.005% of the collisions that occur every second in the CMS detector. Mirroring the previous section considering the L1T, the event reconstruction (Section 5.2.1) and the event selection (Section 5.2.2) by the HLT is now described.

5.2.1 Event reconstruction

The HLT runs on a computing farm consisting of *central processing units* (CPU) and *graphical processing units* (GPU). The latter is a new addition to Run–3, facilitating the acceleration of specific algorithms relating to the online reconstruction of tracks and vertices of the pixel detector and local reconstructions of the ECAL and HCAL. A comparison between the event processing time when CPU is utilised compared to when part of the reconstruction is offloaded to GPU is presented in Fig. 5.4. The upgrade resulted in a significant decrease in processing time, exemplified by the pixel track reconstruction increasing the number of events processed per second by 3 times [48]. Achieving the same type of performance using only CPUs would have increased the cost (by roughly 15%) and power consumption (by about 30%).

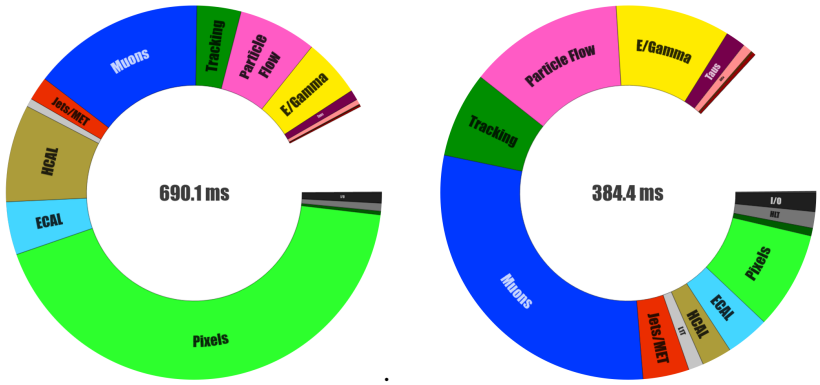


Figure 5.4: Pie chart distributions of the event processing time for the HLT reconstruction running only on CPU (left) and offloading part of the reconstruction to GPU (right). The slices represent the time spent on different object reconstructions; a clear decrease is visible for right with respect to left in the pixel and calorimeter reconstructions. The average processing time per event is displayed in the middle of the chart. Figure taken from Ref. [46].

The farm is subdivided into groups of *processing nodes*, each hosting a pair of *Builder Units* (BU) and *Filter Units* (FU). For events passing the L1T, the DAQ system triggers the transfer of the full detector information to RAM-disks. This processes is facilitated by the BUs. The BUs write custom binary file formats (each containing around 100 collision events) with a header prepended to each event providing information for subsequent event identification. These files are then distributed to FUs. When a FU

receives data from a BU, the data is unpacked in order to reconstruct and filter the events. These actions are performed using a software framework called *CMS SoftWare* (CMSSW). Several independent online reconstruction and filtering processes are executed in parallel, decreasing the latency of the HLT.

Unlike the L1T, the HLT uses information from all sub-detectors during the online reconstruction. A foundation to many of the specific object reconstructions is the PF algorithm discussed in Section 4.3. To meet time constraint, a simplified algorithm is used. The tracking is reduced to three iterations and the time-consuming reconstruction of tracks with low p_{T} or arising from nuclear interactions in the tracker material is dropped. In addition, the electron identification and reconstruction is not included. These modifications lead to a slightly higher jet energy scale for PF jets featuring an electron or a nuclear interaction [39].

5.2.2 Event selection

Similarly to the L1T, the HLT select events through the application of a trigger menu. The trigger menu consists of HLT *paths*; sets of processing steps run in a predefined order. As illustrated in Fig. 5.5, each path contains several modules that both reconstruct physics objects and make selections based on predefined criteria of these objects. Each path has a specific trigger condition based on a set of L1T seeds, which is required to have been accepted before processing of the path begins. For example, a path that reconstructs muons and selects event based on the existence of a muon with p_{T} exceeding 50 GeV can be programmed to only process the event if the event passed a L1T seed requiring a muon with p_{T} exceeding 25 GeV. This reduces the number of times a computationally expensive HLT path is run without the existence of the physics object of interest for that path.

In order to further reduce unnecessary processing, each path is implemented as a sequence of steps where the computationally expensive steps (such as intensive track reconstruction) is performed later in the sequence. An event may be rejected at each step of the path — by processing the computationally cheaper steps first, the computationally expensive steps are only applied when necessary.

Another similarity with the L1T is the application of prescales. Like a L1T seed, a HLT path can have a prescale applied in order to reduce the selected

events of that path. A HLT prescale table sets the prescales of every path as part of the HLT trigger menu.

HLT paths selecting similar physics objects are grouped into *datasets*, with collections of datasets being organised into *streams*. The grouping is illustrated in Fig. 5.5. This is done to facilitate efficient handling of the collision data. By dividing the data into several streams, the processing workload can be efficiently distributed among different computing clusters. In addition, the division allows the CMS experiment to apply initial event selections to events of interest for different physics signatures. For example, some streams are designed to capture events involving specific final states, such as events containing one or more leptons.

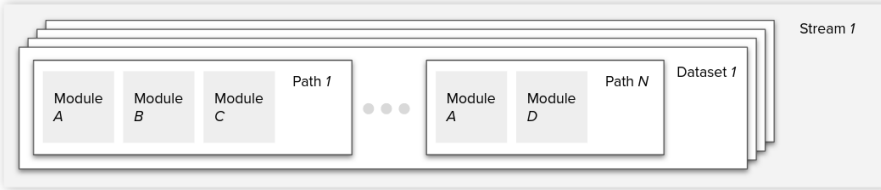


Figure 5.5: Illustration of the grouping of HLT paths into datasets, and datasets into streams. A HLT path is made up of several modules, each executing a processing step involving event reconstruction and output selection.

When an event passes the HLT selection, the full detector readout of the event is sent in a raw data format to be stored on temporary local disks. The raw data is later transferred to a long-term storage facility. By storing the raw detector output, events may be reconstructed several times after data collection — a process referred to as offline reconstruction (discussed in Section 4.6). The offline reconstruction is performed with the same software framework as used online (CMSSW), and like the online reconstruction, the PF algorithms is utilised for the reconstruction of many physics objects.

Trigger study streams

Trigger studies are achieved by recording specific events with the application of two streams; the `ZeroBias` and the `HLTPhysics` stream. The paths of the `ZeroBias` stream process events passing the `L1_ZeroBias` seed with-

out applying any further criteria on the events. The seed selects every event that is not vetoed by its prescale. As a result, the rate depends solely on the number of bunch crossings in the LHC. The stream may therefore be used to study the behaviour of newly developed L1T seeds, without the need to account for a selection bias that would be present if criteria based on the physics objects were applied.

The paths of the `HLTPhysics` stream process events passing any L1T seed without applying any further criteria. As a result, the `HLTPhysics` stream is biased by the L1T selection, but not the HLT selection. The stream may therefore be used to study newly developed HLT paths and datasets given the current L1T trigger menu. For example, the event selection of a dataset may be updated to cover a certain part of phase space through the inclusion of new paths. The `HLTPhysics` stream may then be used to study the effect of the inclusion on the rate and reconstruction time of the dataset. Any modification to the HLT requires careful analysis of the rate and timing in order to avoid harmful impact on the operation of the trigger system. The stream is also used to study the efficiency and performance of the HLT algorithms under varying LHC conditions, such as high and low pile-up. The results of these studies are used to validate and optimise the operations of the HLT.

5.2.3 Operating and monitoring

Various tools and techniques are used to continuously monitor the HLT in order to to ensure its proper functioning and performance. During operation, the CMS trigger system is monitored by a *trigger shifter*. The trigger shifter is a member of the shift crew whose job is to ensure efficient data-taking by the CMS experiment. Real-time monitoring displays are used to visualise and assess the performance of the trigger system, and provide an overview of key metrics such as trigger rates and event processing times. This data allows the shift crew to swiftly identify and rectify any deviations from the expected behaviour of the L1T and HLT.

By monitoring the trigger rates, the trigger shifter ensures that the HLT is selecting events at a rate designed to meet the desired physics goals while not overwhelming the available computing resources. If needed, the trigger shifter may change the prescale column to maintain an appropriate rate.

The trigger shifter is accompanied by the L1T and the HLT *Detector-On-Call* (DOC) shifters. The L1T and HLT DOCs are assigned their positions for two weeks at a time, during which they are responsible for the deployment of all trigger configurations and interact directly with *Run Coordination* and different subsystem experts in order to resolve any arising issues and ensure efficient data collection.

The HLT DOC is responsible for the preparation and testing of HLT trigger menus for special LHC runs, management of L1T and HLT prescales, monitoring of HLT rates and online farm operation, as well as daily and weekly data certification. The HLT DOC maintains comprehensive documentation of the trigger activities, including observations, interventions, and modifications made to the trigger system's configuration. These reports serve as a valuable resource for future reference, and a summary is presented at the *Daily Run meeting* to inform Run Coordination and subsystem experts on the intended operations of the HLT team.

Chapter 6

Data scouting

Data scouting (referred to as "scouting" throughout the rest of this thesis) is a strategy currently used for data analysis at the CMS experiment. Based on trigger-level reconstruction, scouting complements the standard strategy for analysing physics objects that have been reconstructed offline. Unlike the standard approach, which stores the raw detector output, the scouting strategy stores events that have been reconstructed online by the HLT directly to disk. The event size of trigger-level objects are much smaller than the raw detector output, placing a smaller load on the DAQ system.

As discussed in Section 5.2, the bandwidth limitations of the online trigger system impose four primary constraints on the number of events that can be recorded using the standard analysis strategy. By storing the trigger-level reconstruction rather than the raw detector output, the scouting strategy lessens the impact of three of the four main constraints:

1. **The finite bandwidth of the DAQ system**. The size of an event stored as raw data by the standard analysis strategy is of the order of 1 MB. In contrast, the event size of online reconstructed events used by scouting is only about 8 kB. As the bandwidth of the DAQ system imposes limitations on the event size and rate, a smaller event size alleviates this bandwidth constraint on data collection.
2. **The time pressure of reconstructing the recorded collision data**. Scouting events are only reconstructed once (online), removing the need for a prompt reconstruction.
3. **The finite space for permanent data storage**. Due to the smaller

event size, the scouting dataset places a smaller demand on finite storage resources. If the same event is stored by both the scouting and standard trigger strategies, the scouting reconstructed objects account for only approximately 0.8% (8 kB / [8 kB + 1 MB] × 100%) of the storage resources used.

Facilitated by the reduced impact on the DAQ system, the scouting technique increases the rate of events passing the HLT by applying lower trigger thresholds. For example, the scouting strategy achieved an average rate of approximately 22 kHz in 2022, while the standard trigger approach achieved approximately 2 kHz. Accounting for the higher rate of scouting, the scouting reconstructed objects amounts for approximately 8% ([8 kB × 22 kHz] / [8 kB × 22 kHz + 1 MB × 2 kH)] × 100%) of the storage resources used by both the scouting and standard trigger strategies.

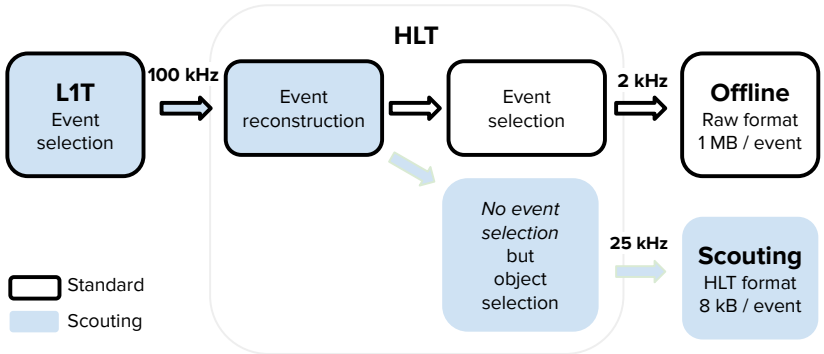


Figure 6.1: Illustration of scouting (blue) and standard trigger (black rim) strategies. The rate following each selection is displayed according to their average values in 2022.

The less stringent selection process of scouting allows physicists to detect, analyse and explore events that would not previously have been allowed to pass the HLT. This provides opportunities for analysis outside the boundaries of the standard trigger strategy, and exploration of previously unexplored regions of phase space.

While scouting enables a wider range of analyses, the approach is limited

by its inability to store the full event information. Without the raw detector outputs, the event can not be reconstructed after collection, when a better understanding of the detector conditions generally allow for an improved reconstruction. However, this limitation does not always have a significant impact on the sensitivity of the analysis. The negative effects are also lessened by scouting's facilitation of early analyses. Due to the increased number of events available with the scouting strategy, analyses may be performed with the scouting reconstructed objects before sufficient events have been collected with the standard trigger strategy to facilitate analysis. If promising results are observed, the analysis can then be targeted later with the standard strategy.

In this chapter, the scouting strategy used at the CMS experiment is described in detail. Firstly, the motivation for the strategy is presented in Section 6.1. Next, an overview of the strategy in Run–1 and Run–2 is presented (Section 6.2), followed by a detailed description of the strategy in Run–3. This includes descriptions of the scouting event content and rates (Section 6.3.1), streams and datasets (Section 6.3.2) as well as trigger efficiencies (Section 6.3.3) in 2022 and 2023. Due to the focus of this thesis, the descriptions centre around the selection of events targeting hadronic activity.

6.1 Physics motivation

The inability of the SM to address existing physics problems provides strong motivation for experimental searches for new physics. Examples of existing problems are the large gap between the gravitational and electroweak energy scales and the lack of explanation for astronomical observations indicating the existence of dark matter. Despite the broad physics programme of investigation performed by the CMS experiment, no signs of new physics beyond the SM have been observed in the last decade.

It is possible that the lack of new discoveries can be explained by new physics existing beyond the current reach of the LHC. For example, if the LHC was capable of colliding protons at higher energies, new interactions may be detectable. However, the CMS experiment has long been aware of the possibility that new physics is in fact observable at the current collision energy of the LHC. This may include the detection of new particles that are light, feebly-coupled or hidden behind large SM backgrounds. Events in-

cluding such particles might be rejected by the standard trigger protocols, due to their kinematic properties being below nominal trigger thresholds. In order to study these type of events, it is necessary to extend the searches for new physics to unexplored phase space regions. This can be achieved by scouting.

The scouting strategy enables the CMS collaboration to embark on pioneering searches for low-mass resonance, with a particular advantage for searches involving jets. As the majority of proton-proton collisions result in relatively uninteresting low-energy jet production from quark and gluon interactions, the standard trigger strategy must adhere to stringent energy and momentum thresholds to prevent overwhelming the DAQ protocols. In contrast, the scouting strategy offers a notable reduction in these thresholds, providing a more flexible approach and allowing searches for interesting but rare physics processes involving low-energy jets.

6.2 Run–1 and Run–2

The scouting strategy was designed and tested for the first time at the end of 2011, recording a total of 0.13 fb^{-1} integrated luminosity. The strategy initially focused on PF jets, and selected events based on the presence of scalar sum of jet p_{T} (denoted as H_{T}) exceeding 350 GeV. The collision data were used to perform a search for heavy resonances decaying to dijets, and demonstrated sensitivity to resonances with masses between 0.6 and 0.9 TeV [49]. This was an energy scale inaccessible by the then standard trigger approach.

Following the successful trial in 2011, the strategy was revised and implemented for all data collection during 2012. The H_{T} selection was lowered to 250 GeV, however, instead of reconstructing computationally expensive PF jets, calorimeter jets were stored. The collision data, corresponding to 18.8 fb^{-1}, were used to perform a dijet resonance search analogous to that performed during 2011. The search results were interpreted as limits on the mass and coupling of a hypothetical leptophobic Z' resonance decaying to quarks. The limits were the strongest yet obtained for masses between 0.5 and 0.8 TeV [50], improving on the results of previous experiments.

The success of the scouting technique in Run–1 prompted an expansion of the strategy for Run–2. The aim was to maintain the ability to search for low-energy jets, while also providing an event format capable of support-

ing a broader range of scouting analyses. To facilitate this, three streams were deployed; one saving an event content based on calorimeter jets, one saving an event content based on PF jets and one saving event content based on a pair of PF muons.

Data recorded by the calorimeter jet stream in Run–2 was used to perform a search for dijet resonances with masses between 0.6 and 1.6 TeV [51]. The results of the search place new limits on resonances part of several beyond the SM models. A search for a narrow resonance decaying to a pair of muons was performed for masses between 11.5 and 45.0 GeV using the PF muon stream [52]. The results of this search are interpreted in the context of a dark photon, and sets strong constraints on dark photon mass and mixing.

6.3 Run–3 (2022–2023)

The exploration and development of the CMS scouting technique during Run-1 and Run-2 highlighted its value as an innovative trigger strategy and as a successful paradigm for data analysis. During this period, the primary constraint in implementing the scouting strategy was found to be the HLT event processing time. As a consequence, scouting benefited greatly from the online hardware upgrade of Run-3, described in Section 5.2.1. Following the hardware upgrade and increased usage of GPUs, the HLT algorithms were redesigned to harness the capabilities of parallel architectures. As a result, during 2021 and 2022 a new GPU-based approach was developed and fully commissioned for the calorimeter reconstruction, pixel local reconstruction and pixel-based tracking.

In order to take full advantage of the GPU upgrade, a modified version of the PF algorithm was utilised by scouting. The modified version uses pixel tracks instead of tracks reconstructed from both pixel and strip tracker hits (described in Section 4.1, used by scouting in Run–2 and the standard trigger strategy in Run–3) as input. The main advantage of the so-called *pixel-only* tracking is the option to offload the track reconstruction to GPUs, thereby notably accelerating event processing at the cost of a slightly worse track resolution compared to standard tracks (especially evident in high p_{T} tracks where the degradation is more significant) [53]. As low p_{T} tracks are most relevant to scouting, this acceleration particularly benefits the scouting strategy. Moreover, the acceleration of processing time al-

lowed scouting to include dedicated reconstructions of electrons and photons into the event content (as discussed in Section 6.2 the focus of the scouting strategy was jets and muons in Run–1 and Run–2).

As a result of both the reduced HLT processing time and the increased resources allocated to the scouting strategy, the scouting rate increased significantly (by about 5 times) from Run–2 to Run–3. This improvement is illustrated in Fig. 6.2, where the average rate of scouting in Run–1 to Run–3 is compared with the average rate of the standard trigger strategy. The decrease in the scouting rate between 2022 and 2023 was a result of additional constraints on the event selection based on photon and electron objects. The constraints were added following dedicated studies performed on scouting data recorded in 2022 and enhance the efficiency for low-p_{T} objects, as well as improve the usage of HLT resources.

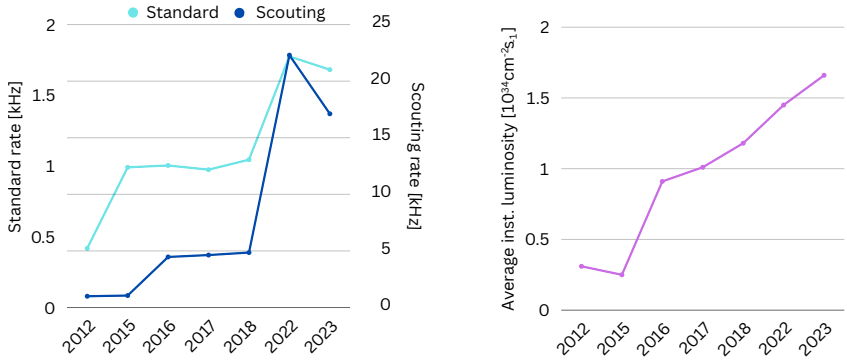


Figure 6.2: The average rate of the standard trigger and scouting streams as a function of time (left). The standard and scouting rates are displayed on the left and right y-axis, respectively. The instantaneous luminosity as a function of time (right). The rate and luminosity has been computed as an average over a representative LHC fill of each year.

6.3.1 Event content

As previously mentioned, an offline reconstruction of the data recorded by the scouting strategy is not possible. As a consequence, the focus of the scouting strategy is to capture key information about the reconstructed objects which are sufficiently detailed for most physics analyses. By storing only the essential information per event, the scouting stream achieves a significant reduction in data size. A comparison of the event size and

bandwidth of the scouting and standard trigger strategies during a typical run are presented in Table 6.1. As the trigger system is a dynamic entity, frequently evolving to adapt to the different data-taking conditions (on a short term) and different physics goals (on a longer term), the numbers reported are not absolute.

Strategy	Rate (Hz)	Event Size	Bandwidth (MB/s)
Standard	2 000	1 MB	2 000
Scouting	25 000	8 kB	200

Table 6.1: Approximate rate, event size and bandwith for the scouting and standard trigger strategies during LHC Fill 8321 Run 361303 at 29th of October 2022.

The reduced event content consists of PF candidates, AK4 PF jets, PF MET, muons, electrons, photons, vertices, tracks and the average energy density in the event (ρ). Selection criteria on the objects mandate that jets and PF candidates have $|\eta| < 3.0$ as well as $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $p_{\mathrm{T}} > 0.6\,\mathrm{GeV}$, respectively. The scouting object information primarily consists of kinematic variables, including its energy and momentum components in three spatial directions. In Run–3, several track-related variables were added to the scouting PF candidates. The storage of PF candidates and the selection of stored variables allows the clustering of PF jets of any distance parameter during offline analysis and training of neural networks for tasks such as determination of jet origins.

6.3.2 Streams and datasets

ScoutingPF

The main scouting stream (`ScoutingPF`) functions to record events for scouting-based analyses. In 2022, the stream contained two datasets:

1. `DST_Run3_PFScoutingPixelTracking`
2. `DST_HLTMuon_Run3_PFScoutingPixelTracking`

The first dataset selects events based on a set of L1T seeds targeting events containing one photon (γ) or one electron (e), two or more γs or es, two

or more muons, one or two jets or a moderate amount of H_{T}. In 2023, the dataset split into four separate datasets to simplify data management and to optimise object and event selections. The four datasets target events with (a) two muons, (b) hadronic activity, (c) two γs or es with $p_{\mathrm{T}} > 16\,\mathrm{GeV}$ and (d) one γ or e object with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$.

Table 6.2 summarises the L1T seeds utilised by scouting in Run–3 to target events based on the presence of a single jet or a moderate amount of H_{T}. The lowest unprescaled seed selecting events based on the presence of H_{T} had a threshold of 360 GeV and 280 GeV in 2022 and 2023 respectively. The threshold was decreased in 2023 by unprescaling a seed with the corresponding threshold. The activation of the seed was facilitated by the decrease of the output rate, resulting from the additional constraints on scouting γ and e objects. The main scouting stream contains two more currently disabled seeds that may be activated when facilitated by a rate decrease.

L1 seed name	Selection criterion	2022 Prescale	2023 Prescale
`L1_HTT200er`	$H_T > 200\,\mathrm{GeV}$	0	0
`L1_HTT255er`	$H_T > 255\,\mathrm{GeV}$	0	0
`L1_HTT280er`	$H_T > 280\,\mathrm{GeV}$	0	1
`L1_HTT320er`	$H_T > 320\,\mathrm{GeV}$	0	1
`L1_HTT360er`	$H_T > 360\,\mathrm{GeV}$	1	1
`L1_HTT400er`	$H_T > 400\,\mathrm{GeV}$	1	1
`L1_HTT450er`	$H_T > 450\,\mathrm{GeV}$	1	1
`L1_SingleJet180`	One jet with $p_T > 180\,\mathrm{GeV}$	1	1
`L1_SingleJet200`	One jet with $p_T > 200\,\mathrm{GeV}$	1	1

Table 6.2: List of L1T seeds targeting events based on the presence of a single jet or a moderate amount of H_{T} in the trigger condition of the `ScoutingPF` dataset in 2022 and 2023. The selection criterion of each seed together with their respective prescales are displayed.

The second of the two datasets part of the main scouting stream selects events passing any HLT trigger related to muon activity, allowing jet based scouting analyses to perform calibrations on an *orthogonal* set of events. Two sets of events are orthogonal if they are selected independently of each other, such as events selected based on muon or jet triggers.

ScoutingPFMonitor

In contrast to the main stream, the secondary stream (`ScoutingPFMonitor`) contains only one dataset. This dataset has the purpose of monitoring the collection of scouting events and facilitating calibration studies of the scouting objects. The dataset contains both the scouting and offline reconstructed objects for a subset of events available in the main stream. Storing both reconstructed events allows for comparisons facilitating (a) the validation of the scouting reconstruction and (b) the computation of calibration factors. In order to facilitate a wide range of calibration studies, the trigger condition of this dataset contains triggers targeting all scouting objects. The triggers are presented in Table 6.3.

Trigger name	Selection criteria	Prescale
`DST_Run3_PFScouting-PixelTracking`	Logical 'OR' of all L1T seeds present in the main dataset	1000
`HLT_Ele115_CaloIdVT-_GsfTrkIdT`	A well-reconstructed electron with $p_{\mathrm{T}} > 115$	12
`HLT_Ele35_WPTight_Gsf`	An well-reconstructed electron with $p_{\mathrm{T}} > 35$	200
`HLT_IsoMu27`	An isolated muon with $p_{\mathrm{T}} > 27$	150
`HLT_Mu50`	A muon with $p_{\mathrm{T}} > 50$	50
`HLT_PFHT1050`	$H_T > 1050$	10
`HLT_Photon200`	A photon with $p_{\mathrm{T}} > 200$	10

Table 6.3: A list of the triggers present in the trigger condition of the `ScoutingPFMonitor` dataset in 2022. The selection criteria of each trigger together with their respective prescales are also displayed.

As both the raw detector information and the scouting reconstruction are stored for each event, the event size of this dataset is closer to that of the standard trigger approach. The triggers used to record events are therefore heavily prescaled in order to avoid straining the bandwidth of the DAQ system. Consequently, the rate is approximately 35 Hz (0.1% of the main scouting stream and 2.6% of the nominal standard trigger strategy stream). While necessary, the presence of the prescales increases the statistical uncertainty of the studies conducted with the dataset and impairs direct comparisons with simulation due to the difficulty of accurately modelling prescales in simulation.

6.3.3 Trigger efficiency

The scouting strategy enables lower hadronic trigger thresholds than the standard strategy, which relies on offline reconstructed data. As listed in Table 6.2, in 2023 the scouting trigger condition included unprescaled seeds targeting events based on the presence of at least one jet with p_{T} exceeding 180 GeV or H_{T} exceeding 280 GeV. In comparison, the lowest unprescaled triggers of the standard trigger strategy required jet p_{T} or H_{T} to exceed 500 GeV and 1050 GeV, respectively. In this section, a comparison between the performance of the jet selection (quantified as the trigger efficiency) is used to demonstrate the different trigger thresholds of the two strategies. In order to compare the events selections as a function of offline reconstructed jet observables, the scouting efficiency is computed with the `ScoutingPFMonitor` dataset (Section 6.3.2).

The trigger efficiency is measured with the reference method, as explained in Section 5.1.3. The efficiency is measured using an unbiased sample of events, collected with a single-muon trigger and containing only one well-identified and isolated muon outside of the jet cone. Events with additional muons are excluded. At least one well-reconstructed PF jet is required in the event, and jets must also pass identification criteria that reject poorly reconstructed jets or jets arising from detector noise. The AK4 PF jets are required to have $|\eta| < 2.5$ and $p_{\mathrm{T}} > 30$ GeV, whereas the AK8 PF jets require $|\eta| < 2.5$ and $p_{\mathrm{T}} > 170$ GeV. The efficiency is defined as the ratio of the number of events where an offline reconstructed PF jet is selected by the data scouting or standard triggers, relative to the total number of events with an offline reconstructed PF jet.

Figure 6.3 presents the efficiencies of collision data collected in 2023 as a function of the offline reconstructed AK4 jet p_{T} and H_{T} for the unprescaled L1T seeds listed in Table 6.2. The curves display a noticeable turn-on, due to the limited p_{T}-resolution of the L1T. The logical 'OR' expression of all considered seeds is fully efficient from approximately 300 GeV and 600 GeV for the selection of AK4 jets and H_{T}, respectively.

In order to compare the jet selection between the scouting and the standard trigger strategy, the efficiency of event selection based on the presence of at least one energetic jet or sufficiently energetic H_{T} is compared. The result is presented in Fig. 6.4 for collision data collected in 2022, as a function of the offline reconstructed AK4 jet p_{T}, AK8 jet p_{T} and H_{T}. The low thresholds of the scouting triggers are visible in the plot of each jet observ-

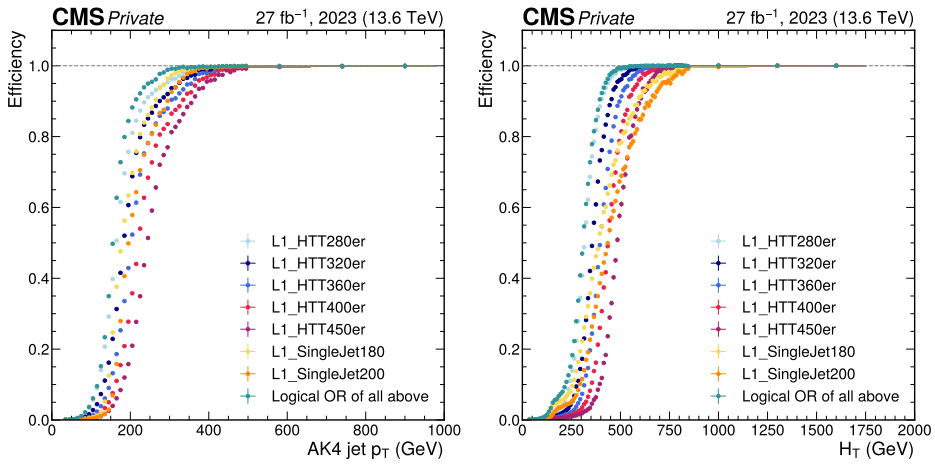


Figure 6.3: Trigger efficiency of the unprescaled L1T seeds targeting events based on the presence of a single jet or a moderate amount of H_{T} in 2023 as a function of the AK4 jet p_{T} (left) and H_{T} (right). A logical 'OR' of all unprescaled seeds is displayed in green. The uncertainties are entirely statistical and calculated as Clopper-Pearson intervals.

able. The efficiency to select AK4 and AK8 scouting jets is approximately 100% for $p_{\mathrm{T}} > 300\,\mathrm{GeV}$. In contrast, the standard trigger is only fully efficient from around 700–800 GeV. Similarly, data scouting is fully efficient for $H_{\mathrm{T}} > 600\,\mathrm{GeV}$, compared to roughly 1300 GeV for the standard trigger. As a result, jet-based analyses relying on the scouting technique are able to probe regions of phase space inaccessible with the standard trigger strategy. By lowering the H_{T} threshold from 360 GeV to 280 GeV in 2023, as discussed in Section 6.3.2, the scouting trigger improves even further the CMS acceptance to hadronic resonances.

6.3.4 Limitations

The scouting strategy in Run-3 is intricately linked to the performance of the pixel tracker, particularly due to a modified version of the PF algorithm (Section 6.3). In July 2023, 27 pixel modules in layer 3 and 4 were rendered inactive due to synchronisation problems in the internal clock of the signal supply tube. Consequently, a decrease in jet selection efficiency is observed within the region $-1.4 \leq \phi < -0.6$ and $-1.5 < \eta < 0$. The graphical representation of this inefficiency is provided in Fig. 6.5.

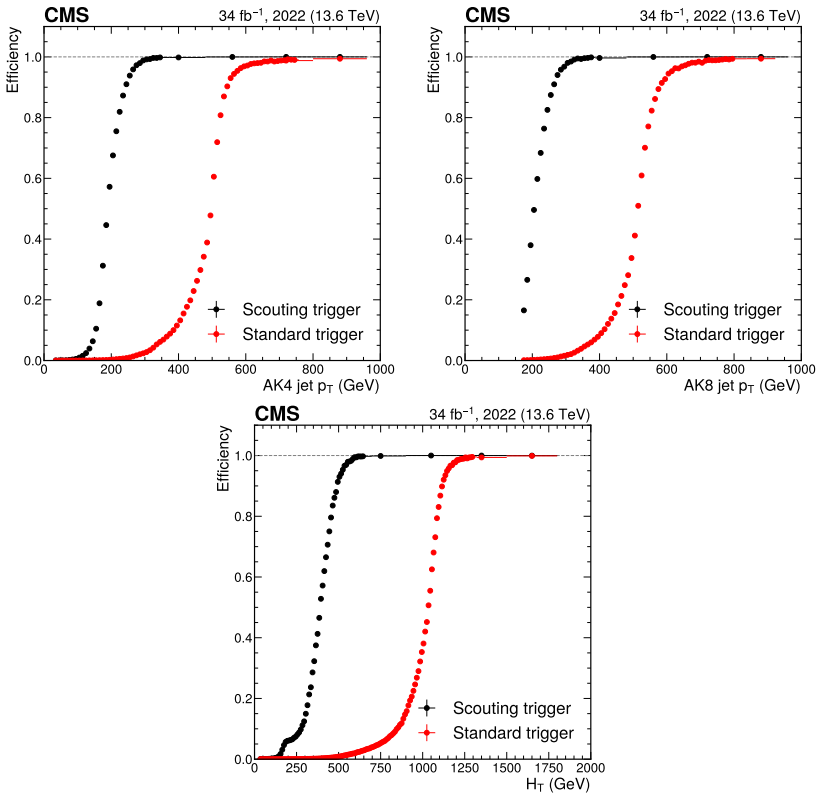


Figure 6.4: Trigger efficiency comparison for the scouting and and standard trigger strategy in 2022 as a function of the AK4 jet p_{T} (upper left), AK8 jet p_{T} (upper right) and H_{T} (lower). The uncertainties are entirely statistical and calculated as Clopper-Pearson intervals. Author's contribution to Ref. [54]

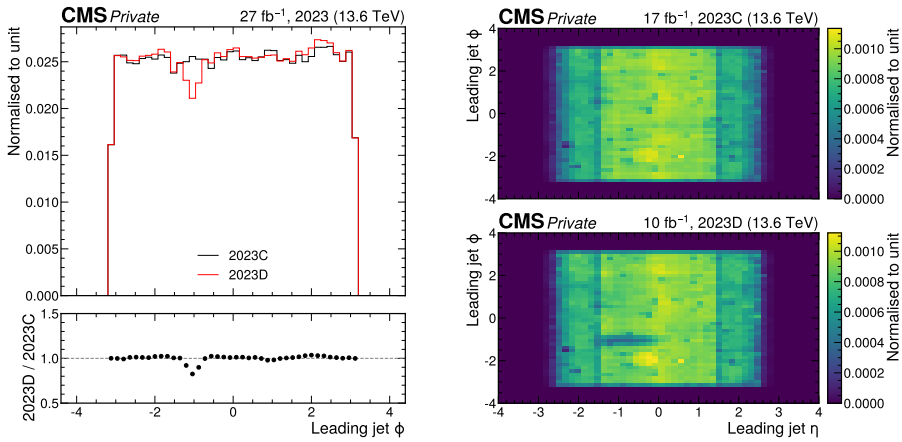


Figure 6.5: The loss of scouting jets due to problems in the pixel tracker estimated over collision data from that period. The labels "2023C" and "2023D" refer to the period of time before and after the pixel incident, respectively. A comparison of the leading jet ϕ distributions of 2023C and 2023D (left). A discrepancy is visible around $-1.4 \leq \phi < -0.6$. A comparison of the leading jet ϕ as a function of the leading jet η of 2023C and 2023D (right). A discrepancy is visible around $\phi = -1$ for $-1.5 < \eta < 0$.

The pixel modules remained masked for the remain of 2023, resulting in a scouting jet loss of approximately 7%. This loss was quantified by comparing the number of jets within the range $-1.4 \leq \phi < -0.6$, normalised by integrated luminosity, before and after the incident.

At present, the underlying issue remains unresolved and appears to necessitate intervention at the pixel hardware level. The proposed corrective operation involves the following sequence of steps: (a) warming up the entire tracker volume, (b) repositioning a segment of the pixel tracker to access the problematic area, (c) transporting the affected portion to the surface, (d) conducting repairs in a clean room environment, and (e) reversing the aforementioned steps to restore the pixel tracker. This operation is estimated to take between several weeks and several months, and poses a risk of damage to operational components of the pixel tracker.

In the interim, an alternative software-based solution has been proposed to restore efficiency. This solution involves integrating information from the strip tracker, in addition to the pixel tracker, during the seed generation for track reconstruction. The feasibility of this approach will be explored upon the availability of the software implementation for the scouting technique.

Subsequent investigation in this direction is scheduled after the conclusion of this thesis.

Chapter 7

Performance of Run–3 data scouting

Although the online reconstruction algorithms used by the HLT are similar to those used offline, the reconstructed physics objects resulting from the two procedures are not expected to be the same. While the offline reconstruction can undergo several reconstructions and improvements, the scouting reconstruction is only performed once during data collection. It is therefore imperative to assess the quality of the scouting objects and take action to improve the quality by amending the event reconstruction if it is deemed to be poor.

The offline reconstruction is continuously monitored by groups at the CMS experiment dedicated to the validation of its reconstructed physics objects. It is therefore possible, and useful, to demonstrate the validity of the scouting reconstruction by comparing the distributions of physics observables between the two reconstructions.

Such a comparison can be made with the jet energy. The energy of jets is a crucial observable in many physics analyses at the CMS experiment, including searches for new particles, measurements of the properties of known particles, and tests of the SM. The jet energy is not directly measured, but rather is reconstructed from the energy deposits in the detector. As a result, it is affected by various factors such as the choice of reconstruction method. Comparing the jet energy between the two reconstructions facilitates a direct evaluation of the scouting reconstruction. The comparison is achieved by deriving and comparing the jet energy scale (JES) and

jet energy resolution (JER) from each reconstruction.

In this chapter, the derivation of the JES and JER is presented. Firstly, the definitions of JES and JER used here are explained in Section 7.1. Next, the general methodology of deriving the JER and JES is described. This involves descriptions of the event selection criteria (Section 7.2), the general analysis method (Section 7.3) and the biases affecting the methodology (Section 7.4). The specific techniques of deriving the JES and JER are then presented together with the results in Section 7.5 and 7.6, respectively. Finally, conclusions of the measurements are presented in Section 7.7

7.1 Jet response

Particle-level

In general, as a result of detector noise and sub-optimal detector resolution, the jet momentum reconstructed from detector signals is not equal to the momentum of the particle that the jet originated from. This effect is quantified by the particle-level jet response defined as

$$\mathcal{R}^{\text{particle}} = \frac{p_T^{\text{reconstructed jet}}}{p_T^{\text{particle}}}. \tag{7.1}$$

$\mathcal{R}^{\text{particle}}$ provides a measure to compare the reconstructed jet momentum to the true momentum of the particle it originated from.

Generation-level

In simulated events, the momentum of the particle from which the jet originated corresponds to the momentum of the generated jet clustered from stable particles after hadronisation and decay. As a result, the particle-level jet response can be determined as

$$\mathcal{R}^{\text{simulation}} = \frac{p_T^{\text{reconstructed jet}}}{p_T^{\text{generated jet}}}. \tag{7.2}$$

$\mathcal{R}^{\text{simulation}}$ is referred to as the simulated jet response. Typically, the JES and JER are defined as the average simulated jet response $\langle \mathcal{R}^{\text{simulation}} \rangle$ and the width of the response distribution, respectively. However, since the goal of this chapter is to compare the scouting and offline reconstructed jets (not the reconstructed and generated jets) the JES has a different definition. Here, it is defined as

$$\text{JES} = \frac{\langle p_T^{\text{scouting}} \rangle}{\langle p_T^{\text{offline}} \rangle} \tag{7.3}$$

In contrast, the definition of the JER remains the same. Instead, a comparison between the scouting and the offline reconstructed JER is achieved by taking the ratio of $\text{JER}_{\text{scouting}}$ to $\text{JER}_{\text{offline}}$.

7.2 Dijet sample

Dataset and jet definition

In order to facilitate a fair comparison between the scouting and offline reconstructions, the `ScoutingPFMonitor` dataset (Section 6.3.2) is used to access both types of reconstructed objects for each event.

The selected jets are AK4 PF jets that are corrected with detector response corrections derived from simulation to adjust the measured response of reconstructed jets towards that of generated jets on average. The corrections applied to the scouting jets are derived specifically for online reconstructed jets, while those applied to the offline jets are derived for offline reconstructed jets. The same corrections are applied to jets reconstructed from simulation and collision data, with no *in-situ* corrections (such as JES calibration) applied to the latter. Before clustering, pileup is mitigated using the PUPPI technique for offline reconstructed jets and using the CHS technique for the scouting jets (Section 4.4.3).

Trigger selection

The events are collected using an array of L1T seeds and HLT triggers that select events containing at least one jet with a p_{T} exceeding a certain thresh-

old. The list of seeds and triggers are presented in the legend of Fig. 7.1. The figure displays the trigger efficiency calculated with the tag-and-probe method as explained in Section 5.1.3. Here, the *tag* and the *probe* are randomly assigned to the leading and sub-leading jet of the event. The leading and sub-leading scouting jets are matched to two offline jets by requiring an angular distance $\Delta R \leq 0.2$. The scouting tag is required to have a p_{T} above the trigger threshold. The efficiency is then defined as the ratio of events where a scouting probe is above the trigger threshold to the total number of events passing the tag requirement.

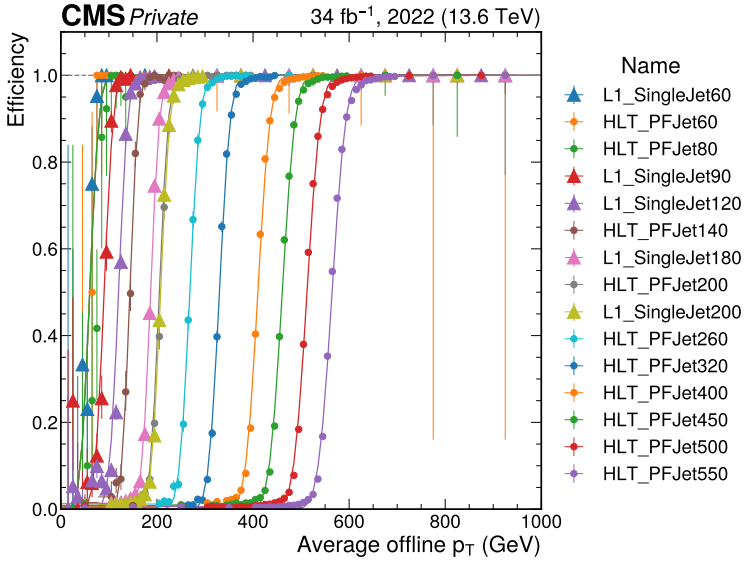


Figure 7.1: Trigger efficiency as a function of average offline p_{T} for each L1T seed and HLT trigger displayed in the legend. The uncertainty is entirely statistical and calculated as Clopper-Pearson intervals.

The efficiency is calculated as a function of the average offline jet p_{T} defined as

$$p_T^{average} = \frac{p_{T,1} + p_{T,2}}{2}, \tag{7.4}$$

where $p_{T,1}$ and $p_{T,2}$ refer to the p_{T} of the two leading offline jets.

A trigger threshold, denoted $p_T^{ave,95\%}$, is computed for each trigger by fitting the data points of the trigger efficiency curve with

$$f(p_T^{average}, a_1, a_2) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{p_T^{average} - a_1}{\sqrt{2}a_1}\right)\right), \tag{7.5}$$

where erf() refers to the error function while a_1 and a_2 are parameters of the function.

The value of $p_T^{ave,95\%}$ for each trigger is selected by finding the transverse momentum that results in $f(p_T^{ave,95\%}) = 0.95$. Each bin corresponds to a different trigger selection, with $p_T^{ave,95\%}$ equal to the bin's minimum value. To avoid any bias in the event selection, each bin is filled only by the events selected by the corresponding trigger (events selected by other triggers are ignored). For example, the bin ranging from 500 GeV to 550 GeV is filled by events selected by `HLT_PFJet450` whose $p_T^{ave,95\%}$ is approximately 500 GeV.

Dijet selection

As described in Section 4.4, a dijet event refers to a specific type of collision event comprising of two high-energy jets. In order to select such events, at least two well-reconstructed jets must be in the final state. The two leading jets of the event are then required to be back-to-back. This is achieved by requiring a minimum angular separation of 2.7 radians in the (x, y)-plane (illustrated in Fig. 7.2).

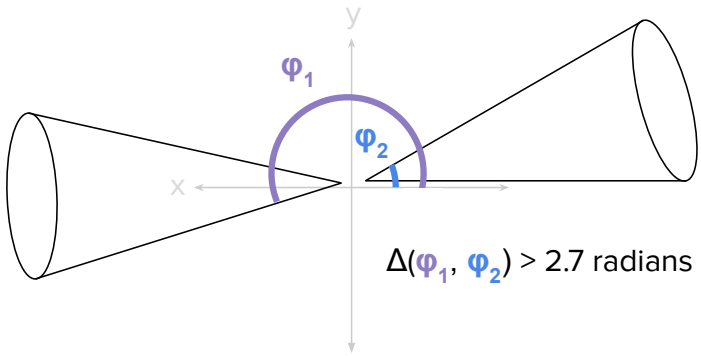


Figure 7.2: Illustration of a dijet event. The angular separation in the (x, y)-plane between the two leading jets is required to be larger than 2.7 radians.

7.3 Dijet p_T-balancing

The JES and JER measurements are performed with the dijet p_{T}-balancing method that utilises the principle of momentum conservation. As the protons part of the collision have no initial momentum in the transverse plane, the transverse momentum of the collision products must sum to zero. The p_{T} of the two jets in a dijet event must therefore be balanced [55]. In the absence of biases, a p_{T}-imbalance may stem from inaccuracies in reconstruction of the jet energy. By computing the p_{T}-imbalance while accounting for known biases, the accuracy of the reconstruction may be evaluated.

Due to varying detector responses in different η regions of the detector, the measurements are performed in bins of η defined as

$$|\eta_{1,2}| \leq 1.3, \tag{7.6}$$
$$1.3 < |\eta_{1,2}| \leq 2.5, \tag{7.7}$$

where $\eta_{1,2}$ refer to the η of the two leading jets.

The event is discarded if the two leading jets are located in different η regions. In this chapter, the region defined by Eq. 7.6 is referred to as ”barrel” while the region defined by Eq. 7.7 is defined as ”endcap”.

7.4 Biases

All methods based on collision data are affected by inherent biases related to physics properties and detector effects. The two most important biases related to the dijet p_{T}-balancing method are discussed: the radiation imbalance bias and the resolution bias.

7.4.1 Radiation imbalance bias

The measurement of p_{T}-imbalance is affected by the presence of extra jet activity, such as jets resulting from initial- and final-state radiation. The effect of extra jet activity can be demonstrated by assuming an estimator of the measured response defined as

$$\mathcal{R}^{measured} = \frac{p_{T,1}}{p_{T,2}}, \tag{7.8}$$

where $p_{T,1}$ and $p_{T,2}$ refer to the measured transverse momenta of the two jets in the dijet event [56].

The measured transverse momenta relate to the true transverse momenta (p_T^{true}) through the true response ($\mathcal{R}^{true}$) as follows

$$p_{T,1} = \mathcal{R}_1^{true} \times p_{T,1}^{true}, \tag{7.9}$$
$$p_{T,2} = \mathcal{R}_2^{true} \times p_{T,2}^{true}. \tag{7.10}$$

In the presence of extra jet activity $p_{T,2}^{true} = p_{T,1}^{true} - \Delta p_T$, where Δp_T quantifies the imbalance due to the additional radiation. By combining all equations above, the estimator $\mathcal{R}^{measured}$ can be redefined as

$$\mathcal{R}^{measured} = \frac{p_{T,1}/p_{T,1}^{true}}{p_{T,2}/p_{T,2}^{true}} \left(1 - \frac{\Delta p_T}{p_{T,2}^{true}}\right). \tag{7.11}$$

When $\Delta p_T \to 0$, resulting in $p_{T,1}^{true} \equiv p_{T,2}^{true}$, this relation becomes Eq. 7.8. As a result, the p_{T}-ratio between the two jets of the dijet event serves as a useful estimator of the measured response when the jets contributing to the extra jet activity are soft and negligible.

In order to control for the effect of extra jet activity, the JES and JER measurements are performed by considering

$$\alpha = \frac{p_T^{average}}{p_{T,3}}, \tag{7.12}$$

where $p_{T,3}$ is the p_{T} of the third leading jet. The equation equals zero if the event contains exactly two jets.

In the case of the JES measurement, events are required to have a small α in order to minimise the extra jet activity and the effect of the radiation imbalance bias. For the JER measurement, the bias is accounted for by

extrapolating the extra jet activity to 0. This is achieved by computing the JER several times, each time requiring α to be smaller than a certain threshold. The extra jet activity is then controlled for by extrapolating linearly to $\alpha = 0$.

7.4.2 Resolution bias

The measurement of the jet energy response is typically performed by comparing a jet to a reference object chosen on the basis of high resolution. For example, when deriving the p_{T}-imbalance between a photon (γ) and a jet, the γ object is selected as reference due to its superior p_{T} resolution [56]. However, in the case of dijet p_{T}-balancing, the two jets have comparable resolutions. In this case, the measurement of the jet energy response is biased by the object with the inferior resolution.

The reconstructed jet resolution is worsened due to the difference between reconstructed and true jet momenta. A reconstructed jet p_{T} bin does not only include jets whose true transverse momenta lie within that bin. Instead, jets whose reconstructed momenta have fluctuated slightly lower or higher than their true momenta may also occupy the same bin. As the p_{T} distribution of proton-proton collisions is a steeply falling spectra, the number of reconstructed jets with lower true p_{T} that fluctuated up is more than the number of jets with a higher true p_{T} which fluctuated down. As a result, the measured response is systematically higher.

In the dijet p_{T}-balancing method, the phenomena described above affects both jets. In order to mitigate the effect of this bias, the measurement of the response is computed in bins of average p_{T} defined by Eq. 7.4. If both jets have comparable resolutions, the bias is cancelled on average [56]. This is the case for the method used to measure the JER in this chapter. The effect it has on the JER measurement is quantified in Fig. 7.3.

7.5 Jet energy scale

In this chapter, the purpose of the JES is to relate the energy reconstructed for a scouting jet to the energy of the corresponding offline reconstructed jet. In the following text, the JES is measured separately from collision data and simulation, and a comparison of $\mathrm{JES}_{\mathrm{data}}$ to $\mathrm{JES}_{\mathrm{simulation}}$ is provided to

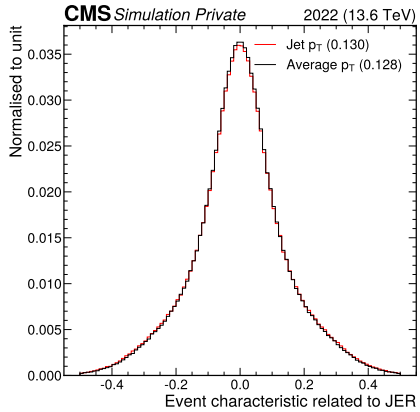


Figure 7.3: Illustrations of the resolution bias: an event characteristic relating to the JER in a jet p_{T} bin (red) and an average p_{T} bin in (black). The width of each distribution in the right plot is displayed in the parenthesis of the legend. A smaller width corresponds to a better JER.

showcase the level of agreement.

7.5.1 Methods

Dijet events are selected as outlined in Section 7.2. The dijet p_{T}-balancing technique is used, in which the leading scouting jet is chosen as *probe* and the sub-leading jet as *tag*. The scouting tag and probe are paired with two offline jets by requiring that the angular distance $\Delta R \leq 0.2$. The offline jets paired with the scouting probe and tag are referred to as offline probe and tag, respectively. In order to reduce the impact of the radiation imbalance bias (Section 7.4.1) a selection is applied on α, requiring $\alpha < 0.05$.

In order to derive JES (Eq. 7.3), four quantities are needed:

$$\langle \frac{p_T^{\text{scouting,probe}}}{p_T^{\text{offline,tag}}} \rangle \text{ as a function of } p_T^{\text{offline,tag}}, \tag{7.13}$$

$$\langle \frac{p_T^{\text{offline,probe}}}{p_T^{\text{offline,tag}}} \rangle \text{ as a function of } p_T^{\text{offline,tag}}, \tag{7.14}$$

$$\langle p_T^{\text{scouting,probe}} \rangle \text{ as a function of } p_T^{\text{offline,tag}}, \tag{7.15}$$

$$\langle \frac{p_T^{\text{scouting,probe}}}{p_T^{\text{offline,probe}}} \rangle \text{ as a function of } p_T^{\text{offline,tag}}. \tag{7.16}$$

The JES is then derived by following the three steps outlined below.

1. Eq. 7.13 (displayed in Fig. 7.4) is divided by Eq. 7.14. The division results in Eq. 7.3 as a function of $p_T^{\text{offline,tag}}$.
2. Eq. 7.15 (displayed in Fig. 7.5) is then used to map the result of Step 1 from $p_T^{\text{offline,tag}}$ to $\langle p_T^{\text{scouting,probe}} \rangle$.
3. Finally, the standard deviation of Eq. 7.16 is used to assign the uncertainty on the result of Step 2.

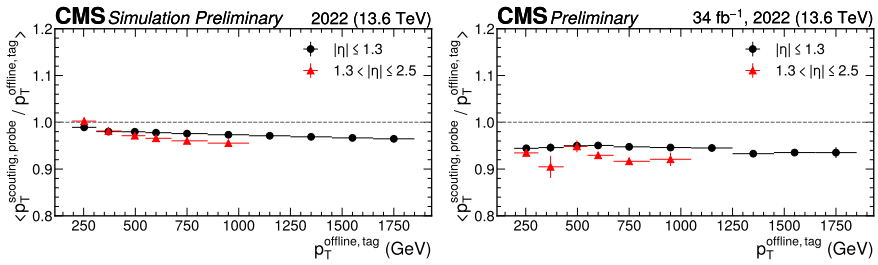


Figure 7.4: $\langle p_T^{\text{scouting,probe}} / p_T^{\text{offline,tag}} \rangle$ as a function of $p_T^{\text{offline,tag}}$. Created with simulation (left) and collision data (right). The uncertainty is entirely statistical and defined as the standard deviation divided by the square root of the total number of events.

Figure 7.4 shows that on average the scouting probe has a lower p_T than the offline tag. As the scouting probe is chosen as the leading scouting jet, and the offline tag is paired with the sub-leading scouting jet, scouting jets on average have a lower p_T than their offline reconstructed counter parts. A similar trend (particularly for data at high momentum) is visible in Fig. 7.5, which showcases the mean scouting probe p_T as a function of the offline tag p_T.

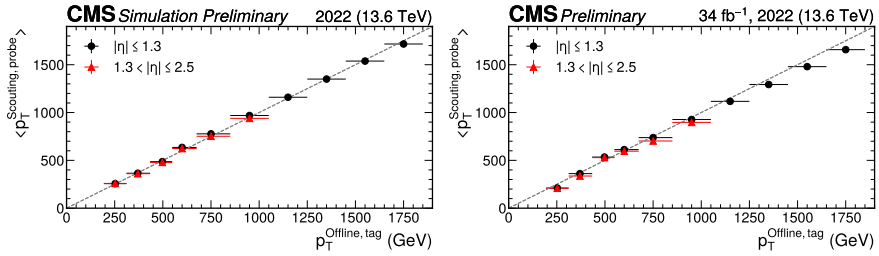


Figure 7.5: $\langle p_T^{\text{scouting,probe}} \rangle$ as a function of $p_T^{\text{offline,tag}}$. Created with simulation (left) and collision data (right). The uncertainty is entirely statistical and defined as the standard deviation divided by the square root of the total number of events.

7.5.2 Results

The final result of the JES measurement is presented in Fig. 7.6. The creation of $\langle p_{\text{T}}^{\text{scouting}} \rangle$ involves a mapping from $p_{\text{T}}^{\text{offline}}$, resulting in varying bin widths across different η regions, as well as discrepancies between bin widths of simulation and recorded data.

As illustrated below, the JES is similar for both simulation and collision data; approximately 0.97 for $|\eta| \leq 1.3$, and 0.98 for $1.3 < |\eta| \leq 2.5$. The ratio of the JES derived from collision data to the JES derived from simulation is presented in Fig. 7.7. The ratio approximates to 1, indicating a good level of agreement between simulation and collision data.

7.6 Jet energy resolution

The JER describes the resolution at which the energy of a jet can be measured and quantifies the uncertainty in the reconstructed energy of the jet. In the following text, the JER of scouting and offline reconstructed jets are derived and compared by calculating their ratio. In parallel with the JES measurement, a comparison of JER_{data} to $\text{JER}_{\text{simulation}}$ is provided to showcase the level of agreement.

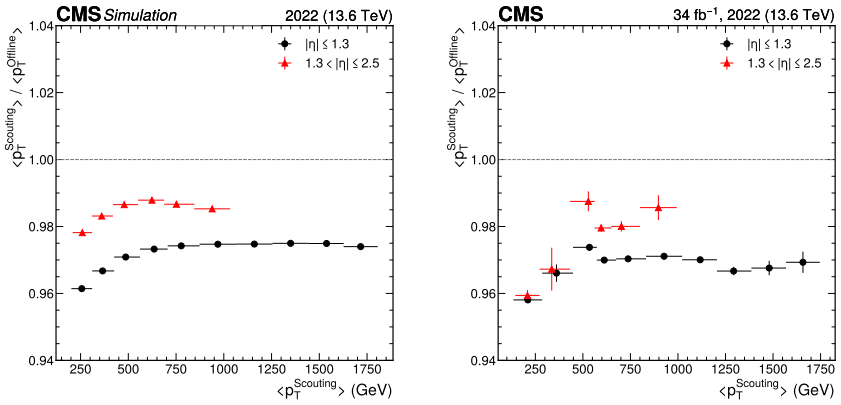


Figure 7.6: The JES as a function of $\langle p_T^{\text{scouting,probe}} \rangle$. Created with simulation (left) and collision data (right). The uncertainty is entirely statistical and is defined as the standard deviation of Eq. 7.15. Author's contribution to Ref. [57].

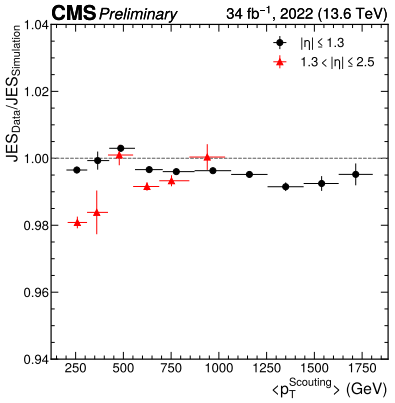


Figure 7.7: The ratio of the JES derived from collision data to the JES derived from simulation. The uncertainty is entirely statistical and is computed through error propagation of the numerator and denominator whose uncertainty is defined as the standard deviation of Eq. 7.15. Author's contribution to Ref. [57].

7.6.1 Methods

Dijet events are selected as outlined in Section 7.2. To account for the resolution bias (Section 7.4.2), the JER is measured in bins of average p_{T}. The measurement is performed using the dijet *asymmetry* technique [55], which exploits the dijet p_{T}-balancing method.

Asymmetry

The asymmetry of a dijet event is defined as

$$A = \frac{p_{T,1} - p_{T,2}}{p_{T,1} + p_{T,2}}, \tag{7.17}$$

where $p_{T,1}$ and $p_{T,2}$ refer to the randomly ordered transverse momenta of the two leading jets.

The asymmetry is expected to be close to 0, but may vary due to measurement biases and errors in reconstruction. The distribution of asymmetries mimics a Gaussian distribution with long tails, and is exemplified in Fig. 7.8.

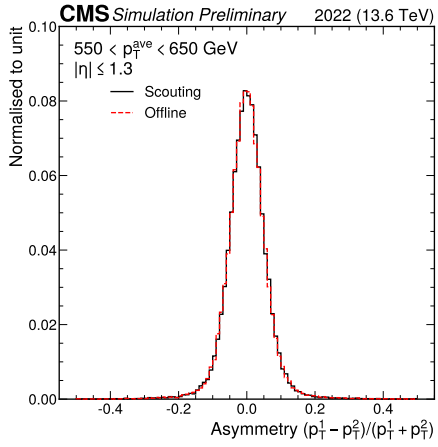


Figure 7.8: Comparison of asymmetry distributions of the scouting and offline reconstructions for simulated events with $550 < p_T^{\text{average}} < 650$ and $|\eta| < 1.3$. The histograms are normalised to unit. Author's contribution to Ref. [57].

Width of the asymmetry

Ignoring the tails, the asymmetry is approximately Gaussian distributed with a standard deviation defined as

$$\sigma(A) = \left|\frac{\delta(A)}{\delta(p_{T,1})}\right| \sigma(p_{T,1}) + \left|\frac{\delta(A)}{\delta(p_{T,2})}\right| \sigma(p_{T,2}). \tag{7.18}$$

In the ideal case where the two jets of the dijet event are located in the same η region and have perfectly balanced momenta, the following relationships are true:

$$\langle p_{T,1}\rangle = \langle p_{T,2}\rangle = \langle p_T\rangle, \tag{7.19}$$
$$\sigma(p_{T,1}) = \sigma(p_{T,2}) = \sigma(p_T). \tag{7.20}$$

These relationships allow for a simplification of Eq. 7.18. The new equation relates the JER to the width of the asymmetry distribution according to

$$\frac{\sigma(p_T)}{p_T} = \sigma(A) \times \sqrt{2}, \tag{7.21}$$

where $\sigma(A)$ is the width of A and $\sigma(p_T)$ is the JER.

There are several methods that can be used to compute the width of the asymmetry distribution. A commonly used approach is to fit a Gaussian function to the asymmetry distribution and define $\sigma(A)$ by the parameters of this function. However, this method does not account for the presence of tails in the distribution. As a consequence, a more robust method is used in this section. This method computes $\sigma(A)$ as the *effective resolution* (described in Ref. [58]) and is achieved by finding the smallest interval containing 68% ($\pm 1\%$) of the events, and dividing that interval by 2.

Simulated jet response

In order to account for the missing calibration of the JES (Section 7.2), it is necessary to account for the simulated jet response ($\mathcal{R}^{\text{simulation}}$, Eq. 7.2)

when measuring the JER. This is a common practice when estimating the JER to avoid a bias due to an imperfect JES calibration.

$\langle \mathcal{R}^{\text{simulation}} \rangle$ as a function of the generated jet p_{T}, is displayed in Fig. 7.9. The response is stable at high p_{T} with a value of approximately 1.004 and 1.005 for offline jets in the barrel and endcap respectively. The response is slightly worse for scouting jets, which is expected as the corrections applied are derived for online jets but not specifically scouting jets (Section 7.2). The response is approximately 0.991 and 1.01 for scouting jets in the barrel and endcap respectively.

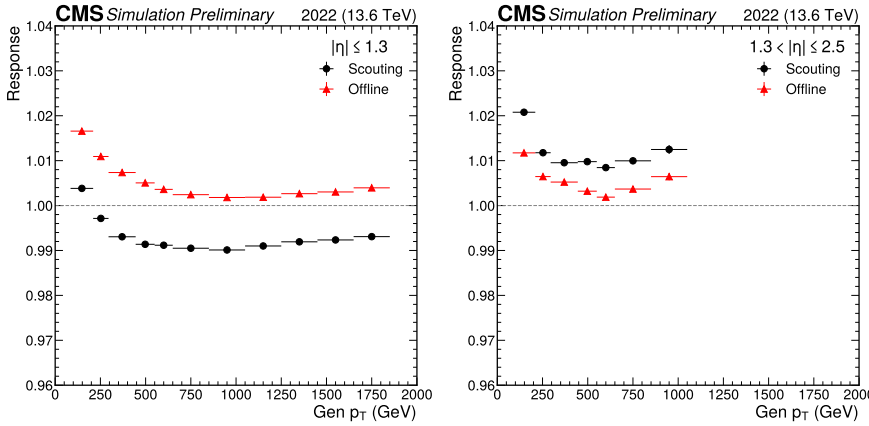


Figure 7.9: Average simulated jet response as a function of generated jet p_{T}, for the barrel (left) and endcap (right) The uncertainties are entirely statistical and defined as the standard deviation divided by the square root of the total number of events. Author's contribution to Ref. [57].

The $\mathcal{R}^{\text{simulation}}$ is accounted for in the calculation of JER as

$$\frac{\sigma(p_T)}{p_T} = \frac{\sigma(A) \times \sqrt{2}}{\langle \mathcal{R}^{\text{simulation}} \rangle}. \tag{7.22}$$

Eq. 7.22 is the final formula needed to measure the JER.

α extrapolation

To account for the radiation imbalance bias (Section 7.4.1), the measurement of the JER is performed four times with decreasing amounts of extra

jet activity. The JER is then extracted by extrapolating the extra activity to zero. The variable α represents the extra jet activity, and the four inclusive α bins used are $\alpha < 0.2$, $\alpha < 0.15$, $\alpha < 0.1$, and $\alpha < 0.05$.

An example of the extrapolation procedure is displayed in Fig. 7.10. The point of intersection with the y-axis is considered to be the JER without extra jet activity, and is identified by applying a linear fit to the data points.

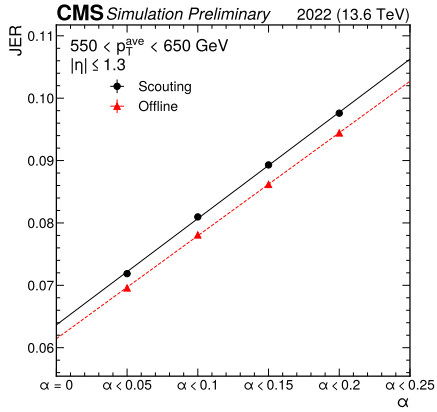


Figure 7.10: The value of $\sigma(A)$ as a function of α for simulated events with $550 < p_T^{\text{average}} < 650$ and $|\eta| < 1.3$. A linear fit is used to determine the point of intersection with the y-axis. The uncertainty is entirely statistical and defined as the square root of the variance divided by the total number of events. Author's contribution to Ref. [57].

7.6.2 Results

The final result of the JER is presented in Fig. 7.11. The results show a good level of agreement between the scouting and offline reconstructed jets. The resolution is stable from around 500 GeV, with a value of approximately 0.05 and 0.06 in the barrel and endcap regions respectively.

Figure 7.12 displays the ratio of the JER derived from scouting reconstructed events to the JER derived from offline reconstructed events. The ratio shows a disagreement of approximately 10% and 2% below and above 500 GeV respectively. Finally, Fig. 7.13 displays the ratio of the JER derived from collision data to the JER derived from simulation. The ratio is consistently above 1, which is expected as simulation always assumes a better detector performance than is observed in collision data. The only exception to this

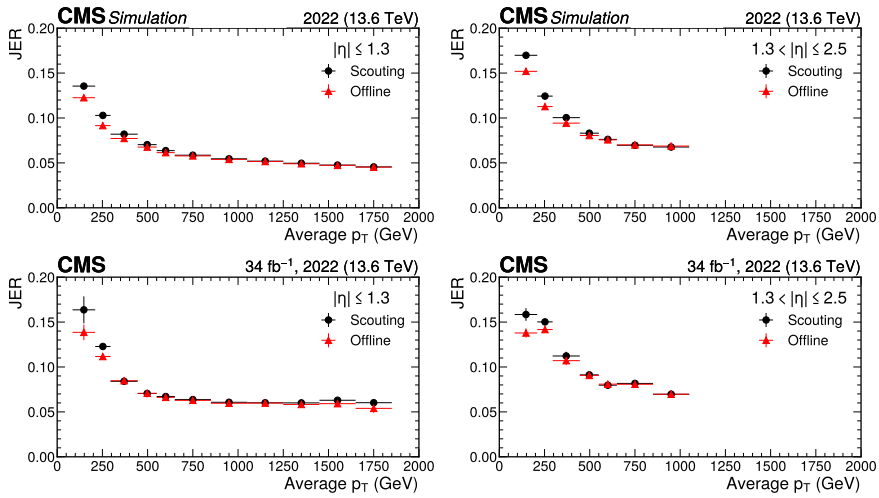


Figure 7.11: JER as a function of average p_{T}. Computed from simulation (upper) and collision data (lower), with events from the barrel (left) and endcap (right). The uncertainties are taken from the linear fit of the extrapolation procedure, and account for the number of events in each bin as well as the uncertainty of the fit. Author's contribution to Ref. [57].

is the first p_{T} bin in the endcap. This discrepancy may have arisen due to the reduced number of events in this particular bin as a result of the trigger selection. The large uncertainty of that bin accounts for the discrepancy.

7.7 Conclusion

The JES and JER measurements show a good level of agreement between the scouting and offline reconstructions. In addition, the JES and JER derived from collision data are within the statistical uncertainty of the simulation.

The impact of the differences observed in the JER between the two reconstructions on the results of subsequent analyses can be quantified by

$$\mathrm{JER}_{\mathrm{final}} = (1 + \Delta_{JER}) \times \mathrm{JER}, \tag{7.23}$$

where Δ_{JER} is the discrepancy between the reconstructions (expressed in

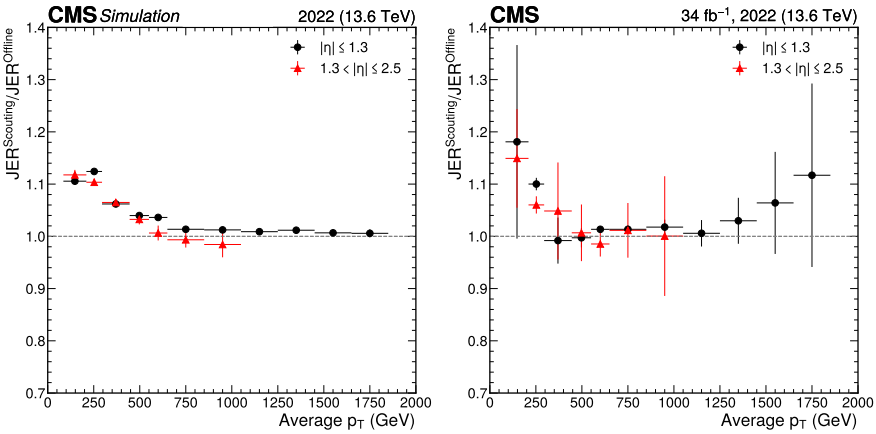


Figure 7.12: The ratio of JER derived from scouting reconstructed events to the JER derived from offline reconstructed events. Created with simulation (left) and collision data (right). The uncertainties are derived with error propagation of the numerator and denominator and account for the number of events in each bin as well as the uncertainty of the linear fit of the extrapolation procedure. Author's contribution to Ref. [57].

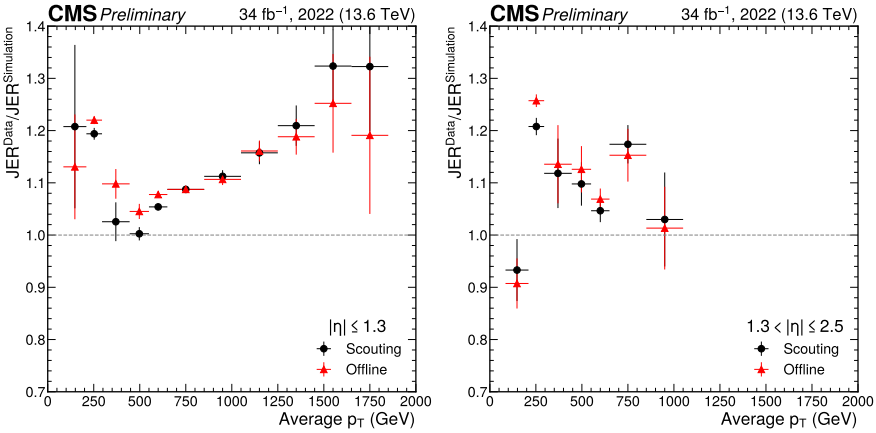


Figure 7.13: The ratio of the JER derived from collision data to the JER derived from simulation with events from the barrel (left) and endcap (right). The uncertainties are derived with error propagation of the numerator and account for the number of events in each bin as well as the uncertainty of the linear fit of the extrapolation procedure. Author's contribution to Ref. [57].

decimals).

Given a JER of 0.05 (as in Fig. 7.11 at high p_{T}) and a Δ_{JER} ranging from 2–10% (as in Fig. 7.12), $\mathrm{JER}_{\mathrm{final}}$ is 0.051–0.055. This change is almost negligible for the purposes of most analyses. This is because the JER is incorporated into analyses involving jets by smearing the simulated jet energy according to a parameterised Gaussian function that represents the JER. As it has previously been demonstrated that the JER in collision data is larger than in simulation [56], it is necessary to smear the simulated jets to accurately model the collision data jets. To quantify the impact of the JER uncertainty, the analysis is repeated multiple times using different versions of the Gaussian smearing function. This effectively accounts for the possible variations in reconstructed jet energies due to the JER. This uncertainty is then quantified by calculating the spread of the analysis results. The larger the spread, the larger the impact of the JER uncertainty on the analysis. This systematic uncertainty is often one of the dominant uncertainties in analyses involving jets, so a change of 2–10% in the JER is insignificant to the final uncertainty associated with it.

In conclusion, the results presented here indicate that the scouting reconstructed jets are suitable for almost all analyses that are statistically limited, but not systematically limited, such as precision measurements that require the best possible JER.

7.7.1 Future perspectives

It was noted during the measurements of the JES and JER that the trigger condition of the `ScoutingPFMonitor` dataset could be further optimised for future studies of this nature. While the measurements require several triggers targeting events containing at least one jet with a p_{T} exceeding different thresholds, as discussed in Section 6.3.2, the dataset only recorded events containing hadronic activity based on the presence of $H_T > 1050\,\mathrm{GeV}$. This ultimately prevented the study of JES and JER below a p_{T} of 200 GeV.

In order to improve future measurements, a study to improve the trigger condition was conducted. The study compared the replacement of `HLT_PFHT1050` with an array of triggers selecting events based on the presence of at least one jet with p_{T} exceeding varying thresholds (denoted as "Array of `HLT_PFJet`"). The study was performed with a subset of events recorded by the `HLTPhysics` stream (Section 5.2.2) in 2023 and the results

are displayed in Fig. 7.14 as a function of leading jet p_{T} and average p_{T}. While the events are few, the problem with the old trigger condition is clear: low-p_{T} events are not selected.

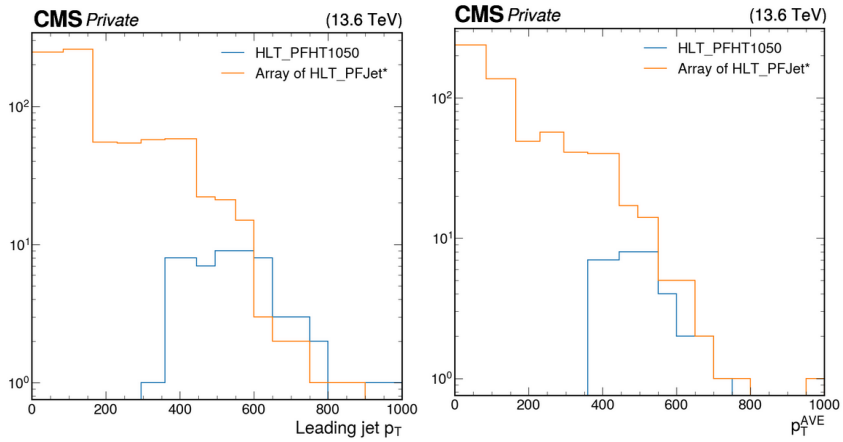


Figure 7.14: Comparison of number of events selected with the old (blue) and new (orange) `ScoutingPFMonitor` trigger condition as a function of leading jet p_{T} (left) and average p_{T} (right).

While this study was successful and steps were made to provide a more appropriate trigger condition, it was not possible to repeat the measurement with the improved condition before the submission of this thesis due to a premature shut down of the LHC accelerator.

Part II

Data analysis

Chapter 8

Theoretical background

This chapter serves as an introduction to the theoretical underpinnings of the physics analysis presented in this thesis. First, an overview of the central concepts of the Standard Model is provided (a comprehensive review can be found in Ref. [59]). The Higgs mechanism, which provides particles with mass and gives rise to the Higgs boson, is then discussed. Finally, the production and decay modes on the Higgs boson are presented with special emphasis of the production and decay modes crucial to the physics analysis presented later in this thesis.

8.1 Overview of Standard Model

The *Standard Model* (SM) of particle physics is a theory that describes all known elementary particles and their interactions at the fundamental level. Based on an internal property referred to as *spin*, the elementary particles are divided into two distinct categories. *Fermions* are distinguished by their half integer spin, and make up the matter of our universe. *Bosons* have a integer spin, and mediate the interactions between the fermions.

The fermions include 12 elementary particles along with their respective *antiparticles*. The antiparticles have opposite charge but the same mass and spin. The fermions among the elementary particles are further divided into *quarks* and *leptons*, which exist in three generations of different mass scales. For quarks, each generation consists of a quark doublet, while for leptons, each generation consist of one massive and one extremely light

lepton. While the first generation is stable, the particles of the two other generations are short-lived and decay into lighter particles.

All quarks carry a colour charge, and are combined to form colour neutral composite particles called *hadrons*. The three colours are denoted as *red*, *blue* and *green*. In addition to colour charge, quarks also carry electric charge. The three up-type quarks (up, charm and top) carry an electric charge of $+2/3$, while the three down-type quarks (down, strange and bottom) have an electric charge of $-1/3$.

The leptons also carry electric charge. The three charged leptons (electron, muon and tau) have an electric charge of -1. The charged leptons are accompanied by corresponding neutral leptons called neutrinos (electron neutrino, muon neutrino and tau neutrino). While the SM predicts that the neutrinos have zero mass, the neutrino masses have been experimentally measured to be very small but not zero [37].

Particle	Generation	Mass	Electric charge
Quarks			
up (u)	1	2.2 MeV	$+2/3$
down (d)	1	4.7 MeV	$-1/3$
charm (c)	2	1.27 GeV	$+2/3$
strange (s)	2	93.4 MeV	$-1/3$
top (t)	3	172.7 GeV	$+2/3$
bottom (b)	3	4.18 GeV	$-1/3$
Leptons			
electron (e)	1	0.511 MeV	-1
electron neutrino (ν_e)	1	< 0.8 eV	0
muon (μ)	2	106 MeV	-1
muon neutrino (ν_μ)	2	< 0.19 MeV	0
tau (τ)	3	1.777 GeV	-1
tau neutrino (ν_τ)	3	< 18.2 MeV	0

Table 8.1: List of fermonic particles in the SM. Masses are obtained from Ref. [37].

The SM explains three of the four fundamental forces governing our universe; the strong force, the weak force and the electromagnetic force. The last two forces appear very different at low energies, but merge into a single force (called the the electro-weak force) above the *electro-weak scale* of approximately 100 GeV.

Each force is carried by gauge bosons, also known as vector bosons. The chargeless photon (γ) mediates the electromagnetic (EM) force, the eight flavours of massless and electrically neutral gluons (g) mediate the strong force, and the charged $W^{\pm}$ and neutral Z bosons mediate the weak force. The Higgs boson, a spin-0 particle, gives mass to fundamental particles.

Particle	Mass	Electric charge
photon (γ)	0	0
gluon (g)	0	0
$W^{\pm}$	80.38 GeV	± 1
Z	91.19 GeV	0

Table 8.2: List of gauge bosons in the SM. Masses are obtained from Ref. [37].

8.2 Gauge groups

The SM is built upon the principles of quantum field theory, where particles are represented as excitations of quantum fields. Each fundamental particle corresponds to a specific quantum field that permeates all space. A *gauge group* defines the type of *symmetry transformations* that the field can undergo while leaving the physical behaviour of the system unchanged. The SM is built from three types of gauge groups.

1. **U(1)** is associated with the electromagnetic force and is represented by the photon. The gauge group is described by the theory of quantum electrodynamics (QED) that models interactions between charged particles.
2. **SU(2)** describes the weak nuclear force, which is responsible for processes such as beta decay and neutrino interactions. The gauge bosons associated with this gauge group are the $W^{\pm}$ and Z bosons.
3. **SU(3)** is associated with the strong nuclear force, governed by gluons. The strong force is responsible for forming hadrons, such as protons and neutrons, out of their constituent quarks.

Each of these gauge groups is described by a specific mathematical framework called a *gauge theory*. By combining the theories associated with the

three gauge groups listed above, a gauge group describing all fundamental particles and their interactions is obtained. This group is denoted as

$$U(1) \times SU(2) \times SU(3). \tag{8.1}$$

8.3 Higgs mechanism

The combination of the weak and electromagnetic force, known as the electroweak force, is represented by combining the gauge group U(1) and SU(2). The requirment of a local gauge symmetry in the group forces the gauge bosons (the $W^{\pm}$ and Z bosons, and photon) to be massless. However, experimental observations clearly indicate that the $W^{\pm}$ and Z bosons are in fact massive [60, 61]. To address this discrepancy, a mechanism involving the Higgs field was proposed.

The Higgs field is a quantum field that permeates all space, and when acquired a nonzero value, breaks the electroweak symmetry. The symmetry is broken spontaneously by introducing the *Englert-Brout-Higgs-Guralnik-Hagen-Kibble mechanism* [62–67], which generates masses for the $W^{\pm}$ and Z bosons through the interactions with the particles and the Higgs field. In addition, the mechanism provides a way for fermions to acquire mass, through the so-called *Yukawa interactions*.

A particle referred to as the Higgs boson emerges as a quantum excitation of the Higgs field, and can be detected experimentally. The Higgs boson's discovery by the CMS [33, 34] and ATLAS [68] experiments in 2012 confirmed the existence of this mechanism and provided a crucial piece of evidence for the SM. So far, all measurements of the properties of the Higgs boson suggest that the particle is compatible with the SM expectation [69, 70].

8.4 Higgs production modes

The probability of a certain particle process occurring is described by the so-called *cross section*. The cross section of a Higgs boson production mode is determined by a combination of factors related to the nature of the particles involved and the available energy in the collision. The total cross

section for the production of the SM Higgs boson in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$ is 54 ± 2.6 pb [71]. This results in the production of roughly one Higgs boson every second at an instantaneous luminosity of $2 \times 10^{34}\,\text{cm}^{-2}\text{s}^{-1}$ [69].

At the LHC, the production of Higgs bosons is dominated by a processes referred to as *gluon-gluon fusion* (ggF, illustrated by a Feynman diagram in Fig. 8.1). In this process, two gluons from the colliding protons interact. These gluons are excitations of the quantum field associated with the strong nuclear force, which momentarily transform into virtual quarks through a quantum process called *loop correction*. A virtual particle is a concept within the framework of quantum field theory and represents an intermediate state in particle interactions. They are not directly observed as free particles, but are instead considered to be fluctuations in the quantum field.

In the process of ggF, the virtual quarks generated through loop correction emit a virtual Higgs boson. This Higgs boson can be thought of as emerging from the interaction of the quarks and the Higgs field. Due to the large cross section of the production mode, ggF is often targeted by analyses studying rare decays.

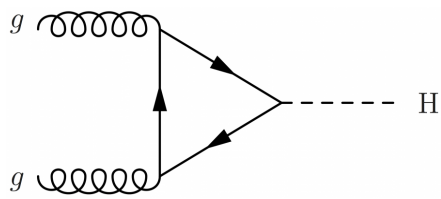


Figure 8.1: Feynamn diagram of Higgs boson production through gluon-gluon fusion, where g denotes a gluon and H a Higgs boson.

Vector boson fusion (VBF, illustrated in Fig. 8.2) is the second largest Higgs boson production mode at the LHC. In VBF, two quarks from the colliding protons radiate virtual vector bosons. As these vector bosons interact with each other, a Higgs boson is emitted as a quantum excitation of the Higgs field. The initial quarks that radiated the vector bosons travel mostly along their initial directions, being deflected only very slightly. Due to the presence of two forward-going quarks, VBF provides a distinctive experimental signature that is reconstructed as two forward-going jets by the CMS detector. A study of the VBF and ggF production modes are presented in detail in Chapter 11.

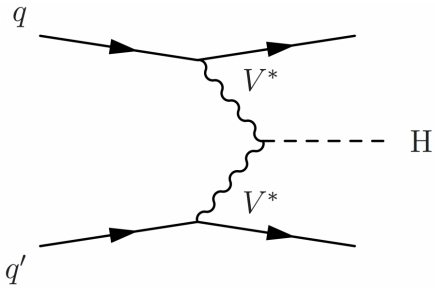


Figure 8.2: Feynamn diagram of Higgs boson production through vector boson fusion, where q denotes a quark, V^* a Z or $W^{\pm}$ boson and H a Higgs boson.

There are two additional prominent Higgs boson production modes at the LHC; *Higgs strahlung* (VH) and *top quark fusion* (ttH). While VH involves the interaction of a quark with a virtual vector boson, ttH occurs through the interaction of two top quarks. Both interactions result in the emission of a Higgs boson.

While ggF is the predominant production mode at the LHC, the fraction of Higgs bosons generated through other production modes is affected by the p_{T} of the Higgs boson. This is illustrated in Fig. 8.3, where the relative contribution to the cumulative cross sections as a function of the Higgs boson p_{T} threshold is displayed for each production mode described in previous paragraphs. While ggF contributes 87% of the cross section at $\sqrt{s} = 13\,\mathrm{TeV}$ when considering the full p_{T} range [72], the relative contribution decreases to 50% at $p_{\mathrm{T}} > 450\,\mathrm{GeV}$. At $p_{\mathrm{T}} > 1200\,\mathrm{GeV}$, the ggF and VH rates are comparable, with each contributing about 35% of the total Higgs boson production cross section.

8.5 Higgs decay modes

The relatively short lifetime of the Higgs boson (1.6×10^{-22} seconds as predicted by the SM [73]), is influenced by its interactions with a large number of particles. The Higgs boson interactions occur through *gauge couplings* for gauge bosons and *Yukawa couplings* for fermions. These couplings determine the strength of the interaction between the Higgs boson and a given particle, and is related to the masses of the particles involved. The Higgs boson couples to gauge bosons with an amplitude proportional

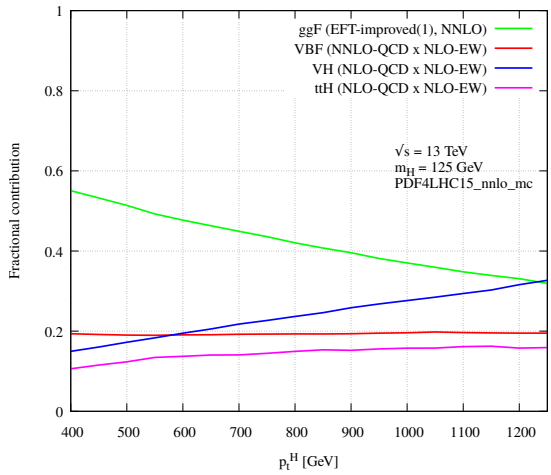


Figure 8.3: The relative contribution to the cumulative Higgs boson cross section due to the ggF (green), VBF (red), VH (blue) and top-quark fusion (magenta) production modes as a function of the Higgs boson p_{T} threshold. Figure taken from Ref. [72].

to the gauge boson mass squared, and to fermions with an amplitude proportional to the fermion mass [69].

The decay of the Higgs boson has to obey energy and momentum conservation. Consequently, the Higgs boson cannot decay to the heaviest known elementary particle, the top quark. While the combined mass of two $W^{\pm}$ or two Z bosons are larger than that of one Higgs boson, it is possible for the Higgs to decay into a pair of each boson. In such cases, one of the gauge bosons is a virtual particle whose existence suppresses the likelihood of the decay. As a result, the majority of Higgs bosons decay to the next most massive particle — the bottom quark (denoted as $H \to b\bar{b}$).

Despite being the most common Higgs boson decay mode, detection of $H \to b\bar{b}$ at the CMS experiment can be challenging due to the large number of other processes that also produce quarks. Distinguishing $H \to b\bar{b}$ events from other processes involving quarks requires sophisticated analysis techniques. In contrast, decay modes resulting in leptonic final states are much easier to detect due to their distinctive signatures in the detector. While the Higgs boson may directly decay into a pair of Z or $W^{\pm}$ bosons, the decay into massless photons are only possible via quantum-loop processes involving particles such as $W^{\pm}$ bosons and top quarks.

The probability of a certain decay occurring can be quantified by the *branching fraction*, defined as the fraction of particles (of a given type) that decay through an individual decay mode with respect to the total number of such particles which decay. Figure 8.4 displays the primary decay modes of the Higgs boson in order of their branching fractions as a function of the Higgs boson mass.

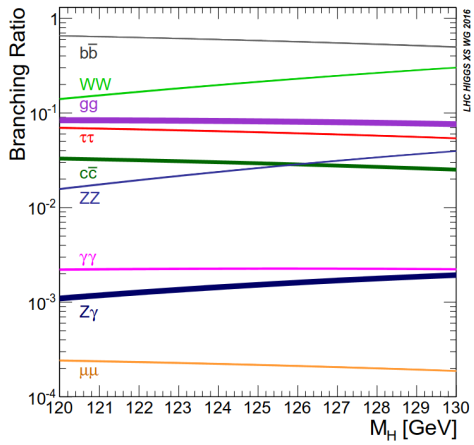


Figure 8.4: Higgs boson branching ratios around the mass range of 125 GeV. Figure taken from Ref. [71].

Chapter 9

Statistical methods

The physics analysis presented in this part of the thesis aims to compare the observed yield of collision data to the expected yield of simulation in order to establish the presence of a particular physics process. If an excess of events corresponding to the presence of the physics process of interest (so-called *signal*) is observed, the likelihood of it existing due to the signal model and not a statistical fluctuation is estimated. If there is no excess corresponding to the presence of the signal, a likelihood of excluding the signal model is instead estimated. In both cases, the parameter of interest is the amount of signal, represented by the *signal strength*.

The analysis is performed over a distribution that discriminates between the signal and other physics processes present in the data (collectively referred to as *background*). The distribution is referred to as the *summary statistic*, and can be any distribution such as the reconstructed jet mass or the discriminator of a machine learning classifier. Distributions of the summary statistic, referred to as *templates*, are created for each signal and background process. A nominal template corresponds to the nominal yield of the distribution, while up- and down-varied templates are created by varying the event yield in each bin up and down.

In this chapter, the relevant statistical methods applied in the analysis presented in Chapter 11 and 12 are discussed. First, the signal strength is described (Section 9.1). The likelihood function and maximum likelihood estimation are then introduced in Section 9.2 and 9.3, respectively. The use of test statistics is discussed in Section 9.4 and 9.5. Finally, the treatment of *systematic uncertainties* are discussed in Section 9.6.

9.1 Signal strength

The agreement between the observed yield and the theoretical expectation can be quantified by introducing a so-called signal strength parameter. The *signal strength modifier* (μ) is a multiplicative factor to the expected signal yield (N_{S}) according to

$$N_{\mathrm{exp}} = \mu \cdot N_{\mathrm{S}} + N_{\mathrm{B}}, \tag{9.1}$$

where N_{exp} and N_{B} are the total and background expected yields, respectively.

The signal strength modifier quantifies the strength of the signal sought in the observed data, and is chosen such that

$$\mu = \begin{cases} 0, \text{ background model;} \\ 1, \text{ background+signal model,} \end{cases} \tag{9.2}$$

where the background model (referred to as B in the following text) dictates that the observed data can be explained solely by known background processes, while the background+signal model (referred to as $B+S$) represents the presence of the signal process in addition to known background processes.

The signal strength modifier is adjusted during analysis to best fit the observed data by introducing it as a parameter of interest in the statistical fit. The value of μ that best fits the observed data is obtained by studying the likelihood function.

9.2 Likelihood function

The likelihood function returns the probability density for a given observed data sample, as a function of its statistical parameters. The likelihood, $\mathcal{L}$, of observing N_{obs} events in bin i, when the expected yield is N_{exp}, is described by the Poisson probability distribution. The combined Poisson probability for all bins is defined as the product of independent Poisson probabilities according to

$$\mathcal{L}(\{N_{\text{obs}}\}|\mu) = \prod_i \frac{(N_{\text{exp},i})^{N_{\text{obs},i}}}{N_{\text{obs},i}!} e^{-(N_{\text{exp},i})}, \tag{9.3}$$

where $\{N_{\text{obs}}\}$ represents all bins and $N_{\text{obs},i}$ denotes the observed events in bin i.

9.3 Maximum likelihood

The likelihood function, when evaluated for a given data sample, indicates the most likely parameter values based on the probability of observing the data sample. The statistical method used to estimate the parameters of the likelihood function is known as *maximum likelihood estimation*. In maximum likelihood estimation, the parameters are chosen to maximise the likelihood that the assumed model results in the observed data.

In Eq. 9.3, the value of μ that maximises this likelihood function for the observed data is termed the maximum-likelihood value μ_{ML}. For computational effectiveness, maximisation of the likelihood function can be performed as the minimisation of the corresponding negative log-likelihood function.

9.4 Test statistic

In order to investigate the measure of agreement between the observed data and a given hypothesis, a function of the measured variables called a *test statistic* is constructed [74]. To compare the compatibility of the observed data with the B and $B + S$ model hypotheses, the *profile likelihood ratio* described in Ref. [75] is used. The profile likelihood ratios is constructed by normalising the likelihood function by its maximum likelihood value, resulting in

$$\tilde{q}_\mu = -2\ln\frac{\mathcal{L}(\{N_{\text{obs}}\} \mid \mu)}{\mathcal{L}(\{N_{\text{obs}}\} \mid \mu_{\text{ML}})}. \tag{9.4}$$

$\tilde{q}_\mu$ is subject to the constraint $0 \leq \mu_{\text{ML}} \leq \mu$, where the lower bound is dictated by physics (ensuring a non-negative signal rate). In contrast, the upper bound is manually imposed to establish a one-sided confidence interval. In practical terms, this implies that upward fluctuations of the data, such that $\mu_{\text{ML}} \geq \mu$, are not considered as evidence against the $B+S$ model.

When quantifying the probability for the background to fluctuate and result in an excess of events as large or larger than what is observed in the data, as described in Ref. [75], the test statistic is constructed as

$$q_0 = -2\ln\frac{\mathcal{L}(\{N_{\text{obs}}\} \mid B \text{ model})}{\mathcal{L}(\{N_{\text{obs}}\} \mid B+S \text{ model})}, \tag{9.5}$$

$$= -2\ln\frac{\mathcal{L}(\{N_{\text{obs}}\} \mid N_B)}{\mathcal{L}(\{N_{\text{obs}}\} \mid \mu N_S + N_B)}, \tag{9.6}$$

$$= 2 \cdot (\mu N_S + N_B) \cdot \ln(1 + \mu N_S/N_B) - 2\mu N_S. \tag{9.7}$$

The probability is then evaluated as the p-value of the upward fluctuation of the B hypothesis. The p-value is often converted into the significance Z as described in Ref. [75]. In the limit of a large data sample, the profile likelihood ratio follows a non-central χ^2-distribution. As a result the significance may be approximated to

$$Z = \sqrt{q_0}. \tag{9.8}$$

9.5 Number-counting significance

As a higher significance relates to a higher likelihood of the $B+S$ model, Eq. 9.8 may be leveraged when optimising the event selection of an analysis. This is achieved by adjusting the event selection to maximise the significance. Such a procedure is referred to as maximising the *number-counting significance*. Before maximising, the $B+S$ model hypothesis is assumed, resulting in $\mu = 1$ and Eq. 9.8 becoming

$$Z = \sqrt{2 \cdot (N_S + N_B) \cdot \ln(1 + N_S/N_B) - 2N_S}. \tag{9.9}$$

9.6 Systematic uncertainties

Imperfections within the experimental setup and data processing methods are accounted for by assigning systematic uncertainties. These are added to the statistical analysis as a set of *nuisance parameters* $\vec{\theta}$. The nuisance parameters can alter event yields in each bin. When the uncertainties are taken into account the event yields become functions of $\vec{\theta}$: $N_{\text{exp}} \rightarrow N_{\text{exp}}(\vec{\theta})$.

Different sources of uncertainty, corresponding to different nuisance parameters, can be treated as fully correlated (100% correlation), anti-correlated (–100%), or independent (0%). The appropriate assignment of correlations depends on the specific uncertainties present. Partially correlated uncertainties are treated by splitting them to fully correlated or anti-correlated sub components.

The nuisance parameters are added to Eq. 9.3 by adjusting the function to depend on both the signal strength modifier μ and the full set of nuisance parameters $\vec{\theta}$. The adjusted function is described by

$$\mathcal{L}(\{N_{\text{obs}}\}|\mu,\vec{\theta}) = \prod_i \frac{(\mu N_{\text{exp},i}(\vec{\theta}))^{N_{\text{obs},i}}}{N_{\text{obs},i}!} e^{-(\mu N_{\text{exp},i}(\vec{\theta}))} f(\theta|\tilde{\theta}), \tag{9.10}$$

where $f(\theta|\tilde{\theta})$ denotes the probability density function (pdf) of the nuisance parameters and $\tilde{\theta}$ is the default value of the nuisance parameter.

If the nuisance parameter affects all bins in the same way, such that the effect equals a multiplication of the total event yield by a given factor, it is referred to as a *normalisation uncertainty*. In contrast, if the effect is defined separately for each bin, such that the nuisance parameter can affect both the shape and normalisation of the distribution, it is called a *shape uncertainty*.

Chapter 10

Existing Higgs boson measurements

The discovery of the Higgs boson by the CMS and ATLAS experiments in 2012 marked a monumental breakthrough in particle physics, confirming the existence of the last missing elementary particle predicted by the SM. The construction of the SM spanned 60 years of theoretical and experimental work [1, 2]. In the years following the discovery, significant progress has been made in expanding our understanding of the particle. In this chapter, the discovery of the Higgs boson is presented together with current knowledge of its properties. The chapter concludes with a consideration of remaining unanswered questions relating to the boson.

10.1 Discovery of the Higgs boson

At the time of the Higgs boson discovery, the CMS experiment was targeting five decay modes with the aim of achieving an observed statistical significance of 5.0 standard deviations (σ). Individually, the two most sensitive modes ($H \to \gamma\gamma$ and $H \to ZZ \to 4l$) achieved 4.1 observed σ and 3.2 observed σ, respectively [33]. Combined, however, a local significance of 5.0 σ (at a mass near 125 GeV) and a global significance of 4.6 σ was achieved [33] — signalling the production of a new, Higgs boson-like, particle. The invariant mass distributions of the two decay modes mentioned above are presented in Fig. 10.1.

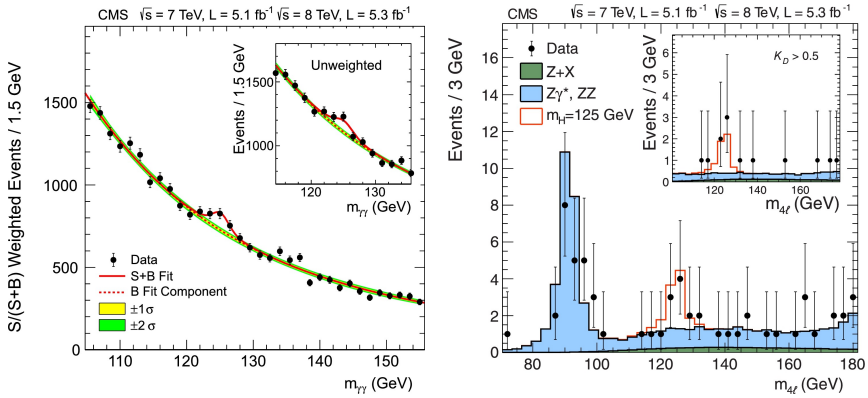


Figure 10.1: The diphoton (left) and four-lepton (right) invariant mass distribution of the $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ analyses, respectively. Figures taken from Ref. [33].

The existence of a peak over the background expectation in each figure indicates the presence of a new particle, and the position signifies the mass of that particle. At the time of the Higgs boson discovery, a significant excess in the range of $121.5 < m_{\mathrm{H}} < 128\,\mathrm{GeV}$ could not be excluded at a 95% confidence level (CL) [33]. This was suggestive of a Higgs boson-like particle in this mass range. The combined signal strength modifier μ between all production and decay modes was then measured to be 0.87 ± 0.23 [33]. The CMS experiment has since refined this measurement to $\mu = 1.002 \pm 0.057$ [69], which aligns with the SM expectation. The uncertainties in the new measurement correspond to an improvement in precision by more than a factor of 4 compared with what was achieved at the time of discovery.

10.2 Production and decay modes

As discussed in Section 8.5, the SM predicts that the strength of the Higgs boson's couplings scale with the mass of the particles it couples to. As a result, the strength of the couplings can be precisely determined by inserting the previously measured particle masses and effectively provide SM expectations. In this way, experimental measurement of the couplings to each individual particle provides direct tests of the SM. Moreover, these measurements impose stringent constraints on theories beyond the SM, which typically predict different coupling strengths.

To date, all measurements of coupling strengths performed by the CMS experiment have been consistent with the expectations of the SM. Measurements from data recorded in Run 2 by the CMS detector are presented in Fig. 10.2. The figure displays μ extracted for various production and decay modes. The production modes presented in Section 8.4 are all observed with a significance of 5 σ or larger. The measurements also show that the Higgs boson directly couples to bottom quarks (5.6 σ) and tau leptons (5.9 σ) [69].

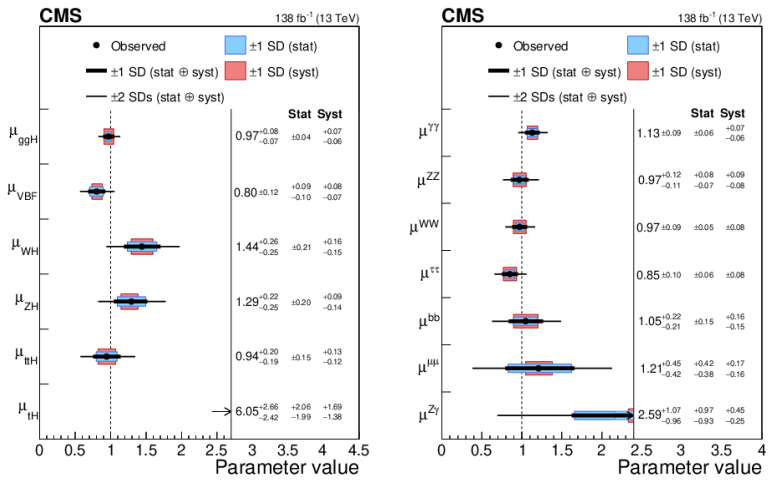


Figure 10.2: The signal strength modifier extracted for various production (left) and decay (right) modes. The thick and thin black lines indicate the 1-σ and 2-σ confidence intervals, respectively, with the systematic and statistical components of the 1-σ interval indicated by the red and blue bands, respectively. The vertical dashed line at unity represents the value of the SM. Figures taken from Ref. [69].

10.2.1 Self-coupling

The SM predicts that, as Higgs bosons have mass, they interact with themselves. The probability of self-interaction is determined by properties of the Higgs field — properties describing conditions immediately after the Big Bang. Increasing knowledge of the Higgs boson self-interaction may therefore advance our understanding of the dynamics of the early universe. Moreover, some theories explain the abundance of matter over antimatter

by requiring that the Higgs self-interaction diverges from the SM prediction [76].

The most promising and direct way to measure the Higgs self-coupling involves identifying a pair of Higgs bosons in the final state. Unfortunately, the rate of such events is approximately 1000 times lower than that of single Higgs boson production [77], rendering its measurement at the LHC challenging. As of the latest analysis using data recorded by the CMS detector during Run 2, the cross section for the production of Higgs boson pairs is measured to be below 3.4 times the SM expectation at a 95% CL [69].

10.3 Beyond the Standard Model

Theories of physics beyond the SM modify the predicted rate of Higg boson production and decay modes. To probe such deviations from the predictions of the current SM theory, the κ framework [78] is used. In this framework, a set of parameters are introduced that affect the Higgs boson coupling strengths without altering the kinematic distributions of Higgs boson interactions. The product of the cross section and the branching fraction for an individual measurement is parameterised in terms of the multiplicative *coupling strength modifier* κ. In the SM, all κ values are equal to one.

The coupling modifiers for various Higgs boson interactions are measured using data recored in Run 2 by the CMS detector. The results are presented in Fig. 10.3 — all measured values are compatible with the SM expectations within 1.5 σ [69].

10.4 Unanswered questions

While understanding of the Higgs boson has advanced in the years since its discovery, current knowledge remains incomplete. For example, many properties of the Higgs boson have been determined with accuracies around 10%; this precision is insufficient to examine theories that differ only slightly from the SM. In the future, the precision of these measurements can be increased by recording more collision data.

A prominent question in Higgs boson research concerns self-interaction.

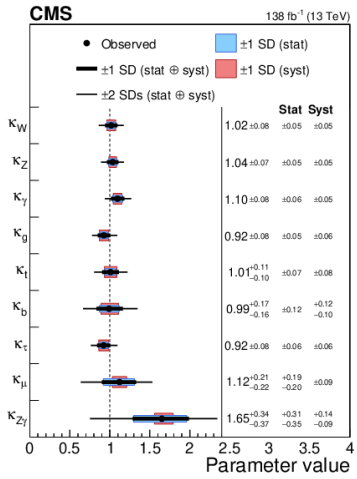


Figure 10.3: Coupling modifiers extracted for various processes. The thick and thin black lines indicate the 1-σ and 2-σ confidence intervals, respectively, with the systematic and statistical components of the 1-σ interval indicated by the red and blue bands, respectively. The vertical dashed line at unity represents the values of the SM. Figure taken from Ref. [69].

As discussed earlier, advances in this area have the potential to provide insights into the early universe and the matter-antimatter imbalance. Moreover, if self-interaction differs substantially from the SM prediction, it may suggest that the universe does not exist in the energy state currently assumed. Related to Higgs boson couplings, couplings to lighter particles such as muons and charm quarks are yet to be observed with 5 σ. Any discrepancies arising from studies of these couplings, or other precision measurements, may suggest new physics beyond the SM.

Finally, the nature of the Higgs boson is explored in several theories that extend the SM, discussed in detail in Ref. [79]. Some suggest that the Higgs boson, like the proton, is not fundamental but is composed of other particles [80]. Other theories predict the existence of multiple Higgs bosons, each sharing similarities but differing in characteristics such as charge or spin. The Higgs boson discovered in 2012 has zero spin and no electric charge, but other Higgs particles could have different characteristics. Some phenomena that could be explained by additional Higgs particles include dark matter, neutrino masses and the imbalance of matter and antimatter in the universe [79].

Chapter 11

Search for boosted Higgs boson production

Experimental exploration of the Higgs boson is still in its infancy, as eluded to in Chapter 10. Analyses aiming to (a) increase the precision of previously observed interactions and (b) detect further, unobserved interactions are still required to enhance current understanding of the Higgs boson. This chapter presents an analysis pertaining to unobserved interactions, which aims to expand the search for Higgs bosons produced with high transverse momentum (so-called *boosted* Higgs bosons) via VBF and ggF. As boosted Higgs boson production can be sensitive to physics beyond the SM (particularly momentum-dependent anomalous couplings [81]), its study has become an important part of the CMS physics programme. The lead investigator of the analysis presented in this chapter is Dr. Jennet Dickinson — the author of this thesis worked on this analysis as a member of Dr. Dickinson's group.

As most proton-proton interactions at the LHC occur at relatively low energies compared to the centre-of-mass energy of the colliding protons, the production of boosted Higgs bosons is rare. To circumvent this challenge, this analysis targets the $H \to b\bar{b}$ decay mode because of its large branching fraction (Section 8.5). This analysis provides, for the first time, insight towards boosted Higgs bosons decaying through the $H \to b\bar{b}$ decay mode in tandem with the VBF production mode. Existing searches for the boosted $H \to b\bar{b}$ process by the CMS [82] and ATLAS [83] experiments have focussed on inclusive Higgs boson production. Due to the dominance of the ggF production mode, these searches are primarily sensitive to Higgs cou-

plings to top quarks and gluons. In contrast, as a consequence of targeting the VBF production mode, the analysis presented here allows analysis of Higgs couplings to vector bosons. The analysis is performed with data collected from proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$ with the standard trigger strategy in Run–2 and an integrated luminosity of 138 fb^{-1} [84–86].

11.1 Analysis strategy

The signal of the analysis is the production of Higgs bosons through VBF and ggF and their subsequent decay into bottom quark-antiquark pairs (illustrated for the VBF production mode in Fig. 11.1). The background is all other physics processes in the data, with the dominant contribution stemming from QCD multijet production.

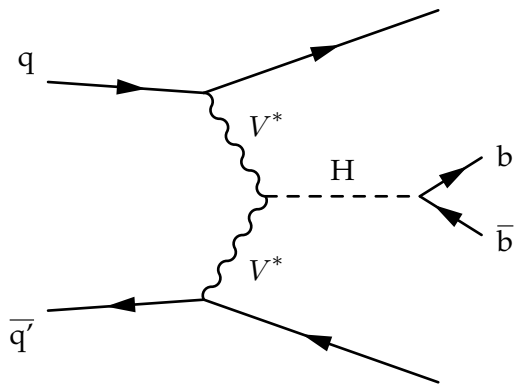


Figure 11.1: Feynman diagram of the VBF production of a Higgs boson and its subsequent decay into a bottom quark-antiquark pair. q denotes a quark, V^* a Z or $W^\pm$ boson, H a Higgs boson and b a b quark. A letter with a bar denotes an anti-particle.

As detailed in Section 4.4.1, when massive particles (such as the top quark, $W^\pm$, Z and Higgs bosons) decay hadronically at low p_T, the decay products are reconstructed as separate jets due to the spatial separation between the partons of the decay. In contrast, when massive particles have a p_T greatly exceeding their mass, the hadronic decay results in highly collimated decay products. These products can be more efficiently reconstructed as a single large-radius jet, as is the case for boosted $H \to b\bar{b}$. This

is exemplified in Fig. 11.2, where a comparison of a hadronically decaying boson produced with low and high p_{T} is provided.

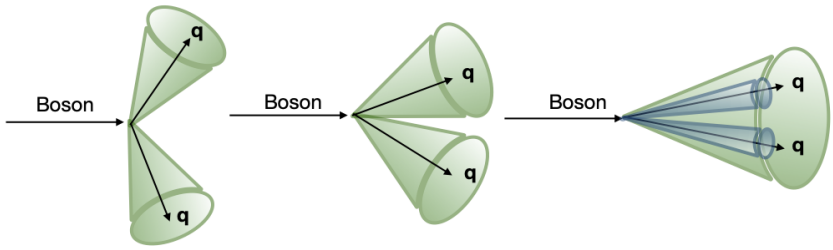


Figure 11.2: Illustration of a boson produced with increasing amounts of p_{T} (left to right) decaying to two quarks. At high p_{T} (right most), the two quarks are emitted collimated, giving rise to one single large-radius jet containing a *subjet* for each quark. Figure taken from Ref. [87].

It is possible to discriminate between decaying massive particles and QCD multijet production on the basis of the properties of these large-radius jets. One such property is the jet mass. The jet mass is defined as the invariant mass of the jet constituents' four-vector sum, and its distribution varies between physics processes. If the decay products of a quickly decaying unstable particle (so-called *resonant* particle) are captured by a jet, the jet mass distribution will centre around the mass of the particle. In contrast, the jet mass distribution of a non-resonant background does not. As a result, analysis of the jet mass distribution allows the resonant particle and the non-resonant background to be distinguished. In the analysis presented in this chapter, the main background comprises non-resonant QCD multijet production. In order to discriminate between boosted $H \to b\bar{b}$ (resonant) and QCD multijet production (non-resonant), the jet mass is selected as the summary statistic.

The analysis presented in this chapter involves a series of steps. These steps are summarised below, and then described in more detail in the following sections.

1. **Signal jet selection**. The large-radius jet most likely to originate from the signal process is selected for each event, and termed the *Higgs candidate jet*.

2. **Event selection**. Consecutive selection steps (based on the event topology of boosted $H \to b\bar{b}$) are applied in order to discard back-

ground events whilst retaining signal events. This maximises the signal efficiency.

3. **Background estimation**. The shape and normalisation of the distributions generated by background processes are estimated in order to make a precise comparison between signal and background events.

4. **Evaluation of systematic uncertainties**. Systematic uncertainties arising from various sources, such as detector performance and theoretical modelling, are quantified. These uncertainties define the precision of the final results.

5. **Statistical analysis**. Statistical methods are used to compare the expected signal and background distributions in order to extract the observed signal strength.

11.2 Signal jet selection

As the decay products of boosted massive particles are often reconstructed as a single large-radius jet, the Higgs candidate jet of each event is selected as the large-radius jet most likely to contain two b quarks. The implementation of this selection is explained below.

11.2.1 Higgs candidate jet

The likelihood of a large-radius jet containing two b quarks is described by the DeepDoubleBvL-v2 (DDB) *tagger* discriminant [88], where a larger tagger DDB discriminant corresponds to a larger likelihood. The task of identifying the origin of jets is referred to as *jet tagging*, and the algorithm employed for this purpose is called a *jet tagger*. In the CMS experiment, a variety of algorithms using modern machine learning methods, such as deep neural networks (DNN), have been developed for this task.

The DDB tagger is a DNN trained to distinguish (a) large-radius jets originating from the decay of a boosted object to a bottom quark-antiquark pair from (b) jets originating from QCD multijet production. As illustrated in Fig. 11.3, the DDB tagger shows a large improvement in performance relative to previous taggers; for a likelihood of misidentifying QCD jets as

signal jets of 1%, there is a 75% likelihood of correctly identifying signal jets.

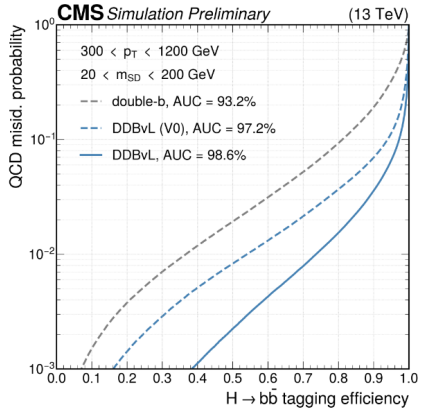


Figure 11.3: Performance of DeepDoubleBvL-v2 identification algorithm (blue line) demonstrating the probability of misidentifying QCD jets as a function of the $H \to b\bar{b}$ tagging efficiency in terms of a ROC curve [88]. For reference, the performance of a previous generation classifier, the double-b tagger (gray dashed), is shown as well as the performance of an earlier prototype of the DeepDoubleBvL classifier (blue dashed) previously shown in Ref. [89] and used in Ref. [90]. Figure taken from Ref. [88].

11.2.2 Soft drop jet mass

After selecting the Higgs candidate jets, a jet *grooming* algorithm is applied to obtain the summary statistic: the groomed jet mass. It is necessary to apply the grooming algorithm in order to remove soft and wide-angle contributions from the jet mass that would otherwise bias the results. These contributions stem from initial-state radiation, underlying event and pile-up. While several jet grooming algorithms exist, the "soft drop" algorithm [91] is used in this analysis.

The "soft drop" algorithm works by iteratively declustering a jet into two subjets using the Cambridge–Aachen algorithm [92, 93]. The energy distributions of the two subjets are evaluated, and if a specific criterion is met, the sum of the two subjets is retained as the final jet. If the criterion is not met, the softer subjet is removed and the procedure is repeated. The criterion is defined as

$$\frac{min(p_{T,1}, p_{T,2})}{p_{T,1} + p_{T,2}} > z_{\text{cut}} \left(\frac{\Delta R_{1,2}}{R_0} \right)^{\beta}, \tag{11.1}$$

where R_0 is the distance parameter of the jet, $\Delta R_{1,2}$ is the angular distance between the two subjets, and β and z_{cut} are parameters of the procedure. The default parameters used by the CMS experiment are $\beta = 0$ and $z_{\text{cut}} = 0.1$.

The mass of the jet returned by the "soft drop" algorithm, here denoted as m_{SD}, is used as the summary statistic of the analysis. Specific corrections to m_{SD} are applied to correct for residual p_{T}-dependence. These are evaluated centrally by CMS and result in a jet mass resolution of approximately 0.1, as documented in Ref. [94].

11.3 Event selection

To maximise signal efficiency, consecutive selection steps are applied to identify boosted $H \rightarrow b\bar{b}$ events. These selections are chosen based on their ability to discriminate effectively between signal and background jets. First, the decay mode of the signal process is targeted. Next, focus is placed on distinguishing the VBF and ggF production modes. A detailed description of the selection steps is provided below.

11.3.1 Trigger selection

Events containing large-radius jets are selected by an array of HLT triggers that select events based on the presence of hadronic activity. In addition, a jet tagger targeting jets originating from b quarks is employed in one of the triggers. The efficiency of each trigger is computed using the reference method as described in Section 5.1.3. The implementation of the reference method is the same as that described in Section 6.3.3.

The trigger efficiency curve as a function of collision data recorded in 2016 is presented in Fig. 11.4. The efficiency of the logical 'OR' of all triggers used is 100% at approximately 500 GeV, and 95% at around 450 GeV. The application of the jet tagger allows the efficiency to plateau at earlier p_{T}

than if the trigger expression was soley based on triggers selecting events based on jet p_{T} and H_{T}. In order to avoid a trigger bias (which might arise by selecting jets on the turn-on of the trigger efficiency curve) Higgs candidate jets are required to have $p_T \geq 450\,\mathrm{GeV}$.

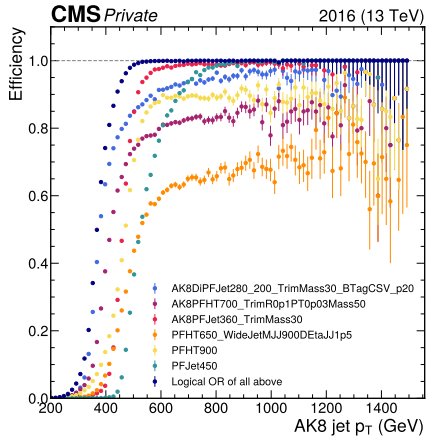


Figure 11.4: Trigger efficiency of the HLT triggers used to select events for collision data collected in 2016, as a function of AK8 jet p_{T}. A logical 'OR' of all the triggers is displayed in dark blue and reaches a plateau around 500 GeV. The uncertainties are entirely statistical and computed as Clopper-Pearson intervals.

11.3.2 Baseline selection

The Higgs candidate jets are corrected with detector response corrections derived from simulation to adjust the measured response of reconstructed jets towards that of generated jets on average. The same corrections are applied to jets reconstructed from simulation and collision data, with *in-situ* corrections (such as JES calibration) applied to the latter. Before clustering, pileup is mitigated using the PUPPI technique (Section 4.4.3).

All Higgs candidate jets are required to have $|\eta| < 2.5$ and $m_{\mathrm{SD}} > 40\,\mathrm{GeV}$. The jets are also required to pass a quality criterion that discards badly reconstructed and noise jets following the procedure outlined in Ref. [94], and to be separated from all photons or charged leptons by an angular distance of $\Delta R > 0.8$.

The DeepCSV algorithm [95], a neural network trained to identify small-

radius jets originating from b quarks, is used to reduce $t\bar{t}$ background contamination. Events are discarded if any of the four leading small-radius jets in the hemisphere opposite the Higgs candidate jet are b-tagged (a typical feature of $t\bar{t}$ events). This selection discards around 40% of the background $t\bar{t}$ events. Since no neutrinos are expected in the final state, events with MET > 140 GeV are also discarded.

11.3.3 Jet ρ

A selection is made on the QCD dimensionless scaling variable ρ, defined as

$$\rho = 2 \ln \left(\frac{m_{\text{SD}}}{\text{jet } p_{\text{T}}} \right). \tag{11.2}$$

A detailed study performed in Ref. [82] found that to avoid instabilities at the edges of the ρ distribution, events are required to have $-6 < \rho < -2.1$. Jets below $\rho = -6$ are not considered in order to avoid the non-peturbative regime of the m_{SD} calculation. Similarly, jets with a ρ value above -2.1 are discarded because they are impacted by the *finite cone effects* of jet clustering: the p_{T} of the Higgs boson is too low for the two b quarks of the decay to be captured by a single large-radius jet. This selection is fully efficient for the Higgs signal, but reduces the upper edge of the fitted m_{SD} range in the lowest two p_{T} bins.

11.3.4 Jet substructure

The Higgs candidate jets are required to be consistent with the 2-prong substructure of the $H \to b\bar{b}$ decay mode. The term *N-prong* describes a specific decay signature, where N in this case represents the number of subjets that are part of the large-radius jet. In the case of boosted $H \to b\bar{b}$, each b quark stemming from the Higgs boson decay produces a separate jet that can be identified as a subjet. This is illustrated in Fig. 11.2, where the subjets are coloured blue and are contained within the large-radius jet, coloured green.

The identification of jets with 2-prong substructures is achieved by selection based on the jets' *energy correlation function* (ECF) [96, 97]. The ECF

uses information about the energies and pair-wise angles of particles within a jet in order to identify the N-prong substructure, and is defined as

$$ {}_{o}e_{N}^{\beta} = \sum_{1<\ldots<i_N} \prod_{a=1}^{n} z_{i_a} \times \min \left(\prod_{\text{pairs}\{i_1,\ldots,i_n\}}^{o} \Delta R^{\beta} \right), \tag{11.3} $$

where i is a particle within the jet, N is the number of subjets considered, z the energy fraction of the particle, ΔR is the angular distance between two particles, o denotes the order of the angular factor and β is related to angular weighting.

A detailed study performed in Ref. [97], found that the following ratio of ECFs is effective when distinguishing boosted Higgs boson decays

$$ N_2^1 = \frac{{}_2e_3^1}{({}_1e_2^1)^2}. \tag{11.4} $$

In order to minimise the dependence of the substructure selection on the p_{T} and mass of the Higgs candidate jets, a variant of N_2^1 is defined as a function of p_{T} and ρ. This procedure is referred to as *decorrelating* and the updated version of N_2^1 is referred to as a *designed decorrelated tagger* (DDT) [98]. The procedure is derived for a specific background efficiency (ϵ); the application of the DDT tagger results in the removal of $100\% - \epsilon\%$ of the QCD events. For a given p_T and ρ, the value of X that results in a certain ϵ, when requiring $N_2^1 < X$, can be written as $X_\epsilon(p_T, \rho)$. The transformation to decorrelate N_2^1 is then defined as

$$ N_2^{1,\mathrm{DDT}}(p_T, \rho) = N_2^1 - X_\epsilon(p_T, \rho). \tag{11.5} $$

In this analysis, ϵ is selected to be 26% in order to maximise the signal sensitivity. The value is estimated by maximising the number-counting significance as described in Section 9.5. A 2D map of $N_2^{1,\mathrm{DDT}}$ is constructed using a sample of simulated QCD events, and is presented in Fig. 11.5 for data recorded in 2016. As a result, by requiring that the Higgs candidate jets have $N_2^{1,\mathrm{DDT}} < 0$, 74% of QCD jets wrongly identified as originating from the signal process are discarded. In contrast, only about 48% of signal jets are removed.

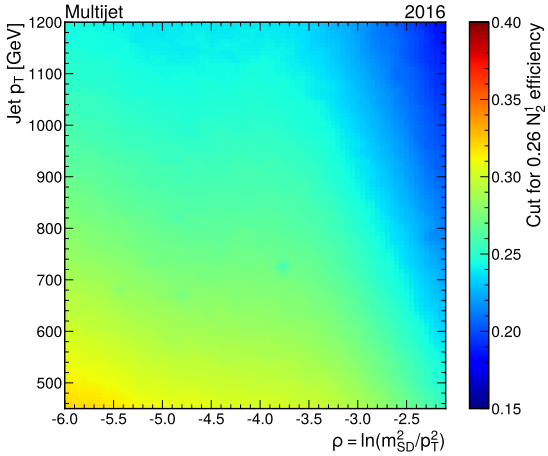


Figure 11.5: Transformation map of $N_2^1 \rightarrow N_2^{1,\mathrm{DDT}}$, corresponding to a background efficiency of 26%, as a function of p_T and ρ. The z-axis correspond to $X_\epsilon(p_T, \rho)$ in Eq. 11.5. The inflection in the distribution around $\rho = -2$ stems from the finite cone effect, and represents a natural limit to the mass of a QCD jet for a given p_T.

11.3.5 Higgs boson production mode

The final event selection separates Higgs candidate jets into two categories: those originating from the ggF and from the VBF production mode. As the VBF mode is rare compared to the ggF mode, the categorisation process focuses on maximising the number of VBF events identified.

To isolate the VBF production mode, the two forward jets characterising the VBF mode (as discussed in Section 8.4) are targeted. To be identified as a VBF event, two leading small-radius jets are required to have an angular separation of $|\Delta\eta| > 3.0$ and an invariant mass $m_{\mathrm{jj}} > 1$ TeV. The threshold values are chosen to optimise sensitivity to the VBF mode by maximising the number-counting significance. Events not meeting these criteria, or that contain fewer than two small-radius jets, are classified as ggF events. This selection correctly identifies more than 90% of jets (remaining from the previous selections) as originating from the ggF and VBF production modes.

11.4 Background estimation

The background considered in this analysis comprises physics processes classified as non-resonant and resonant. The QCD multijet production is the dominant non-resonant background due to its large cross section (Section 2.3). The next largest non-resonant background stems from top quark processes. In events where a boosted top quark is produced, it is possible that a large-radius jet will capture only two of the three prongs of the top quark decay. As a result, the jet passes the $N_2^{1,\mathrm{DDT}}$ selection. In addition, the presence of a b quark in top quark decays makes the jet likely to pass the DDB selection, leading to misclassification. Since the large-radius jet fails to capture all decay products of the top quark, the invariant mass of the jet constituent does not centre around the top quark mass. In this way, top quark production may erroneously contribute to the non-resonant component of the background.

Significant resonant backgrounds arise from $W^{\pm}$ and Z + jets processes, where the 2-prong decay of a boosted vector boson is mistaken for a signal jet. Electroweak $W^{\pm}$ and Z production contributes mainly to the background in the VBF category because of the presence of two forward jets.

Accurate estimation of the background processes' shape and normalisation is essential because their combined yields dominate the total event yield. For the analysis presented in this chapter, with the exception of the QCD background, the shapes of all background processes are estimated from simulation. The normalisation is further estimated from simulation for all background processes except the QCD and the top quark background. The estimations of these two are described in detail in the following text.

11.4.1 Signal and control regions

To facilitate background estimation, Higgs candidate jets are divided into a *signal region* and a *control region*. These regions are identified by being either above or below a DDB tagger discriminant value, and are chosen to optimise sensitivity to VBF by maximising the number-counting significance. Higgs candidate jets with a DDB tagger discriminant ≥ 0.64 are allocated to the signal region, while Higgs candidate jets with a DDB tagger discriminant < 0.64 are allocated to the control region. The discriminant threshold corresponds to a 40% likelihood of correctly identifying signal

jets and a 0.5% likelihood of misidentifying QCD background jets as signal jets.

Figure 11.6 shows the relative contribution of each Higgs boson production mode discussed in Section 8.4, to the total Higgs signal yield in the signal and control regions. The ggF and VBF signal regions are dominated by jets originating from the ggF and VBF production modes, respectively. Furthermore, the Higgs boson signal in the ggF and VBF signal regions are more than 60% and 75% pure, respectively, in the target production mode. Only about 0.3% of QCD jets are selected into the signal region of either production mode.

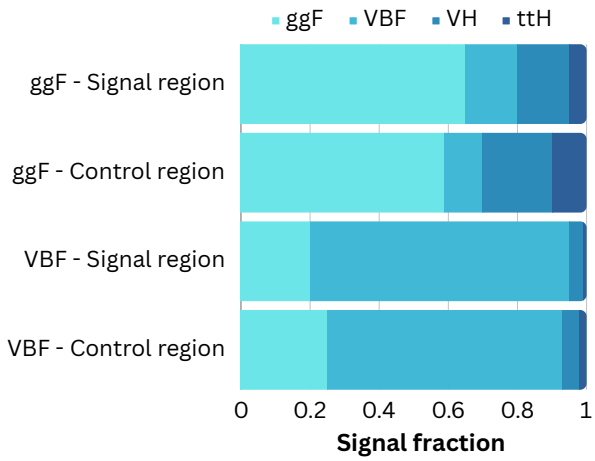


Figure 11.6: Relative contribution of each production mode to the total Higgs signal yield in the signal and control regions after applying all event selections described above. The fractions are computed using simulated samples dedicated to each production mode. The ggF and VBF categories are shown separately.

11.4.2 QCD background

The predicted shape and normalisation of the QCD background in the signal region is derived using the background-enriched control region. The predicted number of QCD events in bin i of the signal region is computed as

$$N_{\mathrm{S}}^{i} = R_{\mathrm{S/C}}^{\mathrm{simulation}}\; N_{C}^{\mathrm{data},i}\; F_{\mathrm{S/C}}^{\mathrm{QCD},i}\; F_{\mathrm{S/C}}^{\mathrm{data},i}, \tag{11.6}$$

where $R_{\mathrm{S/C}}^{\mathrm{simulation}}$ is the ratio of simulated QCD events in the signal region to simulated QCD events in the control region and $N_{C}^{\mathrm{data},i}$ is the number of observed data events in bin i in the control region. $F_{\mathrm{S/C}}^{\mathrm{QCD},i}$ and $F_{\mathrm{S/C}}^{\mathrm{data},i}$ are *transfer factors* (polynomial functions) that quantify the change of the background shape between the signal and control regions.

$F_{\mathrm{S/C}}^{\mathrm{QCD},i}$ controls potential shape effects introduced by the DDB tagger selection, and is derived with a dedicated fit to the m_{SD} distribution of simulated QCD events. $F_{\mathrm{S/C}}^{\mathrm{data},i}$ accounts for discrepancies in DDB tagger performance between collision data and simulation, and is derived with a simultaneous fit to the m_{SD} distribution of the collision data in the signal and control regions.

For each transfer factor, the optimal number of free parameters used by the polynomial is determined by a Fisher F-test [99]. A low order polynomial with p_1 parameters is taken as the baseline function. An alternative function with $p_2 > p_1$ parameters is tested against the baseline, and adopted as the new baseline if it provides a significantly better goodness of fit (test statistic that describes how well a statistical model fits a set of observations).

11.4.3 Top quark background

The normalisation of the top quark background is estimated by selecting a sample targeting the event topology of single-μ $t\bar{t}$ events (illustrated in Fig. 11.7). To target the specific event topology, the following selections are applied consecutively.

1. **To target the muon of the leptonically decaying W boson**, events are required to contain exactly one muon with $p_T > 55\,\mathrm{GeV}$ and $|\eta| < 2.1$.
2. **To target the two partons from the hadronically decaying W boson**, at least one large-radius jet with $p_T > 400\,\mathrm{GeV}$, $|\eta| < 2.4$ and $N_2^{1,\mathrm{DDT}} < 0$ is required. The large-radius jet must be separated from the muon by an angular distance of $\Delta\phi > 2\pi/3$.

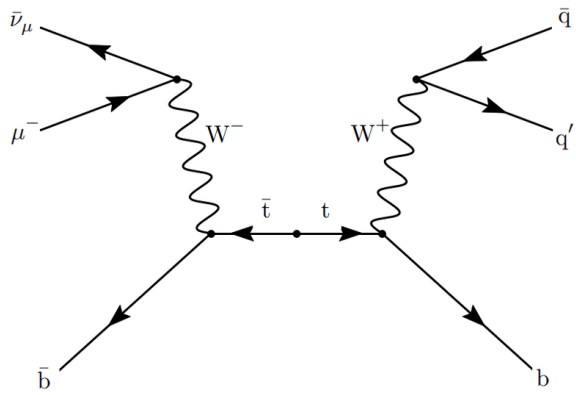


Figure 11.7: Feynman diagram of a single-μ $t\bar{t}$ event.

3. **To target the b quark from either of the decaying W bosons**, at least one small-radius b-tagged jet with $p_T > 50\,\text{GeV}$ and $|\eta| < 2.5$ is required.

The single-μ $t\bar{t}$ sample is separated into a $t\bar{t}$ signal region and a $t\bar{t}$ control region according to the same DDB tagger discriminate criterion as described in Section 11.4.1. The sample is then used to derive two *scale factors* (SF). The first SF constrains the number of events according to

$$SF_{\text{N}} = \frac{N^{\text{data}}_{S \text{ and } C} - N^{\text{simulation,non}-t\bar{t}}_{S \text{ and } C}}{N^{\text{simulation},t\bar{t}}_{S \text{ and } C}}, \tag{11.7}$$

where the denominator is the simulated $t\bar{t}$ events in the signal and control region and $N^{\text{simulation,non}-t\bar{t}}_{S \text{ and } C}$ represents the simulated events from the remaining simulated samples in the signal and control region.

The second SF accounts for the efficiency of the DDB tagger selection according to

$$SF_{\epsilon} = \epsilon_{\text{data}}/\epsilon_{\text{simulation}}, \tag{11.8}$$

where the numerator and denominator is defined as

$$\epsilon_{\text{data}} = \frac{N_S^{\text{data}} - N_S^{\text{simulation,non}-t\bar{t}}}{N_{S \text{ and } C}^{\text{data}} - N_{S \text{ and } C}^{\text{simulation,non}-t\bar{t}}}, \tag{11.9}$$

$$\epsilon_{\text{simulation}} = \frac{N_S^{\text{simulation},t\bar{t}}}{N_{S \text{ and } C}^{\text{simulation},t\bar{t}}}. \tag{11.10}$$

Both SFs are fitted simultaneously with a maximum likelihood fit. The overall predicted number of top quark events in the signal region is then derived as the number of simulated $t\bar{t}$ events in the signal region scaled by $(SF_{\text{N}} \times SF_{\epsilon})$.

11.5 Systematic uncertainties

Systematic uncertainties arise from various sources, including experimental measurements and theoretical assumptions. The impact of each systematic uncertainty is quantified by evaluating its effect on the final results as described in Section 9.6. When combined, the total systematic and statistical uncertainties determine the final uncertainty of the result. In the following text, the main sources of systematic uncertainties in this analysis are discussed.

11.5.1 Trigger uncertainty

Discrepancies between simulation and collision data may result in differences between their respective trigger efficiencies. For example, if simulated jets are consistently reconstructed with a lower p_{T} than their truth p_{T}, the simulated trigger efficiency ($\epsilon_{\text{simulation}}$) curve may reach its plateau slightly earlier than is observed with collision data (ϵ_{data}). The difference between the trigger selection is therefore corrected for, and the limited precision of the correction is taken into account by assigning a systematic uncertainty in the form of a shape uncertainty (Section 9.6). The correction is applied to the simulated Higgs candidate jets with a SF derived as

$$SF(m_{\text{SD}},\ p_T) = \frac{\epsilon(m_{\text{SD}},\ p_T)_{\text{data}}}{\epsilon(m_{\text{SD}},\ p_T)_{\text{simulation}}}, \tag{11.11}$$

where the efficiency for simulation is computed with simulated samples of QCD multijet production.

The SF for the data-taking period of 2016 is presented in Fig. 11.8 together with their associated uncertainty (ΔSF). The uncertainty is propagated to the analysis as a systematic uncertainty by creating one up- and one down-varied template, as described in Section 9.6. To create the up- and down-varied templates the nominal distribution of the summary statistic is multiplied by $1 + \Delta SF$ and $1 - \Delta SF$, respectively.

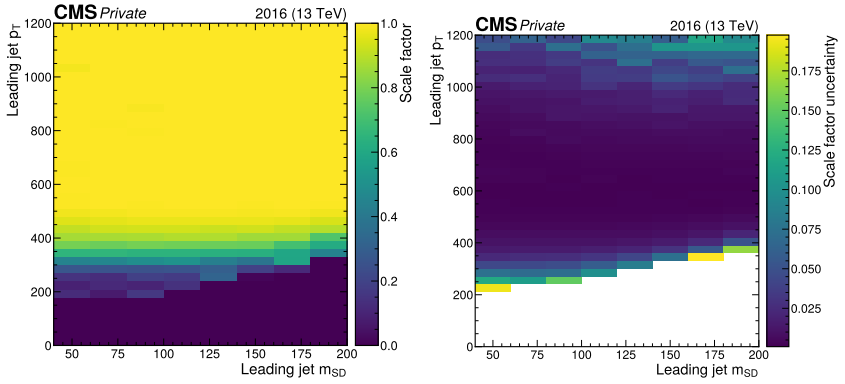


Figure 11.8: Trigger efficiency SF (left) and its associated statistical uncertainty (right) for the 2016 data-taking period as a function of AK8 jet p_{T} and m_{SD}.

11.5.2 Groomed jet mass and substructure uncertainties

When interpreting the results of an analysis involving the groomed jet mass, it is critical to account for the grooming algorithms effect on the jet mass scale (JMS) and jet mass resolution (JMR). In order to compute these properly, the signal process must be well isolated from other processes, such as QCD multijet production. While it is challenging to isolate processes based on the jets of the event, the desired process can be efficiently selected using leptons. The energy deposited by electrons is more tightly clustered than that deposited by hadrons, leaving a cleaner signature in the calorimeter. For muons, the large amount of material between the interaction point and the muon chambers acts as an absorber for almost all SM particles except muons; if a signal is detected in the muon chambers, it is likely a muon. In addition, fewer other SM processes mimic the signatures of leptonically decaying particles than mimic the signatures of hadronic

channels. Particle decays producing at least one lepton are therefore effective tools for jet calibration.

As a consequence, the JMS and JMR are estimated with the use of a proxy process that mimics the event topology of boosted $H \to b\bar{b}$: boosted $W \to q\bar{q}$. This is achieved by selecting a single-μ $t\bar{t}$ sample, largely following the selection steps outlined in Section 11.4.3. The muon from the leptonically decaying W boson is targeted in order to select the event, while the groomed jet mass from the decay of the hadronically decaying W boson is used to determine the JMS and JMR.

The leading large-radius jet of each event is selected as the *hadronically decaying W candidate*. The candidate jets of the simulated samples are then divided into four categories based on the angular distance ΔR between the candidate jets and the generated W boson as well as the candidate jet substructure. The categorisation is illustrated in Fig. 11.9.

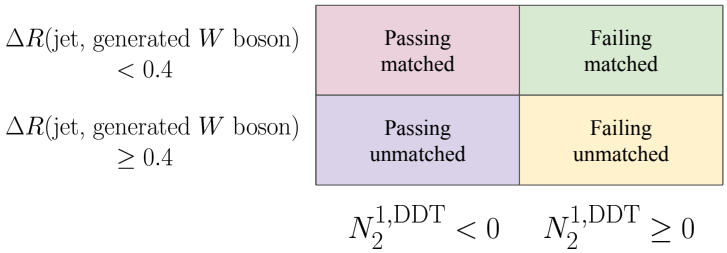


Figure 11.9: Illustration of hadronically decaying W candidate jet categorisation. "Jet" here refers to the candidate jet. If the angular distance ΔR between the candidate jet and the generated W boson is less than 0.4, the candidate jet is classified as *matched*, otherwise as *unmatched*. The *passing* region constitutes hadronically decaying W candidate jets that have $N_2^{1,\mathrm{DDT}} < 0$, while the *failing* region contains all other candidates.

The m_{SD} is chosen as the summary statistic, and up- and down-varied templates are created for the simulated samples. The JMS up- and down-varied templates are derived by shifting the mass scale up or down by 5%. In contrast, the JMR up template is obtained by smearing the mass value by 10% with a Gaussian function, while the JMR down variation is the nominal distribution. The systematic uncertainty of JMS and JMR is accounted for in the form of shape uncertainties.

The categorisation into two regions depending on the jet substructure facilitates the substructure selection efficiency (ϵ) to be estimated and its uncertainty accounted for. This is achieved by including a parameter for the efficiency into the maximum likelihood fit according to

$$\epsilon_{\text{passing}} = N_{\text{passing}}/N_{\text{failing}}, \tag{11.12}$$

$$\epsilon_{\text{failing}} = 1 - \epsilon_{\text{passing}}, \tag{11.13}$$

where N_{passing} and N_{failing} are the number of events with $N_2^{1,\text{DDT}} < 0$ and $N_2^{1,\text{DDT}} > 0$, respectively.

Both $\epsilon_{\text{passing}}$ and $\epsilon_{\text{failing}}$ are allowed to vary between $[0, 3]$ times their nominal values in the simulated samples, and the systematic uncertainty is accounted for in the form of normalisation uncertainty (Section 9.6) The substructure selection is then derived simultaneously as the JMS and JMR estimation.

The fitted values of the parameters for the data-taking period of early 2016 are provided in Table 11.1, and are representative of the SFs of the other data-taking periods.

$\epsilon_{\text{passing}}^{\text{data}}/\epsilon_{\text{passing}}^{\text{simulation}}$	δ_m (GeV)	$\sigma_m^{\text{data}}/\sigma_m^{\text{simulation}}$
0.85 ± 0.14	-1.50 ± 0.45	0.98 ± 0.04

Table 11.1: Corrections for the jet substructure selection ($\epsilon_{\text{passing}}$), JMS ($\delta_m$) and JMR ($\sigma_m$) for the data-taking period of early 2016.

11.5.3 Experimental uncertainties

Experimental uncertainties, including those related to the determination of the integrated luminosity [84–86], variations in the amount of pile-up and the isolation and identification of leptons are also considered. In addition, the effect of the limited statistics of the simulated samples and background estimation are also included.

Additional systematic uncertainties are applied to the event yields to account for the uncertainties due to the jet energy scale and resolution. The

efficiency and corresponding uncertainty for the AK8 DDB tagger selection are measured centrally by the CMS, and have a value of 1 and 30%, respectively. All experimental uncertainties are considered to be fully correlated across all data-taking periods, with the exception of a correlated component of the luminosity uncertainty.

11.5.4 Theoretical uncertainties

In addition to experimental uncertainties, theoretical uncertainties are included in the final fit to account for imprecision in the modeling of SM processes. The dominant theory uncertainty applied to all Higgs signal processes is due to the choice of QCD renormalization and factorization scales. The uncertainties is calculated by varying the renormalization and factorization scales up and down by a factor of two, according to the prescription in Ref. [100], and amounts to approximately 20% on ggF and 5% on VBF.

Additional theory systematics are applied to account for imprecise knowledge of the strong coupling constant and uncertainties on the parton distribution functions according to Ref. [101]. All theoretical uncertainties are considered to be fully correlated across all data-taking periods.

11.6 Statistical analysis

A binned maximum likelihood fit to the observed m_{SD} distribution is performed over the simulated signal and background contributions. The binning differs for the two production modes, and is chosen to maximise the expected significance in each category. The binnings are chosen as listed in the following text.

- **ggF category**, six differential bins in jet p_{T}.
- **VBF category**, two differential bins in invariant mass of the two leading small-radius jets (m_{jj}).

Three separate signal strengths are conducted to scale the event yields of ggF, VBF and $Z \to b\bar{b}$: μ_{ggF}, μ_{VBF} and μ_{Z}. As a result, the full likelihood function (Section 9.6) takes the form

$$
\begin{aligned}
&\mathcal{L}(\{N_{\text{obs}}\} \mid \mu, \theta) = \\
&\quad \prod_{i,j} \text{Poisson}(N^{\text{obs}}_{\text{ggF, C},i,j} | N^{\text{bkg}}_{\text{ggF,C},i,j} + \mu_Z N^{\text{Z}}_{\text{ggF,C},i,j} + \mu_{\text{ggF}} N^{\text{ggF}}_{\text{ggF,C},i,j} + \mu_{\text{VBF}} N^{\text{VBF}}_{\text{ggF,C},i,j}) \\
&\times \prod_{i,j} \text{Poisson}(N^{\text{obs}}_{\text{ggF,S},i,j} | N^{\text{bkg}}_{\text{ggF,S},i,j} + \mu_Z N^{\text{Z}}_{\text{ggF,S},i,j} + \mu_{\text{ggF}} N^{\text{ggF}}_{\text{ggF,S},i,j} + \mu_{\text{VBF}} N^{\text{VBF}}_{\text{ggF,S}}, i, j) \\
&\times \prod_{i,j'} \text{Poisson}(N^{\text{obs}}_{\text{VBF,C},i,j'} | N^{\text{bkg}}_{\text{VBF,C},i,j'} + \mu_Z N^{\text{Z}}_{\text{VBF,C},i,j'} + \mu_{\text{ggF}} N^{\text{ggF}}_{\text{VBF,C},i,j'} + \mu_{\text{VBF}} N^{\text{VBF}}_{\text{VBF,C},i,j'}) \\
&\times \prod_{i,j'} \text{Poisson}(N^{\text{obs}}_{\text{VBF,S},i,j'} | N^{\text{bkg}}_{\text{VBF,S},i,j'} + \mu_Z N^{\text{Z}}_{\text{VBF,S},i,j'} + \mu_{\text{ggF}} N^{\text{ggF}}_{\text{VBF,S},i,j'} + \mu_{\text{VBF}} N^{\text{VBF}}_{\text{VBF,S},i,j'}) \\
&\times f(\theta | \tilde{\theta}),
\end{aligned}
\tag{11.14}
$$

where N^{obs} denotes the observed number of events in a given bin. The first subscript indicates the category and the second subscript, S and C, represents the signal and control regions, respectively. The subscript i runs over the bins of the summary statistic m_{SD}. The subscript j runs over the six p_{T} bins in the ggF category, and the subscript j' runs over the two m_{jj} bins in the VBF category. The final term is the constraints on all nuisance parameters in the likelihood.

The test statistic chosen to determine the signal yield is based on the *profile likelihood ratio* described in Section 9.4. The nuisance parameters are added to Eq. 9.4 by adjusting the function to depend on both μ and the full set of nuisance parameters $\hat{\theta}$. The adjusted function is described by

$$
\tilde{q}_\mu = -2\ln \frac{\mathcal{L}(\{N_{\text{obs}}\} \mid \mu, \hat{\theta}_\mu)}{\mathcal{L}(\{N_{\text{obs}}\} \mid \hat{\mu}, \hat{\theta})} \tag{11.15}
$$

where $\hat{\theta}$ and $\hat{\mu}$ maximises the likelihood, and $\hat{\theta}_\mu$ refers to the conditional maximum likelihood estimators of θ.

The signal strength modifier is evaluated from a scan of $\tilde{q}_\mu$, performed with a parametric bootstrap as in Ref. [102]. The 68% CL intervals for μ are evaluated from $\tilde{q}_\mu = 1.00$.

11.7 Results

The combined signal strength for the VBF process is determined as $5.0^{+2.1}_{-1.8}$. The corresponding significance, calculated with the ggF signal strength freely floating, is observed to be 3.0 σ (with expected at 0.9 σ). Similarly, the combined signal strength for the ggF process is measured as $2.1^{+1.9}_{-1.7}$, yielding an observed and expected significance of 1.2 σ and 0.9 σ, respectively. The largest sources of uncertainty stem from the QCD background estimation, the theory uncertainties on Higgs boson production, the size of simulated signal samples, jet energy scale and the uncertainty on the DDB tagger selection. However, overall, the precision of the results is constrained by the statistical uncertainty inherent in the measurements. To enhance the reliability of the results, additional data collection is necessary to mitigate the impact of statistical fluctuations.

The observed data and fitted distributions of the m_{SD} in the VBF and ggF category are presented in Figs. 11.10 and 11.11, respectively. These results are aggregated over all differential bins and data-taking periods. The total background is decomposed into contributions from various processes, and the total uncertainty is represented by a red band. The background-enriched control region is depicted on the left, and the signal region is shown on the right. The fitted ggF and VBF distributions are overlaid in red and green, respectively. The apparent discontinuity at high mass in the ggF category is attributed to the selection of jet ρ as described in Section 11.3.3.

11.8 Discussion

The analysis presented in this chapter focuses on the exploration of boosted Higgs boson production through the $H \to b\bar{b}$ decay mode in conjunction with the VBF and ggF production modes. The rarity of such processes, coupled with the VBF process' unique sensitivity to Higgs couplings with vector bosons, makes this study an important addition to the CMS physics programme. The results, as outlined in the previous section, reveal a combined signal strength for the VBF process of $5.0^{+2.1}_{-1.8}$ with a corresponding observed significance of 3.0 σ. For the ggF process, the combined signal strength is measured as $2.1^{+1.9}_{-1.7}$, yielding an observed significance of 1.2 σ. These findings provide valuable insights into the behavior of boosted

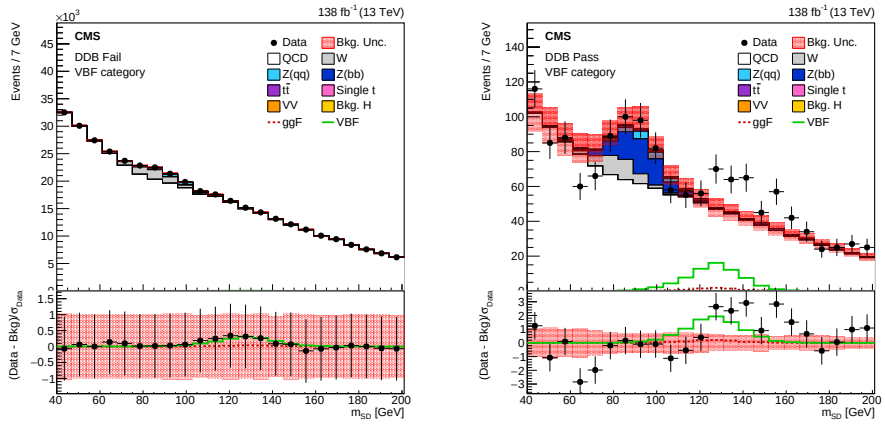


Figure 11.10: Data and fitted m_{SD} distribution in the VBF category, summed over all m_{jj} bins and data-taking periods. The control (left) and signal (right) regions are shown. Figures taken from Ref. [103].

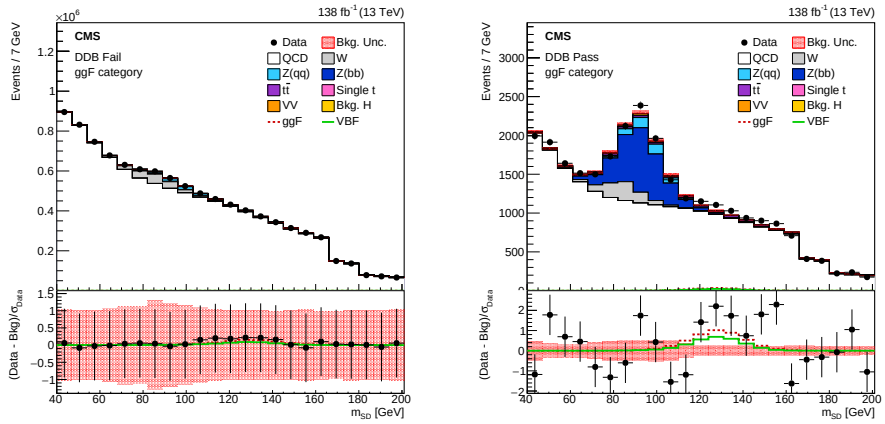


Figure 11.11: Data and fitted m_{SD} distribution in the ggF category, summed over all jet p_{T} bins and data-taking periods. The control (left) and signal (right) regions are shown. Figures taken from Ref. [103].

Higgs bosons and contribute to the broader understanding of Higgs boson interactions.

While the VBF production mode is a known process, the analysis presented here explores it for the first time at high transverse momentum ($p_{\mathrm{T}} > 450\,\mathrm{GeV}$). The measured signal strength of 5, indicates that the observed events are five times more abundant than what would be expected based solely on known particle interactions. This, coupled with the observed significance, provides evidence for the presence of boosted Higgs bosons produced through VBF. However, it is necessary to acknowledge the statistical uncertainty inherent in the measurements, which constrains the precision of the results. The range of $5.0^{+2.1}_{-1.8}$ implies that the true signal strength lies within this interval with a 68% CL. The observed significance of 3.0 σ is indicative of a noteworthy result, but additional data collection is essential to enhance the reliability of these findings and to mitigate the impact of statistical fluctuations. The significance of 0.9 σ for the ggF process also underscores the need for further data to strengthen the evidence supporting the observation.

The higher than expected signal strength for the VBF process raises possibilities regarding potential anomalous Higgs couplings, particularly those involving interactions with vector bosons. While the SM provides a well-established framework for understanding particle physics, deviations from the expected behaviour could signal the presence of new physics. Anomalous Higgs couplings, such as modifications in the coupling strength between the Higgs boson and vector bosons, could be a manifestation of new physics phenomena. In addition, the anomaly may be linked to momentum-dependent couplings, a phenomenon explored in the context of boosted Higgs boson production. Detailed investigations into the behaviour of the boosted Higgs boson, coupled with a larger dataset and improved experimental techniques, will be crucial for confirming or refuting the existence of anomalous couplings.

In conclusion, the exploration of boosted Higgs boson production through the $H \to b\bar{b}$ decay mode in VBF and ggF processes represents a significant step in advancing our understanding of Higgs boson interactions. The observed signal strengths in the VBF and ggF categories, while promising, necessitate further scrutiny and additional data to strengthen their statistical significances. The results presented here contribute valuable information to the broader field of particle physics, laying the foundation for future studies of the Higgs boson and its role in the fundamental forces of

the universe.

Chapter 12

Potential application of data scouting in highly energetic boson decay into hadronic final states

Chapter 11 showcased promising results for the search for boosted Higgs boson decays. However, as discussed, further investigation is still necessary. This chapter therefore explores the potential application of the scouting technique as a tool to enhance the efficiency of searches for highly energetic boson decay into hadronic final states. While Section 6.2 demonstrates the value of employing scouting jets for resonance searches in hadronic final states, the prospect of employing scouting jets to enhance searches for *boosted* hadronic resonances remains unexplored. This chapter investigates this prospect, focusing on boosted massive particles decaying into bottom quark-antiquark pairs.

Section 12.1 presents the physics motivation for using large-radius scouting jets in searches for boosted bosons decaying to jets. The potential of this technique is then explored by examining the jet tagging performance (Section 12.2) and jet mass regression (Section 12.3). Next, two prototypical searches are performed, starting with the search for boosted $Z \to b\bar{b}$ (Section 12.4). The purpose of this analysis is to utilise a well-understood physics process as a reference to consider the viability of the scouting technique. The methodology established for this search is then adopted to

conduct a more exploratory study — the search for boosted $H \to b\bar{b}$ in the ggF production mode (Section 12.5). While this preliminary analysis focuses on the ggF production mode, the scouting technique has the potential to study other production modes as well. Finally, the findings of the two searches are discussed in Section 12.6.

12.1 Physics motivation

As discussed in Section 6.1, because QCD multijet production is the predominant outcome of proton-proton collisions, the standard trigger strategy is required to adhere to strict energy and momentum thresholds to suppress such events and maintain a sustainable data acquisition rate. The scouting strategy offers a notable reduction in these thresholds (discussed in Section 6.3.3) enabling investigations at lower energy scales. In the analysis outlined in the preceding chapter, the decision to only study jets with $p_{\mathrm{T}} > 450\,\mathrm{GeV}$ (Section 11.3.1) is determined by the trigger thresholds of the standard strategy rather than a shortage of signal events at lower p_{T}. In fact, no CMS analysis has examined the hadronic decay of Higgs bosons below a p_{T} of 450 GeV. This is, for example, illustrated in the measurement and interpretation of differential cross sections for Higgs boson production (reported in Ref. [104]) obtained by combining the $H \to \gamma\gamma$, $H \to ZZ$, and $H \to b\bar{b}$ processes. As presented in Fig. 12.1, the hadronic process is contributing to only the two final bins of the measurement of the total differential Higgs boson cross section as a function of $p_{\mathrm{T}}^{\mathrm{H}}$.

Notably, the signal jets below 450 GeV can be retained by utilising the scouting technique, as shown in the following text. In simulation, boosted Higgs boson events are identified based on the requirement that the particle-level Higgs boson, together with its decay products (the bottom quark and antiquark), have a maximum angular distance $\Delta R < 0.8$ from the leading AK8 jet. The events are then required to satisfy a logical 'OR' expression of (a) jet-based scouting triggers or (b) triggers part of the standard strategy designed for the selection of boosted Higgs boson events through the usage of a DNN. The efficiency of these trigger expressions to select boosted $H \to b\bar{b}$ events is evaluated as a function of leading jet p_{T}, and displayed in terms of number of boosted Higgs boson events and trigger efficiency in Fig. 12.2. An approximate 20% improvement when employing the scouting triggers relative to the standard triggers is demonstrated. The improvement is most noticeable at low p_{T}, as presented to the left in Fig. 12.2. Due

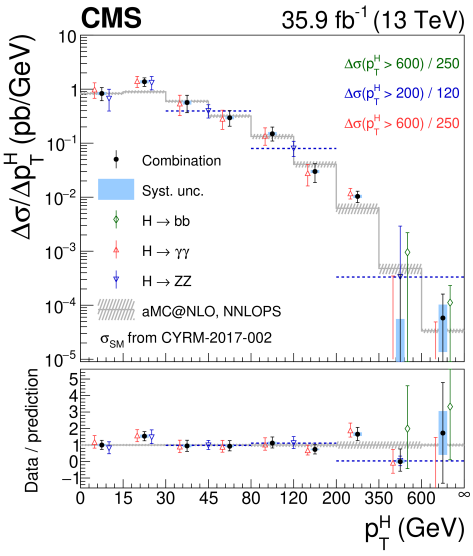


Figure 12.1: Measurement of the total differential cross section as a function of $p_{\mathrm{T}}^{\mathrm{H}}$. The spectra for the $H \to \gamma\gamma$, $H \to ZZ$, and $H \to b\bar{b}$ processes are shown in red, blue, and green, respectively. The combined spectrum is shown as black points. Figure taken from Ref. [104].

to the enhanced selection efficiency, the scouting technique enables the examination of a greater number of events. This is particularly advantageous for the analysis detailed in Chapter 11, where the outcome is constrained by the statistical uncertainty associated with the collision data. In parallel, the trigger efficiency curve to the right in Fig. 12.2 shows that the scouting triggers are 100% efficient from around 300 GeV. In contrast, the standard triggers are fully efficient only at about 500 GeV. As a result, a phase space inaccessible by the standard strategy is made available for examination by the scouting technique.

However, this greater selection obtained with the scouting technique may be significantly affected by the efficiency of identifying the scouting jets as originating from boosted $H \to b\bar{b}$. Identifying the origin of large-radius jets is crucial when exploring boosted topologies. For example, a DNN is leveraged for such a task by the triggers part of the standard strategy in the study described in the preceding paragraph. In order to make a fair statement on the efficacy of scouting-based searches for boosted hadronic resonances, it is necessary to investigate the scouting jet tagging efficiency.

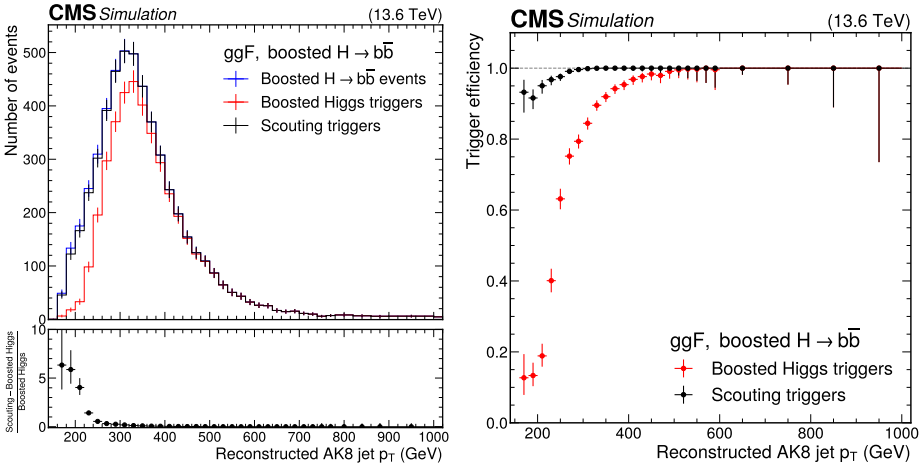


Figure 12.2: Number of events (left) and trigger efficiency (right) for ggF boosted $H \to b\bar{b}$ as a function of AK8 jet p_{T}. The black and red curves correspond to the scouting and the standard trigger selection, respectively. The efficiency is computed from simulation with a projected integrated luminosity of $\mathcal{L} = 100$ fb^{-1}. Figure taken from Ref. [54].

12.2 Jet tagging

A preliminary analysis of the Run-3 scouting jet tagging performance is explored in the following section. The ParticleNet algorithm, a DNN-based algorithm that identifies boosted hadronic decays for a wide range of resonance masses, is used for this task. At its core, the algorithm is constructed from the ParticleNet neural network architecture [105], a type of graph convolution network. For networks of this kind, the inputs are PF candidates that are processed in a permutation-invariant manner. A convolution operation is performed on each particle and its nearest neighbours on the $(\eta,\ \phi)$-plane. The last unit of the network is the soft-max function, which normalises the output and obtains a joint probability distribution for the output classes. The soft-max unit enables *multi-classification*, which in this instance facilitates the classification of a jet as either originating from $H \to b\bar{b}$ or QCD multijet production.

In order to assess the scouting jet tagging performance, the ParticleNet algorithm is trained on a set of large-radius scouting jets originating from $X \to b\bar{b}$ (signal) and QCD multijet (background) production, where X is a variable-mass spin-0 particle. By using simulated decays of a variable-

mass particle instead of the Higgs boson, the network is trained over a wide range of masses; effectively making its prediction mass-invariant. In parallel, jets from both the signal and background samples are re-weighted to yield flat distributions in m_{SD} and p_{T}; ensuring mass and momentum invariance for both signal and background events. This tagger trained on large-radius scouting jets is referred to as $\mathrm{DDB_S}$ in the following text.

The output of the algorithm provides two probability-like scores: $p(X \to b\bar{b})$ and $p(\mathrm{QCD})$. The discriminant used to separate $X \to b\bar{b}$ from QCD jets is the binary classification score defined as

$$\mathrm{D} = \frac{p(X \to b\bar{b})}{p(X \to b\bar{b}) + p(\mathrm{QCD})}. \tag{12.1}$$

While the network is trained with $X \to b\bar{b}$ events, a simulated sample of $H \to b\bar{b}$ is used as signal when evaluating its performance. This can be seen to the left in Fig. 12.3, where the distribution of Eq. 12.1 is displayed for simulated $H \to b\bar{b}$ and QCD multijet events. The discriminant distribution represents the output scores for each event, reflecting the tagger's confidence in identifying a given event as either signal or background. Well-separated distributions, as displayed here, indicate effective discrimination.

An alternative method for evaluating the performance of a jet tagger involves the analysis of the Receiver Operating Characteristic (ROC) curve, as depicted to the right in Fig. 12.3. In the context of discriminating between $X \to b\bar{b}$ and QCD multijet events, the ROC curve visually represents the trade-off between signal jets correctly classified as signal (*signal efficiency* or true positive rate) and background jets wrongly classified as signal (*background efficiency* or false positive rate). The area under the ROC curve (AUC) quantifies the overall performance, with a higher AUC indicating better classification. AUC ranges from 0% to 100%, where a classifier whose predictions are 100% wrong has an AUC of 0%, while one whose predictions are 100% correct has an AUC of 100% [106]. The AUC of the $\mathrm{DDB_S}$ tagger presented in Fig. 12.3 is 97.5%. For comparison, as presented in Figure 11.3, the prototypical version (V0) and the final version of the DDB tagger trained with offline reconstructed jets has an AUC of 97.2% and 98.6%, respectively.

The AUC of these three taggers all correspond to a high classification performance. While the AUC of the $\mathrm{DDB_S}$ tagger is slightly lower than that

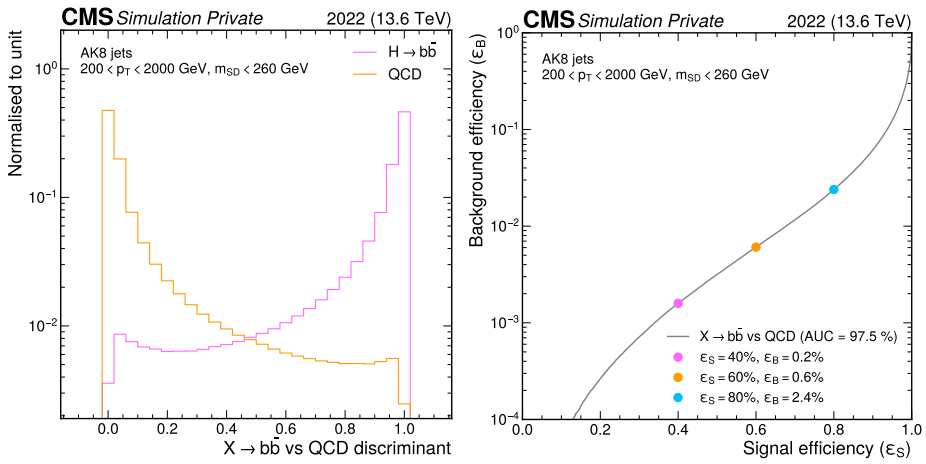


Figure 12.3: The normalised distributions of the $X \to b\bar{b}$ versus QCD discriminant (Eq. 12.1) for $H \to b\bar{b}$ and QCD events displayed in pink and orange, respectively (left). The ROC curve of the $X \to b\bar{b}$ versus QCD jet tagger displayed together with three working points; tight, medium and loose in pink, orange and blue, respectively (right).

of the final DDB version, it is important to note that the DDB_S tagger presented here is a prototype. It is therefore anticipated that, with the completion of further work, the performance of subsequent versions of the tagger will improve. As the DDB_S AUC exceeds (marginally) that of the first version of the DDB, it is considered that the scouting jet tagger has the potential to match or exceed the AUC of jet taggers trained with offline reconstructed jets. Moreover, while the AUC is helpful in providing an overall representation of the classifier's performance, the efficiency of specific working points (WPs) is of more interest from the perspective of a particle search.

WPs represent different trade-offs between maximising signal identification and minimising false positives, allowing the classifier's performance to be tailored to the analysis' objectives. Some analyses require a pure sample, with optimised signal efficiency for a fixed background rejection. In contrast, other analyses necessitate well-behaved background estimates with a certain amount of background events in the signal region (which requires a greater number of false positives).

Three WPs are determined from the discriminant distributions of Fig. 12.3. The selection of WPs involves choosing specific discrimination thresholds

aligning with desired signal or background efficiencies. In this section, the WPs corresponding to the signal efficiencies of 40%, 60%, and 80% (referred to as *tight*, *medium* and *loose* in the following text) are reported. The signal and background efficiencies for each WP are evaluated around the Higgs boson mass ([120, 130] GeV) with simulated events of $H \to b\bar{b}$ and QCD multijet production. The results are listed in Table 12.1. Notably, the tight WP has a signal and background efficiency in near perfect agreement with the DDB tagger (as presented in Section 11.4.1). Therefore, supported by the similar AUC of the DDB_S and DDB taggers, it is considered that the scouting jet tagging performance is comparable to that of a tagger trained with offline reconstructed events.

WP	Signal efficiency (%)	Background efficiency (%)
Loose	80	2.4
Medium	60	0.8
Tight	40	0.6

Table 12.1: Signal and background efficiencies for each considered WP of the DDB_S tagger.

12.3 Jet mass regression

In addition to performing the function of jet tagging, the DNN can also be applied to the task of mass regression. This improves the mass resolution, which increases the sensitivity of a particle search. In the context of boosted $H \to b\bar{b}$, an accurate reconstruction of the Higgs boson decay largely relies on the jet mass resolution. The ParticleNet network architecture is again employed, with the exception of the last soft-max unit. The removal of the soft-max unit allows the output to be a single real number, referred to as the regressed mass (m_reg). The inputs used to train the mass regression are the same as outlined in Section 12.2. The aim of the network is to generate an output as close as possible to the target mass (m_target), defined as

$$m_\text{target} = \begin{cases} m_\text{SD} \text{ of the generated large-radius jet,} \\ \qquad\qquad \text{if QCD sample;} \\ \text{generated } X\text{-particle mass,} \\ \qquad\qquad \text{if spin-0 particle sample.} \end{cases} \tag{12.2}$$

To evaluate the performance of the mass regression, distributions of m_{reg} and m_{SD} are presented and compared in Fig. 12.4 for three distinct p_{T} ranges of simulated $H \to b\bar{b}$ events. The mass regression significantly enhances the performance in comparison to the "soft drop" algorithm, particularly in two aspects. First, the "soft drop" algorithm causes failures in reconstruction, resulting in jet mass values close to 0 (as noted by the tail in Fig. 12.4). In contrast, the mass regression reconstructs jet mass values closer to the generated mass of the Higgs boson, effectively recovering the jets lost in the tails of the jet mass distribution. Second, the "soft drop" algorithm exhibits a pronounced p_{T}-dependence; the tail due to misreconstructions is greater at higher p_{T}. In comparison, no p_{T}-dependence is noticeable in the m_{reg} distribution. The m_{SD} distribution shown here is obtained prior to the application of residual p_{T}-dependent corrections, as mentioned in Section 11.2.2. While these corrections mitigate the aforementioned inefficiencies, no such corrections are required when utilising m_{reg}.

To facilitate further comparison, the jet mass response and jet mass resolution for simulated $H \to b\bar{b}$ events is studied for the m_{reg} and m_{SD} as a function of p_{T}. The results are presented in Fig. 12.5. The mass response is defined as the median of the $m_{\text{reco}}/m_{\text{target}}$ distribution, where m_{reco} is either m_{reg} or m_{SD}. Meanwhile, the mass resolution is estimated by finding half of the minimum interval containing 68% of the events (as described in Section 7.6.1). The response and resolution is computed twice, once over the full mass distribution, and once with mass values within the Higgs boson mass ($[100, 150]$ GeV) as input. The latter allows a comparison omitting the tail of the m_{SD} distribution.

While both the mass response and resolution are stable for the mass regression, the "soft drop" algorithm displays a greater p_{T}-dependence. The dependence is both observed when examining the response of the full and partial m_{SD} distribution, as well as the resolution of the full m_{SD} distribution. Additionally, the results display an improved resolution for m_{reg} with respect to m_{SD} when accounting for the full mass distribution, roughly 0.12 compared to values ranging from 0.14–0.5. However, analogous to the m_{reg} distribution, the m_{SD} resolution improves when examining the partial mass range. Notably, the m_{reg} resolution is similar to the jet mass resolution of the offline reconstruction as presented in Section 11.2.2. While the scouting jet mass resolution has not yet been studied using collision data, the findings presented here suggest that, similar to the scouting jet tagging efficiency, the mass resolution is comparable to that achieved with the of-

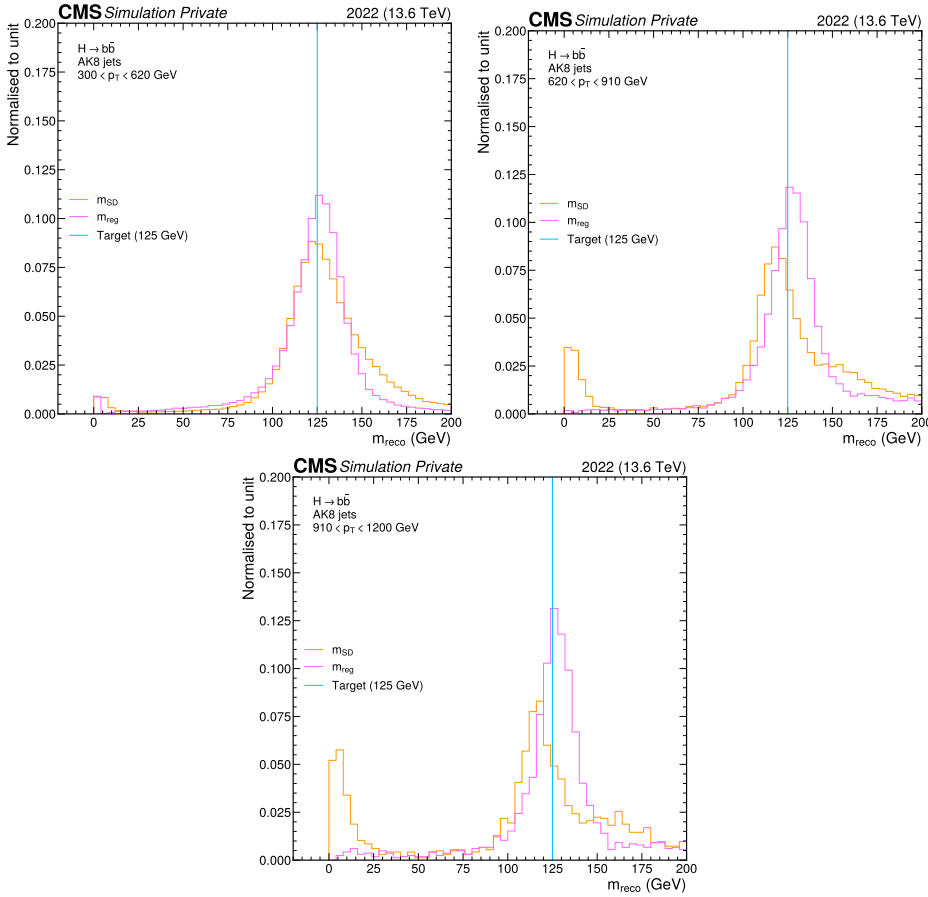


Figure 12.4: Distribution of m_{SD} (orange) and m_{reg} (pink) for simulated $H \to b\bar{b}$ events in three p_{T} ranges: $300 < p_{\mathrm{T}} < 620\,\mathrm{GeV}$ (upper left), $620 < p_{\mathrm{T}} < 910\,\mathrm{GeV}$ (upper right) and $910 < p_{\mathrm{T}} < 1200\,\mathrm{GeV}$ (lower). The target mass of 125 GeV is displayed in blue.

fline reconstruction.

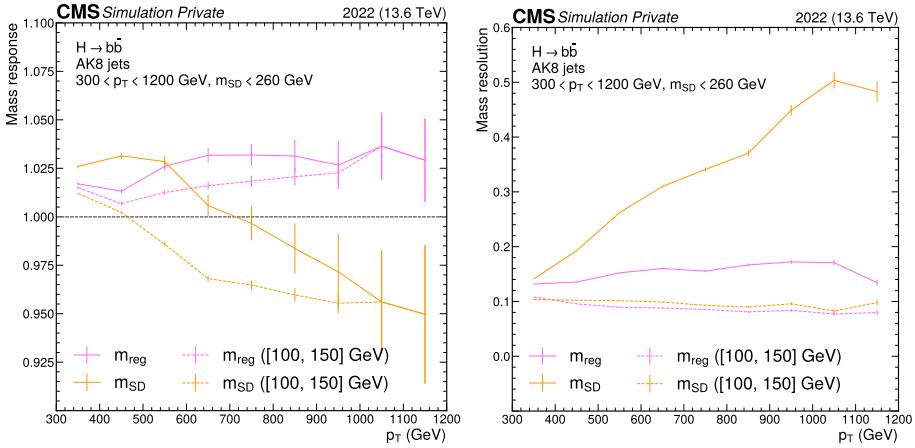


Figure 12.5: The mass response (left) and resolution (right) for simulated $H \to b\bar{b}$ events with $300 < p_{\mathrm{T}} < 1200\,\mathrm{GeV}$. The m_{SD} and m_{reg} are displayed in orange and pink, respectively. The response and resolution computed with the full and partial ($[100, 150]\,\mathrm{GeV}$) mass distribution are displayed with a bold and dashed line, respectively.

12.4 Searching for boosted $Z \to b\bar{b}$

Following the confirmed viability of scouting-based jet tagging and mass regression, the two methods are applied to conduct a prototypical search for boosted Z bosons decaying to bottom quark-antiquark pairs. The analysis presented in this thesis is the first documented search of this kind to be performed with scouting jets. The search broadly follows the methodology outlined in Chapter 11, with a few caveats. The protocol followed by the analysis is described below, along with a presentation of the results.

12.4.1 Event selection

The search uses proton-proton collision data at $\sqrt{s} = 13.6\,\mathrm{TeV}$, recorded by the scouting stream in 2022 and 2023. The Z candidate jet of each event is identified as the large-radius jet most likely to contain two b quarks (by selecting the jet with highest $\mathrm{DDB_S}$ score). The same jet energy calibration as

outlined in Section 7.2 is applied, except the corrections are specifically derived for large-radius jets. As a result of the missing JES calibration, a shift in the jet mass distribution (away from the true mass value of the decaying resonance particle) is expected. Before clustering, pileup is mitigated using the CHS technique (Section 4.4.3).

Events containing large-radius jets are selected by a logical 'OR' of the `L1_SingleJet180` and `L1_HTT360` seeds. This allows the p_{T} threshold to be lowered from 450 GeV to 300 GeV (at which point the logical 'OR' expression is fully efficient as reported in Section 6.3.3 and 12.1). The remaining event selection largely follows that described in Section 11.3, with the exception of that discussed in the following text.

At present, there is a limited number of physicists working on developing the scouting technique. While progress is being made, the small group size means that certain physics objects and observables have not yet been developed. While it is anticipated that these will be created in the future, their current non-existence affects this event selection. Due to the present absence of a neutral network trained to identify scouting small-radius jets originating from b quarks, events are not discarded if a b-tagged AK4 jet is in the opposite hemisphere of the Z candidate jet. As a result, a larger $t\bar{t}$ background contamination (in comparison with the study reported in the previous chapter) is expected. In parallel, as the reconstruction of tau leptons has not yet been developed for scouting, the requirement for the Z candidate jet to be separated from charged leptons by an angular distance of $\Delta R > 0.8$, only includes electrons and muons. As a result, a larger Z and $W^{\pm}$ boson background contribution is expected.

Finally, the $N_2^{1,\mathrm{DDT}}$ selection described in Section 11.3.4 is not applied. As reported in Appendix A, a distortion of the m_{reg} distribution is noticeable after applying the substructure selection. Further work is required to understand the source of this distortion, and consequently, the $N_2^{1,\mathrm{DDT}}$ selection is currently disregarded as it may bias the background estimation. While the selection serves to remove QCD multijet events, its omission does not affect the signal sensitivity (which was studied during its inclusion in the analysis reported in Chapter 11). Its inclusion in the analysis of the preceding chapter instead stems from the methodology of deriving the JMS and JMR as described in Section 11.5.2.

While the derivation of the JMS and JMR for this study will be performed after the completion of this thesis, it is possible to achieve this derivation

using various methodologies that are independent of the $N_2^{1,\mathrm{DDT}}$ selection. For example, the derivation may be accomplished by studying the scale and width of the Z boson peak.

12.4.2 Background estimation

As described in Section 11.4.1, in order to facilitate background estimation, Z candidate jets are partitioned into a signal region and a control region by selecting above and below a tagger discriminant value. Studies of the expected significance as a function of the $\mathrm{DDB_S}$ tagger discriminant showed a significance well above 5 σ for any discriminant value. Consequently, instead of choosing the discriminant value by maximising the number-counting significance (as outlined in Section 9.5), a value that produces a prominent $Z \to b\bar{b}$ peak over the background expectation is selected. The presence of a peak facilitates computation of the JMS and JMR. As a result, Z candidate jets with a $\mathrm{DDB_S}$ tagger discriminant ≥ 0.9945 are allocated to the signal region, while Z candidate jets with a $\mathrm{DDB_S}$ tagger discriminant < 0.9945 are allocated to the control region. This $\mathrm{DDB_S}$ threshold corresponds to signal and background efficiencies of roughly 20% and 0.01%, respectively.

The QCD background estimation is performed in a manner analogous to that described in Section 11.4.2. The transfer factor $F_{\mathrm{S/C}}^{\mathrm{QCD}}$ for the first p_{T} bin obtained from the procedure is displayed in Fig. 12.6. As expected, the shapes of the signal and control regions follow the same pattern (an important characteristic of a successful background estimation of this type). Moreover, the transfer factor (referred to as "Fit" in the figure) follows the pattern as well, indicating a successful estimation. The top quark background estimation will take place after the completion of the thesis.

12.4.3 Statistical analysis and results

A statistical analysis utilising the procedure detailed in Section 11.6 is performed to estimate signal strength (μ_{Z}) and significance. Due to the prototypical nature of the analysis, only certain systematic uncertainties are considered. However, these include uncertainties among those with the greatest impact on the signal strength of the analysis presented in Chapter 11: uncertainties related to the JES and the QCD background estimation. In

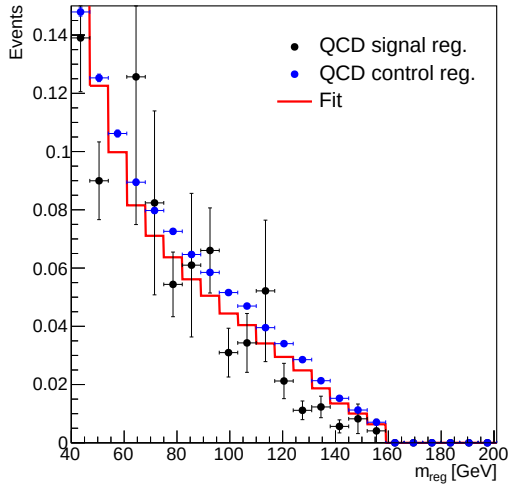


Figure 12.6: Normalised event yield of the simulated QCD multijet sample as a function of m_{reg} for Z candidate jets with $300 < p_{\text{T}} < 350\,\text{GeV}$. The signal and control regions are shown with black and blue points, respectively. The transfer factor $F_{\text{S/C}}^{\text{QCD,0}}$ is shown with a red line.

addition, uncertainties related to the JER, size of simulated samples, luminosity, pile-up and imprecision in the modelling of the SM processes are also included.

12.4.4 Signal strength and significance

During the statistical fit, the μ_{Z} is permitted to vary between $[0, 3]$. In contrast, the $H \to b\bar{b}$ signal strength (μ_{ggF}) is fixed to 1.0 and the Higgs boson mass window ($[120, 130]\,\text{GeV}$) is excluded from the fit. The value of μ_{Z} is evaluated from a scan of Eq. 11.15, performed with a parametric bootstrap as described in Ref. [102]. An illustration of the scan is presented in Fig. 12.7. The observed signal strength obtained from this procedure is determined as $1.1^{-0.1}_{+0.1}$, where the uncertainty is obtained from the 68% CL intervals.

The corresponding significance is observed to be 20 σ (with expected at 16 σ). The observed data and fitted distributions of the m_{reg} are presented in Fig. 12.8. These results are aggregated over all p_{T} bins and the data-taking periods of 2022 and 2023. The distributions per p_{T} bin are presented in

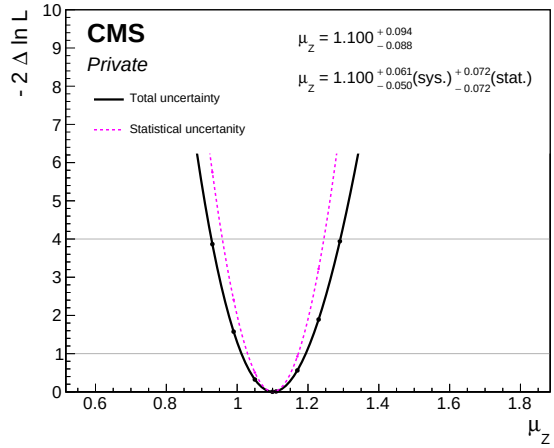


Figure 12.7: Likelihood scan of the observed signal strength for $Z \to b\bar{b}$, with μ_{ggF} fixed to 1.0 and the Higgs boson mass window ($[120, 130]$ GeV) excluded. The 68% CL interval is denoted with the lower horizontal line at $\tilde{q}_\mu = 1.00$. The likelihood scan when accounting for only the statistical uncertainty is marked with a pink dashed line.

Appendix B. The total background is decomposed into contributions from various processes, and the total uncertainty is represented by a red band. The background-enriched control region is depicted on the left, and the signal region on the right. Notably, a pronounced increase in the relative contribution from the $Z \to b\bar{b}$ decay is evident after applying the DDB_S selection. The $Z \to b\bar{b}$ event yield divided by the statistical uncertainty of the collision data is displayed with a dashed blue line in the lower panel.

12.4.5 Jet mass resolution

The prominent $Z \to b\bar{b}$ yield over the background expectation facilitates computation of the jet mass resolution using collision data. This is achieved by fitting a Gaussian function to the data points in the lower panel of the signal region in Fig. 12.8. Figure 12.9 displays this fit. The fitted parameters of the Gaussian function are measured as $\mu = 96.13 \pm 0.57$ GeV and $\sigma = 10.98 \pm 0.63$, where μ is approximately 5% larger relative to the Z boson mass. A shift in the jet mass distribution is therefore observed, however, as discussed in Section 12.4.1, is also expected. The shift may be corrected with dedicated calibration studies. Notably, the ratio $\sigma/\mu \approx 0.11$ is in agree-

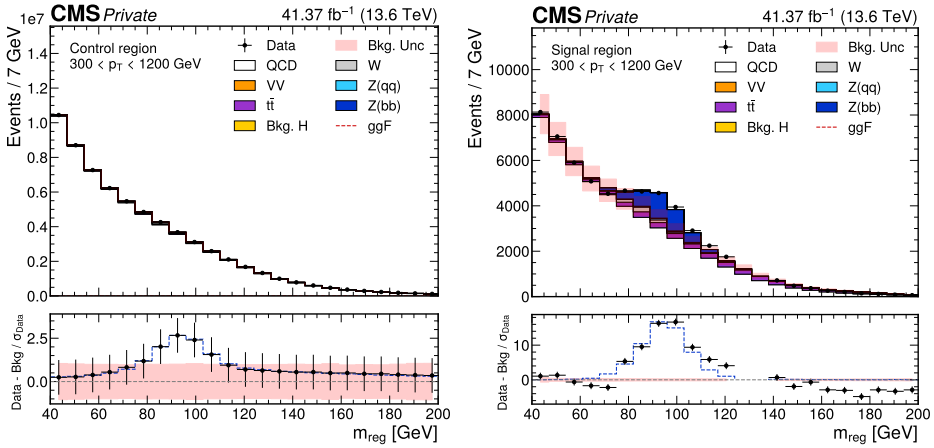


Figure 12.8: Data and fitted m_{reg} distribution, summed over all jet p_{T} bins and data-taking periods. The control (left) and signal (right) regions are shown. The Higgs boson mass window ($[120, 130]$ GeV) in the signal region is concealed. The total background uncertainty is represented by a red band. The lower panel shows the difference between the collision data and the background event yield, divided by the statistical uncertainty of the collision data (black data points). The $Z \to b\bar{b}$ event yield divided by the statistical uncertainty of the collision data is displayed with a dashed blue line.

ment with the m_{reg} resolution reported for simulated large-radius jets in Section 12.3.

12.5 Searching for boosted $H \to b\bar{b}$

Following the successful execution of the search for boosted $Z \to b\bar{b}$ using the scouting technique, the methodology outlined in the previous sections is adopted to conduct a search for boosted Higgs bosons produced through ggF decaying to bottom quark-antiquark pairs. While the analysis primarily targets the ggF production mode, the selection is not exclusive to ggF (which has a relative contribution to the Higgs boson production of more than 50% in the signal region). There is non-negligble contamination from the VBF production mode ($< 25\%$) and minor contributions from VH ($< 15\%$) and ttH ($< 10\%$).

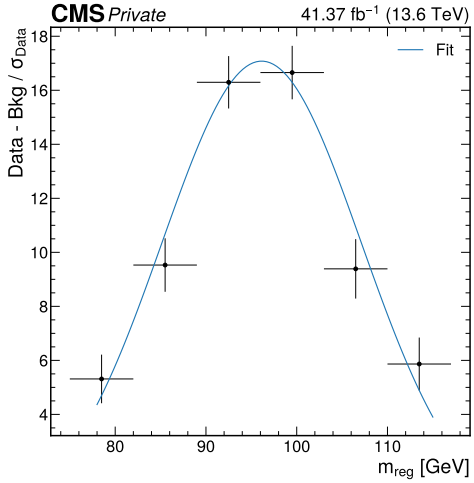


Figure 12.9: The lower panel of the signal region in Fig. 12.8, showing the difference between the collision data and the background event yield, divided by the statistical uncertainty of the collision data (black data points). A Gaussian function fitted to the data points is displayed with a blue curve.

The event selection remains the same, with one addition. As a consequence of shifting the p_{T} threshold, the constraints on jet ρ are changed from $\rho \in [-6, -2.1]$ to $\rho \in [-6, -1.7]$. This change is necessary to avoid discarding Higgs candidate jets around the Higgs boson mass. The effects of this shift is studied in detail and reported in Appendix C.

The signal and control regions are selected with a $\mathrm{DDB_S}$ tagger discriminant score that maximises the number-counting significance for this specific decay signature. The procedure is displayed in Fig. 12.10. The Higgs boson candidate jets with a $\mathrm{DDB_S}$ tagger discriminant ≥ 0.9858 are allocated to the signal region, while Higgs boson candidate jets with a $\mathrm{DDB_S}$ tagger discriminant < 0.9858 are allocated to the control region. This $\mathrm{DDB_S}$ threshold corresponds to a signal and background efficiency of 40% and 0.6%, respectively.

During the statistical analysis, μ_{Z} is allowed to float freely while μ_{ggF} is constrained to $[-9, 10]$. The Higgs boson mass window ($[120, 130]$ GeV) remains concealed. To avoid biasing future extensions of this prototypical study, only the expected signal strength and significance is estimated. The expected signal strength obtained from this procedure is determined as $0.9^{-1.0}_{+1.0}$, where the uncertainty is obtained from the 68% CL intervals. The

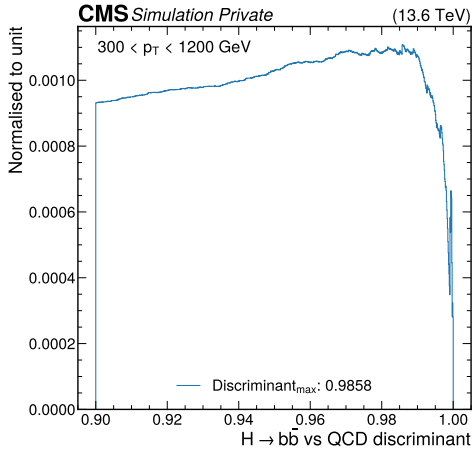


Figure 12.10: Normalised significance as a function of the DDB_S tagger discriminant value. The distribution reaches its maximum value at 0.9858.

corresponding significance is observed to be 1.0 σ.

12.6 Discussion

While there has been significant progress towards an understanding of the physics underpinning boosted decays of Higgs bosons, current knowledge is not yet comprehensive. The application of novel and innovative approaches, such as the scouting technique, is therefore required.

The studies presented in this chapter showcase the promise of scouting to further current understanding of highly energetic boson decay into hadronic final states. This is achieved by proving (a) the viability of scouting as an analysis strategy and (b) the potential of scouting to extend knowledge of the Higgs boson. Considering (a), the successful implementation of scouting-based jet tagging and jet mass regression highlights its comparability with the standard analysis strategy. Moreover, the measured jet mass resolution of 9–11% (as computed with simulation in Section 12.3 and collision data in Section 12.4.5) is in agreement with that measured with offline reconstructed jets. The viability of this new approach is further highlighted by the outcome of the search for boosted $Z \to b\bar{b}$ (a signal strength of $1.1^{-0.1}_{+0.1}$, which is in agreement with the SM expectation).

Considering (b), while the searches presented in Section 12.4 and 12.5 are preliminary, the results demonstrate the potential of scouting to contribute significantly to boosted hadronic searches. Here, measurements of the Z and Higgs boson are conducted in complicated final states that with traditional methods would be over-run by QCD multijet production. The expected significance of 1.0 σ reported for the ggF search marginally exceeds the expected significance for the same production mode as discussed in Chapter 11.6. This is promising, particularly as the integrated luminosity used in this analysis equates to only a third of that used in Chapter 11. However, it is important to acknowledge that the inclusion of further systematic uncertainties and completion of the top quark background estimation may affect the significance, emphasising the need for future work.

Beyond the immediate findings, the searches reported here establish a foundational platform for subsequent analyses. Both searches are the first documented analyses of this kind to be performed with scouting jets. In addition to performing a complete study of boosted $H \to b\bar{b}$, further work may include more extensive searches into diverse boosted decay modes and distinct resonance searches within the high-energy physics regime. In conclusion, the preliminary analyses presented here highlight the effectiveness of the scouting technique and its significant potential to contribute to the understanding of the Higgs boson.

Chapter 13

Summary and outlook

This thesis comprised two complementary parts, each addressing aspects of the application of scouting jets in high-energy physics research. Part I assessed the performance and precision of jets reconstructed using the scouting technique, while Part II focused on the search for boosted Higgs bosons decaying to bottom quark-antiquark pairs. Although the search is first performed with the offline reconstructed jets of the standard trigger strategy, the final chapter delves into the assessment of the viability of integrating scouting jets into the analysis.

In Part I, the effectiveness of using scouting jets to provide access to previously unexplored phase spaces, particularly at low-energy, was demonstrated. The efficiency with which small- and large-radius scouting jets were selected was found to be 100% from approximately 300 GeV. In comparison, the standard strategy was only found to be fully efficient from around 600 and 800 GeV, respectively. As a result of the greater selection efficiency of the scouting technique, scouting-based analyses are able to probe phase spaces inaccessible with the standard trigger strategy — increasing the power of the search for new physics. Moreover, a study of the jet energy showed a good level of agreement between the scouting and offline reconstructions. The study demonstrated a worsening of the jet energy resolution of approximately 10% and 2% below and above 500 GeV, respectively. Despite a slightly degraded performance compared to offline reconstructed jets, scouting jets prove to be reliable tools for various analyses, especially in scenarios where the highest precision is not a primary concern — but instead the statistical uncertainty dominates.

Part II of the thesis presented a comprehensive search for boosted Higgs bosons decaying into bottom quark-antiquark pairs. The analysis, for the first time, provides insights into this physics process in conjunction with the VBF production mode. The combined signal strength for the VBF process was found to be $5.0^{+2.1}_{-1.8}$, with an observed significance of 3.0 σ (expected at 0.9 σ). Similarly, the ggF process yielded a combined signal strength of $2.1^{+1.9}_{-1.7}$, with an observed and expected significance of 1.2 σ and 0.9 σ, respectively. The higher than expected signal strength for the VBF process raises possibilities regarding potential anomalous Higgs couplings, particularly those involving interactions with vector bosons. While promising, the observed significance necessitate further scrutiny and additional data to strengthen the evidence supporting the observation. Notably, this study served as a precursor for the subsequent investigation into the potential integration of scouting jets into the analysis.

Chapter 12 provided a unified synthesis, drawing on the key findings from both parts of the thesis, and lays the groundwork for future research directions. The application of scouting jets to search for boosted Z and Higgs bosons was explored, demonstrating their potential as effective alternatives to offline reconstructed jets. Machine learning techniques applied to scouting jets in this context exhibit comparable performance to their application on offline reconstructed jets. The signal strength for the boosted $Z \to b\bar{b}$ process was found to be $1.1^{+0.1}_{-0.1}$, with an observed significance of 20 σ (expected at 16 σ). As the physics process is already well-understood, the close agreement with the SM expectation strengthens the viability of the scouting technique.

In parallel, the expected signal strength and significance for the ggF process was found to be $0.9^{+0.1}_{-0.1}$ and 1.0 σ, respectively, when utilising scouting jets. While this is promising, particularly given the currently available integrated luminosity of Run–3, it is important to acknowledge that future work is required to strengthen the analysis. While the unexpected delayed start and early shutdown of the LHC in 2021 and 2022, respectively, posed challenges by limiting the recorded data available, the evidence presented within this thesis suggest that scouting jets are indeed valuable assets for boosted hadronic searches. However, the full realisation of this potential will depend on the accumulation of sufficient collision data and further improvement of the methodology for a comprehensive analysis.

Providing a foundational platform for subsequent analyses, the searches reported in this thesis offer more than their immediate findings. As both

searches are the first documented analyses of this kind to be performed with scouting jets, they lay the groundwork for future trigger-level analyses. Looking forward, future research directions should prioritise completing the ongoing study initiated in Part II on the scouting-based analysis of boosted Higgs bosons. This may increase the statistical significance of the results presented in Chapter 11, and will contribute to a deeper understanding of the capabilities and applications of scouting jets in the realm of high-energy physics — particularly in highly energetic boson decay into hadronic final states.

References

[1] S. L. Glashow, "Partial-symmetries of weak interactions", *Nuclear Physics* **22** (1961) doi:https://doi.org/10.1016/0029-5582(61)90469-2.

[2] S. Weinberg, "A model of leptons", *Phys. Rev. Lett.* **19** (1967) doi:10.1103/PhysRevLett.19.1264.

[3] N. Aghanim et al., "Planck 2018 results VI. Cosmological parameters", *Astronomy & Astrophysics* **641** (2020) doi:10.1051/0004-6361/201833910.

[4] L. Evans and P. Bryant, "LHC Machine", *Journal of Instrumentation* **3** (2008) doi:10.1088/1748-0221/3/08/S08001.

[5] CMS Collaboration, "Performance of the CMS Level-1 trigger in proton-proton collisions at 13 TeV", *Journal of Instrumentation* **15** (2020) doi:10.1088/1748-0221/15/10/P10017.

[6] CMS Collaboration, "The CMS trigger system", *Journal of Instrumentation* **12** (2017) doi:10.1088/1748-0221/12/01/p01020.

[7] CMS Collaboration, "Search for Narrow Resonances in Dijet Final States with the Novel CMS Technique of Data Scouting", *Phys. Rev. Lett.* **117** (2016) doi:10.1103/physrevlett.117.031802.

[8] B. Richter, "Very high energy electron-positron colliding beams for the study of weak interactions", *Nuclear Instruments and Methods* **136** (1976) doi:https://doi.org/10.1016/0029-554X(76)90396-7.

[9] CMS Collaboration, "The CMS experiment at the CERN LHC. The Compact Muon Solenoid experiment", *JINST* **3** (2008) doi:10.1088/1748-0221/3/08/S08004.

[10] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *JINST* **3** (2008) doi:10.1088/1748-0221/3/08/S08003.

[11] W. Herr and B. Muratori, "Concept of luminosity", 2006 https://cds.cern.ch/record/941318.

[12] S. Fartoukh, "Achromatic telescopic squeezing scheme and application to the LHC and its luminosity upgrade", *Phys. Rev. ST Accel. Beams* **16** (2013) doi:10.1103/PhysRevSTAB.16.111002.

[13] J. Wenninger et al., "beta* leveling with telescopic ATS squeeze (MD 2410)", 2017 https://cds.cern.ch/record/2285184.

[14] CMS Collaboration, "Presentation layer of CMS Online Monitoring System", *EPJ Web Conf.* **214** (2019) doi:10.1051/epjconf/201921401044.

[15] O. Aberle et al., "High-Luminosity Large Hadron Collider (HL-LHC): Technical design report". CERN Yellow Reports: Monographs. CERN, Geneva, 2020. doi:10.23731/CYRM-2020-0010.

[16] CMS Collaboration, "CMS Luminosity – Public Results", (2023). Summary: proton-proton collisions since 2015 (Run 2 + Run 3).

[17] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV", *Journal of Instrumentation* **10** (2015) doi:10.1088/1748-0221/10/06/p06005.

[18] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV", *Journal of Instrumentation* **13** (2018) doi:10.1088/1748-0221/13/06/p06015.

[19] CMS Collaboration, "Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Journal of Instrumentation* **10** (2015) doi:10.1088/1748-0221/10/08/p08010.

[20] D. Barney, "CMS Detector Slice", (2016). CMS Collection.

[21] S. H. Laurila, "Search for Charged Higgs Bosons Decaying to a Tau Lepton and a Neutrino with the CMS Experiment", 2019.

[22] CMS Collaboration, "The CMS tracker system project: Technical Design Report". https://cds.cern.ch/record/368412, 1997.

[23] CMS Collaboration, "The CMS tracker: addendum to the Technical Design Report". https://cds.cern.ch/record/490194, 2000.

[24] CMS Collaboration, "The CMS Phase-1 Pixel Detector Upgrade". https://cds.cern.ch/record/2745805, 2020.

[25] CMS Collaboration, "Dimuon scouting at CMS", Technical Report CMS-DP-2023-070, CERN, Geneva, 2023. https://cds.cern.ch/record/2869310.

[26] T. C. Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *Journal of Instrumentation* **9** (2014) doi:10.1088/1748-0221/9/10/p10009.

[27] CMS Collaboration, "The CMS electromagnetic calorimeter project: Technical Design Report". https://cds.cern.ch/record/349375, 1997.

[28] CMS Collaboration, "Changes to CMS ECAL electronics: addendum to the Technical Design Report". https://cds.cern.ch/record/581342, 2002.

[29] CMS Collaboration, "Technical proposal for the upgrade of the CMS detector through 2020". https://cds.cern.ch/record/1355706, 2011.

[30] D. Barney, "CMS Document 12030-v2, Materials for CMS ECAL POSTER", https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=12030.

[31] CMS Collaboration, "The CMS hadron calorimeter project: Technical Design Report". https://cds.cern.ch/record/357153, 1997.

[32] CMS Collaboration, "CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter". https://cds.cern.ch/record/1481837, 2012.

[33] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Physics Letters B* **716** (2012) doi:https://doi.org/10.1016/j.physletb.2012.08.021.

[34] CMS Collaboration, "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV", *Journal of High Energy Physics* **81** (2013) doi:https://doi.org/10.1007/JHEP06(2013)081.

[35] CMS Collaboration, "The CMS muon project: Technical Design Report". https://cds.cern.ch/record/343814, 1997.

[36] CMS Collaboration, "Performance of the CMS muon detector and muon re- construction with proton-proton collisions at 13 TeV", *Journal of Instrumentation* **13** (2018) doi:10.1088/1748-0221/13/06/P06015.

[37] P. D. Group, R. L. Workman, V. D. Burkert, and Crede, "Review of Particle Physics", *Progress of Theoretical and Experimental Physics* **2022** (2022) doi:10.1093/ptep/ptac097.

[38] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014) doi:10.1088/1748-0221/9/10/P10009.

[39] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *Journal of Instrumentation* **12** (2017) doi:10.1088/1748-0221/12/10/p10003.

[40] CMS Collaboration, "CMS Physics briefing", (2023). Machining jets.

[41] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k$_t$ jet clustering algorithm", *Journal of High Energy Physics* (2008) doi:10.1088/1126-6708/2008/04/063.

[42] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *The European Physical Journal C* (2012) doi:10.1140/epjc/s10052-012-1896-2.

[43] CMS Collaboration, "Identification techniques for highly boosted W bosons that decay into hadrons", *Journal of High Energy Physics* **2014** (2014) doi:10.1007/jhep12(2014)017.

[44] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup per particle identification", *JHEP* **10** (2014) 059, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013.

[45] CMS Collaboration, "Pileup mitigation at CMS in 13 TeV data", *JINST* **15** (2020) P09018, doi:10.1088/1748-0221/15/09/p09018, arXiv:2003.00503.

[46] CMS Collaboration, "Development of the CMS detector for the CERN LHC Run 3", technical report, CERN, 2023. https://cds.cern.ch/record/2870088.

[47] G. Bagliesi et al., "Debugging Data Transfers in CMS", technical report, CERN, 2010. doi:10.1088/1742-6596/219/6/062055.

[48] A. Bocci et al., "Heterogeneous reconstruction of tracks and primary vertices with the CMS pixel tracker", 2020. doi:10.48550/arXiv.2008.13461.

[49] CMS Collaboration, "Search for Narrow Resonances using the Dijet Mass Spectrum in pp Collisions at sqrt s of 7 TeV", technical report, CERN, Geneva, 2012. http://cds.cern.ch/record/1461223.

[50] CMS Collaboration, "Search for narrow resonances in dijet final states at $\sqrt{s} = 8$ TeV with the novel cms technique of data scouting", *Phys. Rev. Lett.* **117** (2016) 031802, doi:10.1103/PhysRevLett.117.031802.

[51] CMS Collaboration, "Search for narrow resonances decaying to dijets in pp collisions at $\sqrt{s} = 13$ TeV using $12.9\ \mathrm{fb}^{-1}$", Technical Report CMS-PAS-EXO-16-032, CERN, Geneva, 2016. http://cds.cern.ch/record/2205150.

[52] CMS Collaboration, "Search for a narrow resonance decaying to a pair of muons in proton-proton collisions at 13 TeV", Technical Report CMS-PAS-EXO-19-018, CERN, Geneva, 2019. https://cds.cern.ch/record/2684861.

[53] CMS Collaboration, "PF Jet Performances at High Level Trigger using Patatrack pixel tracks", Technical Report CMS-DP-2021-005, CERN, Geneva, 2021. https://cds.cern.ch/record/2765489.

[54] CMS Collaboration, "Trigger efficiency studies of the CMS Run-3 Data Scouting of Jet, Electron and Photon objects and comparison with standard HLT paths during during proton-proton collisions at $\sqrt{s}$ = 13.6 TeV", 2023 https://cds.cern.ch/record/2875709.

[55] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *Journal of Instrumentation* **12** (2017) doi:10.1088/1748-0221/12/02/p02014.

[56] CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in CMS", *Journal of Instrumentation* **6** (2011) doi:10.1088/1748-0221/6/11/p11002.

[57] CMS Collaboration, "Jet energy scale and resolution measurements using data scouting events collected by the CMS experiment in 2022 at $\sqrt{s} = 13.6$ TeV", 2023 https://cds.cern.ch/record/2871592.

[58] CMS Collaboration, "Mass regression of highly-boosted jets using graph neural networks", 2021 http://cds.cern.ch/record/2777006.

[59] I. Aitchison and A. Hey, "Gauge Theories in Particle Physics: A Practical Introduction, Fourth Edition - 2 Volume set". CRC Press, Bristol, UK, 2012.

[60] G. A. et al., "Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c^2 at the CERN SPS Collider", *Phys. Rev. Lett.* **126** (1983) doi:10.1016/0370-2693(83)90188-0.

[61] P. B. et al., "Evidence for $Z^0 \to e^+e^-$ at the CERN pp Collider", *Phys. Rev. Lett.* **129** (1983) doi:10.1016/0370-2693(83)90744-X.

[62] P. W. Higgs, "Broken symmetries, massless particles and gauge fields", *Phys. Rev. Lett.* **12** (1964) doi:10.1016/0031-9163(64)91136-9.

[63] P. W. Higgs, "Broken symmetries and the masses of gauge bosons", *Phys. Rev. Lett.* **13** (1964) doi:10.1103/PhysRevLett.13.508.

[64] C. R. H. G. S. Guralnik and T. W. B. Kibble, "Global conservation laws and massless particles", *Phys. Rev. Lett.* **13** (1964) doi:10.1103/PhysRevLett.13.585.

[65] P. W. Higgs, "Spontaneous symmetry breakdown without massless bosons", *Phys. Rev. Lett.* **145** (1966) doi:10.1103/PhysRev.145.1156.

[66] T. W. B. Kibble, "Symmetry breaking in non-Abelian gauge theories", *Phys. Rev. Lett.* **165** (1967) doi:10.1103/PhysRev.155.1554.

[67] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons", *Phys. Rev. Lett.* **13** (1964) doi:10.1103/PhysRevLett.13.321.

[68] ATLAS Collaboration, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc", *Physics Letters B* **716** (2012) doi:https://doi.org/10.1016/j.physletb.2012.08.020.

[69] CMS Collaboration, "A portrait of the Higgs boson by the CMS experiment ten years after the discovery", *Nature* (2022) doi:10.1038/s41586-022-04892-x.

[70] ATLAS Collaboration, "A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery", *Nature* **607** (2022) doi:10.1038/s41586-022-04893-w.

[71] CERN, "CERN Yellow Reports: Monographs, Vol 2 (2017): Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector", 2017. doi:10.23731/CYRM-2017-002.

[72] K. Becker et al., "Precise predictions for boosted Higgs production", *SciPost Phys. Core* (2024) doi:10.21468/SciPostPhysCore.7.1.001.

[73] CERN, "Handbook of LHC Higgs Cross Sections: 2. Differential Distributions", 2012. doi:10.5170/CERN-2012-002.

[74] G. Cowan, "Statistical data analysis". Oxford University Press, USA, 1998.

[75] ATLAS and CMS Collaborations, and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011", technical report, CERN, 2011. https://cds.cern.ch/record/1379837.

[76] A. Kusenko, L. Pearce, and L. Yang, "Postinflationary Higgs Relaxation and the Origin of Matter-Antimatter Asymmetry", *Phys. Rev. Lett.* **114** (2015) doi:10.1103/PhysRevLett.114.061302.

[77] M. Gouzevitch and A. Carvalho, "A review of Higgs boson pair production", *Reviews in Physics* **5** (2020) 100039, doi:https://doi.org/10.1016/j.revip.2020.100039.

[78] LHCHiggsCrossSectionWorkingGroup Collaboration, "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Section Working Group". CERN Yellow Reports: Monographs, 2013. doi:10.5170/CERN-2013-004.

[79] J. Steggemann, "Extended Scalar Sectors", *Ann. Rev. Nucl. Part. Sci.* **70** (2020) doi:10.1146/annurev-nucl-032620-043846.

[80] O. Witzel, "Review on Composite Higgs Models", 2019.

[81] F. Maltoni, K. Mawatari, and M. Zaro, "Higgs characterisation via vector-boson fusion and associated production: NLO and parton-shower effects", *Eur. Phys. J. C* **74** (2014) doi:10.1140/epjc/s10052-013-2710-5.

[82] CMS Collaboration, "Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **2020** (2020) doi:10.1007/jhep12(2020)085.

[83] G. Aad et al., "Constraints on Higgs boson production with large transverse momentum using $H \to b\bar{b}$ decays in the ATLAS detector", *Phys. Rev. D* **105** (2022) doi:10.1103/physrevd.105.092003, arXiv:2111.08340.

[84] CMS Collaboration, "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS", *Eur. Phys. J. C* **81** (2021) doi:10.1140/epjc/s10052-021-09538-2.

[85] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary, CERN, Geneva, 2018. https://cds.cern.ch/record/2621960.

[86] CMS Collaboration, "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary, CERN, Geneva, 2019. https://cds.cern.ch/record/2676164.

[87] C. Collaboration, "Displays of candidate events in the search for new heavy resonances decaying to dibosons in the all-jets final state in the CMS detector", (2022). CMS Collection.

[88] CMS Collaboration, "Performance of the mass-decorrelated DeepDoubleX classifier for double-b and double-c large-radius jets with the CMS detector", 2022 https://cds.cern.ch/record/2839736.

[89] CMS Collaboration, "Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector", 2018 http://cds.cern.ch/record/2630438.

[90] CMS Collaboration, "Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at 13 TeV", *Journal of High Energy Physics* **12** (2020) doi:10.1007/jhep12(2020)085.

[91] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft drop", *Journal of High Energy Physics* **2014** (2014) doi:10.1007/jhep05(2014)146.

[92] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, "Better jet clustering algorithms", *JHEP* **08** (1997) doi:10.1088/1126-6708/1997/08/001, arXiv:hep-ph/9707323.

[93] M. Wobisch and T. Wengler, "Hadronization corrections to jet cross-sections in deep inelastic scattering", in *Workshop on Monte Carlo Generators for HERA Physics (Plenary Starting Meeting)*. 1998.

[94] CMS Collaboration, "Jet algorithms performance in 13 TeV data", CMS Physics Analysis Summary, CERN, Geneva, 2017. https://cds.cern.ch/record/2256875.

[95] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *Journal of Instrumentation* **13** (2018) doi:10.1088/1748-0221/13/05/p05011.

[96] A. J. Larkoski, G. P. Salam, and J. Thaler, "Energy correlation functions for jet substructure", *Journal of High Energy Physics* **2013** (2013) doi:10.1007/jhep06(2013)108.

[97] I. Moult, L. Necib, and J. Thaler, "New Angles on Energy Correlation Functions", MIT-CTP-4825, 2016 arXiv:1609.07483.

[98] J. Dolen et al., "Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure", *Journal of High Energy Physics* **2016** (2016) doi:10.1007/jhep05(2016)156.

[99] R. A. Fisher, "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P", *Journal of the Royal Statistical Society* **85** (1922) 87–94, doi:10.2307/2340521.

[100] K. Beckeret al., "Precise predictions for boosted Higgs production", 2021.

[101] J. Butterworth et al., "Pdf4lhc recommendations for lhc run ii", *Journal of Physics G: Nuclear and Particle Physics* **43** (2016) doi:10.1088/0954-3899/43/2/023001.

[102] V. Khachatryan et al., "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV", *Eur. Phys. J. C* **75** (2015) doi:10.1140/epjc/s10052-015-3351-7, arXiv:1412.8662.

[103] CMS Collaboration, "Search for boosted Higgs bosons produced via vector boson fusion in the $H \to b\bar{b}$ decay mode using LHC proton-proton collision data at $\sqrt{s} = 13$ TeV", technical report, CERN, Geneva, 2023. http://cds.cern.ch/record/2866501.

[104] CMS Collaboration, "Measurement and interpretation of differential cross sections for Higgs boson production at $\sqrt{s} = 13$ TeV", *Phys. Lett. B* **792** (2019) doi:10.1016/j.physletb.2019.03.059.

[105] H. Qu and L. Gouskos, "Jet tagging via particle clouds", *Physical Review D* **101** (2020) doi:10.1103/physrevd.101.056019.

[106] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition* **30** (1997) doi:https://doi.org/10.1016/S0031-3203(96)00142-2.

Chapter A

Mass distortion following $N_2^{1,\mathrm{DDT}}$ selection

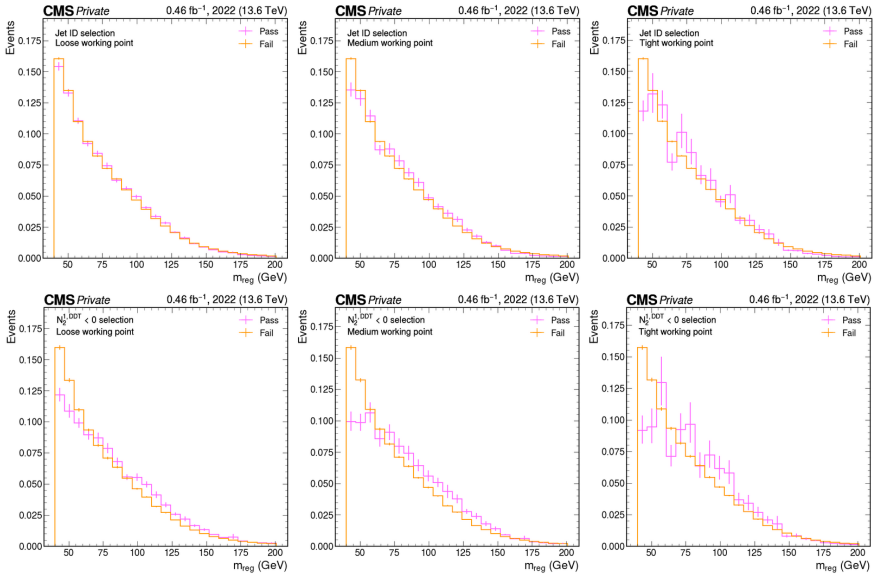
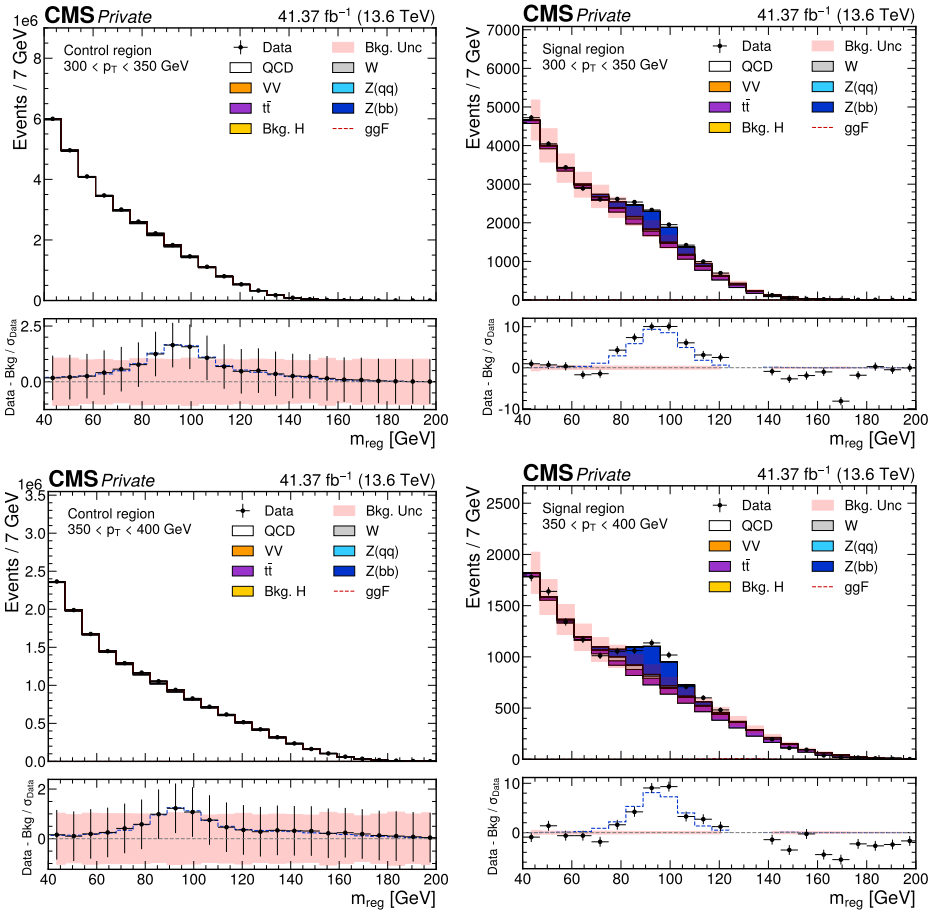


Figure A.1: Normalised event yield as a function of m_{reg} for collision data collected by the scouting stream in 2022. The upper and lower rows display the events before and after requiring $N_2^{1,\mathrm{DDT}} < 0$, respectively. The signal and control regions are shown in pink and orange. The left, middle and right columns display the partitioning into regions by the loose, medium and right $\mathrm{DDB_S}$ WPs, respectively.

As displayed in Fig. A.1, a distortion of the m_{reg} distribution is noticeable after applying the $N_2^{1,\mathrm{DDT}} < 0$ selection. The figure shows the signal and control regions overlaying each other. Before the selection (upper row) the distributions follow each others pattern. In contrast, after the selection (lower row), the distribution shapes of the signal and control regions differ. Notably, the distribution is shifted towards higher mass values for the signal region. The source of this effect is not yet known.

Chapter B

Searching for boosted $Z \to b\bar{b}$



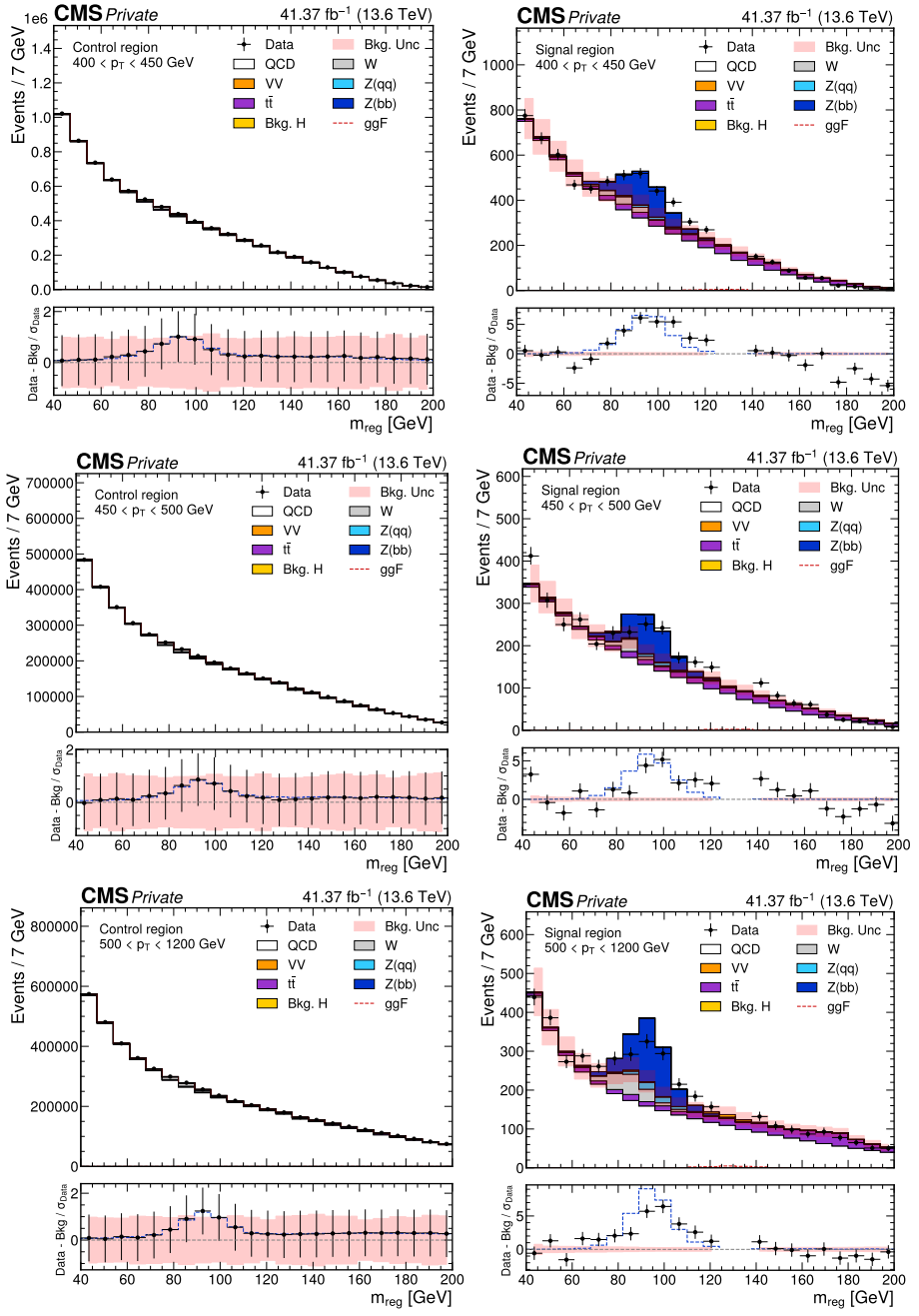


Figure B.1: Data and fitted m_{reg} distribution, summed over all data-taking periods, displayed separate for each p_{T} bin. The control (left) and signal (right) regions are shown. The Higgs boson mass window ($[120, 130]$ GeV) in the signal region is concealed. The $Z \to b\bar{b}$ event yield divided by the statistical uncertainty of the collision data is displayed with a dashed blue line in the lower panel.

Chapter C

Extension of jet ρ region

As a consequence of lowering the jet p_{T} threshold from 450 GeV to 300 GeV in the analyses reported in Chapter 12, jets with masses around 125 GeV (the Higgs mass) are discarded due to the constraints on the jet ρ variable. As discussed in Section 11.3.3, Higgs candidate jets involved in the boosted VBF $H \to b\bar{b}$ analysis are required to have $-6 \leq \rho \leq -2.1$ to avoid instabilities at the edges of the ρ distribution. However, in order to lower the p_{T} threshold, it is necessary to extend the ρ region. In the following text, three concerns regarding this extension are addressed. These concerns are experimental biases potentially introduced due to:

1. Finite cone effects.
2. Unsatisfactory QCD modelling.
3. Degradation of the jet mass scale and resolution.

C.0.1 Finite cone effects

The impact of finite cone effects is studied by computing the angular distance (ΔR) between the two b quarks, stemming from the Higgs boson decay, as a function of the leading jet ρ. The study is performed with simulated Higgs boson events assuming the detector conditions of 2016. The results are presented in Fig. C.1. A horizontal dashed line marks the distance parameter cut-off relevant to AK8 jets (0.8), while two vertical lines denote $\rho = -2.1$ and $\rho = -1.7$. The results show that an extension of the

ρ region from -2.1 to -1.7 greatly increases the number of boosted Higgs boson jets. The extension is therefore not affected by finite cone effects.

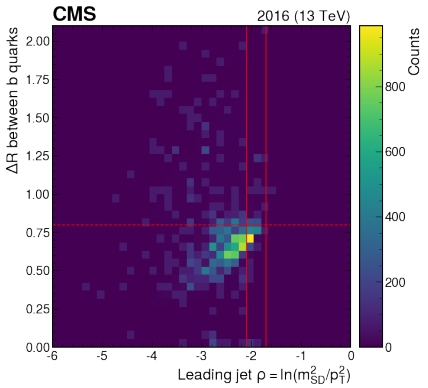


Figure C.1: ΔR between the two b quarks of the Higgs boson decay as a function of the leading jet ρ. The dashed horizontal line denotes the distance parameter (0.8) of the reconstructed AK8 jets. The two solid vertical lines denote ρ equals -2.1 and -1.7.

C.0.2 QCD modelling

The impact of the QCD modelling is assessed by computing the $N_2^{1,\mathrm{DDT}}$ map, as detailed in Section 11.3.4. Maps are constructed using (a) collision data collected in 2016 and (b) QCD multijet events simulated assuming the detector conditions of 2016. The two maps are then compared by considering their difference. The results are shown in Fig. C.2. Notably, no clear difference is evident between $-6 \leq \rho \leq -1.7$. As a result, extending the ρ region to -1.7 is considered to be unaffected by the QCD modelling.

C.0.3 Degradation of the jet mass scale and resolution

The impact of the degradation of the JMS and JMR is assessed by computing the scale and resolution as functions of the leading jet ρ. The analysis is performed with simulated Higgs boson and Z' events assuming the detector conditions of 2016. The Z' sample is simulated over several masses yielding a flat distribution in the jet mass. Fig. C.3 displays the JMS, computed as the jet m_{SD} divided by the relevant particle mass. Meanwhile, Fig. C.4 shows the jet m_{SD} resolution estimated as described in Section 7.6.

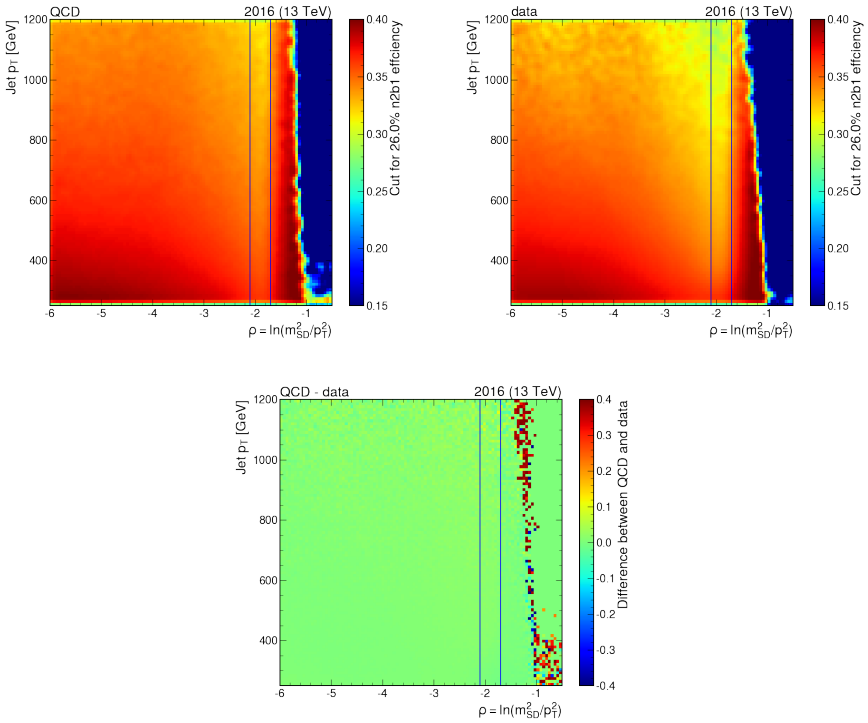


Figure C.2: The $N_2^{1,\mathrm{DDT}}$ map using simulated QCD multijet events (top left), collison data (top right) and their difference (bottom). The two vertical lines indicate $\rho = -2.1$ and $\rho = -1.7$.

Notably, the results do not show a degradation of the JMS or JMR within $-6 \leq \rho \leq -1.7$.

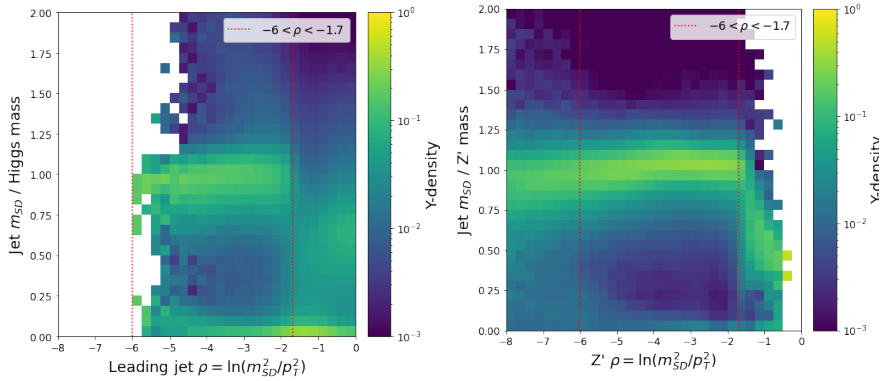


Figure C.3: The normalised event yield as a function of leading jet ρ and m_{SD} of the Higgs boson (left) and Z' (right) samples. The two vertical lines denote $\rho = -6$ and $\rho = -1.7$.

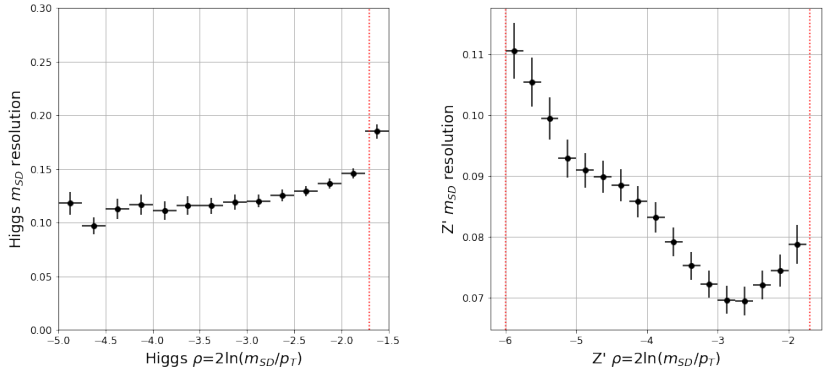


Figure C.4: The m_{SD} resolution as a function of the leading jet ρ of the Higgs boson (left) and Z' (right) samples. The rightmost vertical dashed line denote $\rho = -1.7$.