Bachelor Thesis in Physics

Exploring jet-only final states in the Model Unspecific Search in CMS (MUSiC) using 2018 data

written by

Nils Jonathan Esper

presented to

The Faculty of Mathematics, Computer Science and Natural Sciences
of RWTH Aachen University

III. Institute of Physics A

supervised by

Univ.-Prof. Dr. rer. nat. Thomas Hebbeker

and as second reviewer

Univ.-Prof. Dr. rer. nat. Martin Erdmann

Aachen, August 2023

Abstract

The Model Unspecific Search in CMS (MUSiC) is an analysis that systematically searches for deviations of data recorded at CMS with respect to Monte Carlo (MC) simulations, which describe the processes predicted by the Standard Model. Events are filled in event classes according to the particle multiplicity of the final state. Past MUSiC analyses usually required at least one lepton or photon in the final state. This thesis performs the first study on lifting this requirement and including jet-only classes to MUSiC with Run 2 CMS data collected in 2018 at $\sqrt{s} = 13\,\mathrm{TeV}$, and the first study on this topic in general since 2015. Different QCD MC datasets and jet trigger configurations are used. A significant, systematic deviation between data and MC is observed in the order of 50 %. A large constant fraction of this deviation is identified, as well as dependencies on energy-like quantities (jet transverse momenta sum and invariant mass) and the jet multiplicity of the events. The cause of this deviation could not be identified, however, new physics is regarded as an unlikely origin. A wide jet merging approach and differential pseudorapidity cut, adapted from dedicated CMS jet analyses, is applied and found to significantly reduce both the energy and jet multiplicity dependence. To counter the remaining approximately constant deviation, a normalization approach with an independent dataset is proposed. This strategy leads to improved agreement between data and MC simulation. The signal bias of the analysis steps is regarded as small, however, the performance of a dedicated signal bias study is recommended if the efforts presented in this thesis should be continued.

Zusammenfassung

Die Model Unspecific Search in CMS (MUSiC) ist eine Analyse, die systematisch aufgenommene CMS-Daten und Monte-Carlo-Simulationen (MC), welche die Prozesse im Standardmodell beschreiben, auf Abweichungen untersucht. Die Ereignisse werden gemäß ihrem Teilcheninhalt verschiedenen Ereignisklassen zugeordnet. Vergangene Analysen haben nur Ereignisse mit mindestens einem Lepton oder Photon verwendet. Diese Arbeit führt die erste Studie mit CMS-Daten aus dem Run 2 von 2018 bei einer Schwerpunktsenergie von $\sqrt{s} = 13\,\mathrm{TeV}$ durch, bezüglich Ereignisklassen, welche ausschließlich Jets enthalten. Verschiedene QCD MC Datensätze und Jet-Trigger-Konfigurationen werden verwendet. Eine signifikante, systematische Abweichung zwischen Daten und MC in der Größenordnung 50 % wird beobachtet. Ein großer Teil dieser Abweichung scheint konstant für alle Klassen und Verteilungen zu sein, wobei auch Abhängigkeiten von Energiegrößen (Transversalimpuls-Summe und Masse) sowie der Anzahl an Jets im Event identifiziert werden. Der Grund für diese Abweichung kann nicht genau ermittelt werden, jedoch wird neue Physik als Grund für unwahrscheinlich gehalten. Ein Wide-Jet-Verfahren und eine Pseudorapiditäts-Bedingung, abgeändert von veröffentlichten CMS Analysen mit Jets, wird angewendet. Es zeigt sich, dass die beiden Abhängigkeiten der Abweichung damit deutlich reduziert werden können. Um die verbleibende, konstante Abweichung zu verringern, wird ein Normierungsverfahren mit einem unabhängigen Datensatz entwickelt. Mit dieser Strategie kann eine verbesserte Übereinstimmung zwischen Daten und MC erreicht werden. Die Beeinflussung eines potentiellen Signals wird als gering eingeschätzt, jedoch wird empfohlen, eine Studie bezüglich der potentiellen Beeinflussung durchzuführen, sollten die in dieser Arbeit entwickelten Analyseschritte in Zukunft aufgegriffen und weiterverwendet werden.

Contents

Physical units

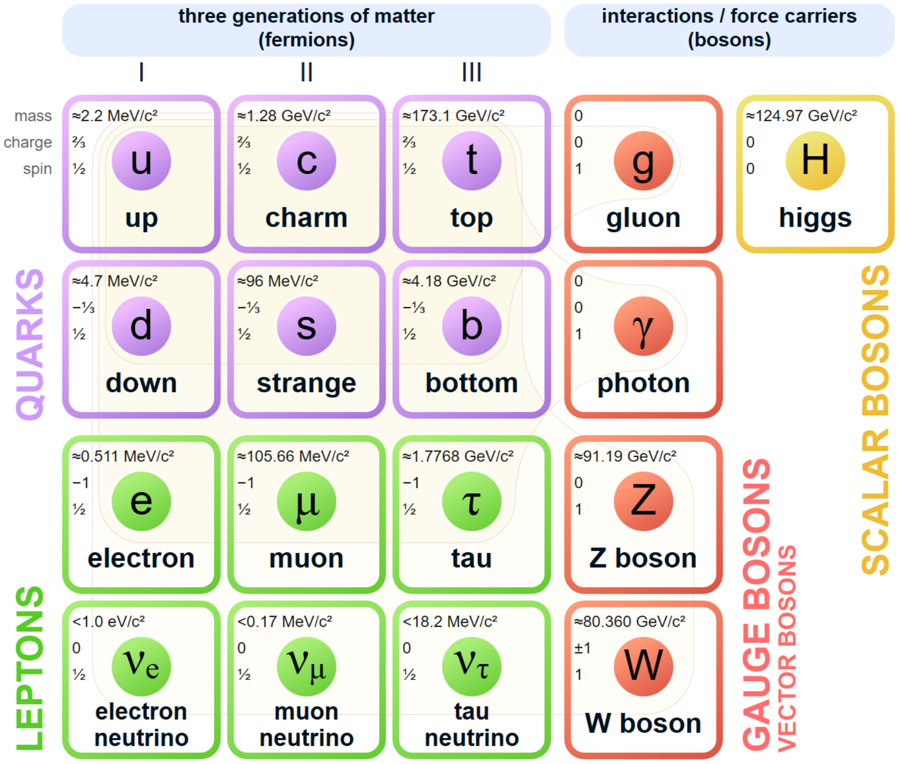
This thesis uses the natural units system, which is conventionally used in particle physics. It is defined by setting the reduced Planck constant and the speed of light to one: $\hbar = c = 1$.

As a consequence, many important measured quantities in particle physics have the unit of energy, namely momentum, and mass. Usually, electron volts (eV), or rather giga electron volts ($1\,\mathrm{GeV} = 10^{9}\,\mathrm{eV}$), are used as an energy unit. The conversion to the SI energy unit Joule is: $1\,\mathrm{eV} \mathrel{\widehat{=}} 1.602\,176\,634 \times 10^{-19}\,\mathrm{J}$ [1].

1 The Standard Model

1.1 Overview

The Standard Model (SM) of particle physics was developed in the 1970s and unifies multiple theories of particles and their interactions [2]. It consists of 17 elementary particles and embeds three fundamental interactions[1]. Over the past decades, all particles predicted by the SM have gradually been discovered [3, 4], the last one being the Higgs boson discovered in 2012 [5, 6]. As the name suggests, all elementary particles are expected to have no substructure, which has also not been observed to date [7]. The theory can explain most of the experimental results in particle physics and is thus considered as very established [3]. Fig. 1.1 shows all particles of the SM sorted in their respective groups.



three generations of matter (fermions)
interactions / force carriers (bosons)
I
II
III
mass
charge
spin
≈2.2 MeV/c² ⅔ ½ u up
≈1.28 GeV/c² ⅔ ½ c charm
≈173.1 GeV/c² ⅔ ½ t top
0 0 1 g gluon
≈124.97 GeV/c² 0 0 H higgs
≈4.7 MeV/c² −⅓ ½ d down
≈96 MeV/c² −⅓ ½ s strange
≈4.18 GeV/c² −⅓ ½ b bottom
0 0 1 γ photon
≈0.511 MeV/c² −1 ½ e electron
≈105.66 MeV/c² −1 ½ μ muon
≈1.7768 GeV/c² −1 ½ τ tau
≈91.19 GeV/c² 0 1 Z Z boson
<1.0 eV/c² 0 ½ νe electron neutrino
<0.17 MeV/c² 0 ½ νμ muon neutrino
<18.2 MeV/c² 0 ½ ντ tau neutrino
≈80.360 GeV/c² ±1 1 W W boson
QUARKS
LEPTONS
GAUGE BOSONS
VECTOR BOSONS
SCALAR BOSONS

Fig. 1.1: List of particles in the Standard Model of particle physics [8, modified]. Matter particles (fermions) are shown in the left block and bosons are on the right. The matter particles are divided into quarks (purple) and leptons (green), while the exchange particles consist of vector bosons (red) and the scalar Higgs boson (yellow). The thin lines in the background indicate groups of particles that interact with the respective exchange particle.

It should be noted that the SM predicts that there exists a respective antiparticle for each particle, which is not shown in the figure. Antiparticles are foretold to have the same properties (e.g. mass and lifetime) except for the physical charges (e.g. electric charge and color charge), which are inverted [9]. They are either referred to with the prefix ”anti-” following the name of the original particle,

[1] Two of the three interactions in the SM, the electromagnetic and the weak interaction, have been combined to the electroweak interaction (see sec. 1.3.1). However, in this section they are referred to as two different interactions.

denoted by the charge (index $^+$ and $^-$), or with a bar over the particle symbol. Particles and their respective antiparticles are counted as only one particle type, yielding the particle count of 17 in the SM. Interactions between multiple antiparticles as well as between "normal" particles and antiparticles follow the same laws as for interactions between "normal" particles [10]. Not all particles have an antiparticle, some particles are considered their own antiparticle. In the following, all particles and fundamental interactions described by the model will be briefly introduced. Since this thesis focuses on the investigation of final states with jets, an additional section in this chapter briefly discusses the mechanisms involved in jet production within the SM. Finally, some general problems with the model will be stated, motivating the search for new phenomena in particle physics.

1.2 Particle content

1.2.1 Matter particles

The SM incorporates 12 matter particles. These particles are fermions, meaning they have a half-odd-integer spin, more precisely all elementary fermions in the SM have spin $1/2$. As fermions, these particles obey the Fermi-Dirac statistics and Pauli's exclusion principle [11]. Matter particles can be divided into two subgroups, six quarks, and six leptons. All quarks (u, d, c, s, t, b) carry an electric charge of $+2/3\ e$ or $-1/3\ e$ while three leptons (e, μ, τ) carry an electric charge of $-1\ e$. Three leptons are electrically neutral, these particles are called neutrinos (ν_e, ν_μ, ν_τ)[2]. In fact, the minimum SM assumes the neutrinos to be massless, however, observed neutrino oscillations suggest non-zero masses [12]. Quarks also carry one of three possible color charges, while leptons do not [9]. Generally, there exist three generations of matter particles. With increasing generation, the masses of the particles increase. Except for the neutrinos, only the first-generation particles are stable, while higher-generation particles quickly undergo decay to lower-generation particles [3, 13]. Because of this, most of the visible matter is, according to current knowledge, composed of first-generation matter particles. Up and down quarks (u, d) form protons and neutrons, the components of the atomic nuclei, while electrons (e) make up the atomic shell [3, 10]. For the matter particles, the corresponding antiparticles are the antiquarks and antifermions. An exception in the described naming scheme for antiparticles exists for the electron, whose corresponding antiparticle is called a positron.

1.2.2 Exchange particles

Four particles in the SM are exchange particles, meaning that they act as carriers of the three interactions described in the model. All exchange particles as well as the Higgs particle are classified as bosons, meaning that they have integer spin and obey the Bose-Einstein statistics. More specifically for the exchange particles, the spin is 1 and they are referred to as vector bosons. Photons (γ) are massless particles that act as exchange particles for the electromagnetic force. Since they have no mass, they move at the speed of light c and have an unlimited lifetime. Gluons (g) are the exchange particles of the strong interaction and are also massless. They carry combinations of color charges, leading to eight different possible gluons. The exchange particles of the weak interaction are the Z^0 and $W^\pm$ bosons which both have a comparably high mass. While the Z^0 boson is electrically neutral, the $W^\pm$ bosons carry an electric charge of $\pm1\ e$ [9]. The $W^\pm$ bosons are the respective antiparticles of each other, while the photon, gluon, and the Z^0 boson are said to be their own antiparticles [9].

[2] e is the elementary charge defined as the absolute value of the charge of an electron. It should not be confused with the symbol for the electron e, which is the same letter.

1.2.3 Higgs boson

The mass of the exchange particles in the SM, namely the massive $W^{\pm}$ and Z^0 bosons, was found to be incompatible with the requirement of gauge invariance. In 1964, a scalar field was introduced, which potentially could explain the particle masses with symmetry breaking [14–19]. Today, this field is known as the Higgs field. The mechanism does not only affect the massive exchange particles but also contributes to the mass of the fermions in the SM, especially for higher mass particles. A consequence of this theory is the addition of another massive, electrically neutral particle with spin 0 (scalar boson) to the SM, the Higgs boson (H^0) [9]. After a multiple-decade search, a new particle with properties consistent with the predicted Higgs boson was finally found by ATLAS [5] and CMS [6] in 2012. The Higgs boson is also considered to be its own antiparticle.

1.3 Fundamental Interactions

1.3.1 Electroweak interaction

As stated in the previous section, the electroweak interaction was initially regarded as two different interactions. Firstly, there is the electromagnetic interaction (also referred to as QED for Quantum Electrodynamics), mediated by photons (γ). All particles that have a non-zero electric charge can participate in this interaction. This includes all quarks, the three charged leptons (e, μ, τ) as well as the $W^{\pm}$ bosons, and of course all corresponding antiparticles. A photon can either be absorbed or emitted, or a particle-antiparticle pair can be created from an existing photon [11]. The electromagnetic interaction has an unlimited range, which is a result of the vanishing photon mass [13, 20]. Since the photon itself is electrically neutral, it can not interact with other photons. The electromagnetic interaction creates bound states, e.g. in the form of stable atoms.

Secondly, there is the weak interaction, which has multiple exchange particles, the $W^{\pm}$ and Z^0 bosons. Because of the mass of these exchange particles, the range of the weak force is reduced to approximately $10^{-17}\,\mathrm{m}$ [13]. The establishment of a theory of weak interaction and the introduction of neutrino particles originate from the search for an explanation for radioactive beta decay [21]. All matter particles including neutrinos can participate in weak interaction processes, even the weak exchange particles themselves as well as all corresponding antiparticles. Interaction processes can be divided into weak charged currents, mediated by the electrically charged $W^{\pm}$ bosons, and weak neutral currents, mediated by the Z^0 boson. It should be noted that the chirality of neutrinos is constrained in the SM, only left-handed neutrinos and right-handed antineutrinos are allowed which puts constraints on weak interaction processes [13]. Unlike the other fundamental interactions, the weak interaction can change the particle flavor [21]. The weak interaction is not known to form any bound states.

In the 1960s, efforts were made to combine the electromagnetic and the weak interaction to one theory of electroweak interaction. Incorporating the idea of the Higgs field (see sec. 1.2.3) to explain the mass of the weak exchange particles, the unification succeded in 1968, resulting in the Glashow-Salam-Weinberg theory of electroweak interaction [21].

1.3.2 Strong interaction

1.3.2.1 Introduction

The strong interaction describes the interaction of quarks and gluons. Quantum Chromodynamics (QCD) is the theory related to this interaction. For the theory to explain all related phenomena that have been observed at accelerator experiments[3], an additional quantum number was introduced, the

[3] Namely, the observed Δ^{++} baryon (consisting of three u quarks with spin $+1/2$) would violate Pauli's exclusion law if no additional degree of freedom would have been added [11].

color charge. Quarks can either have red, blue, or green color charge, for the antiquarks exist the respective anti-colors. QCD predicts the existence of a vector boson, the gluon (g), which carries a combination of color charges and plays the role of an exchange particle. All particles that carry a color charge can participate in this interaction, concretely all quarks (as well as the respective antiquarks) and gluons. Note that because gluons carry color charge, they can also interact with other gluons. The strong interaction also creates bound states referred to as hadrons. Possible bound states include mesons (consisting of a quark-antiquark pair) and baryons (consisting of three quarks or antiquarks) [11, 13]. Although gluons are massless, the range of the strong interaction is limited to approximately 10^{-15} m [13]. The term "confinement" describes the observation that quarks only exist in color-neutral bound states [9]. The theory incorporates this with a potential that increases linearly at long distances [22]. Because of this, the energy to separate quarks in a bound state exceeds the energy to create a new color-neutral bound state out of the vacuum, and therefore free quarks should not exist and have also not been observed to date [21].

1.3.2.2 Jet production

Because of its relevance to this thesis, a summary of jet production and properties within the SM should be given in this section. Jets originate from quarks or gluons that are produced during particle interactions. Quarks, created from an interaction of particle collisions, carry momentum and are moving away from each other. However, they can not emerge freely because of the confinement implied by the strong interaction, therefore many color-neutral hadrons are formed along the orientation of the dispersing original quarks. The creation of these particle showers from the colored quarks in the reaction's final state is called hadronization[4]. The resulting showers of quarks approximately preserve the four-momentum of the origin quark [9, 23]. When studying the hadronization, the possible decay of the resulting hadrons in the shower should be considered, adding an extra level of complexity [23]. The term jet describes one of these collimated particle showers.

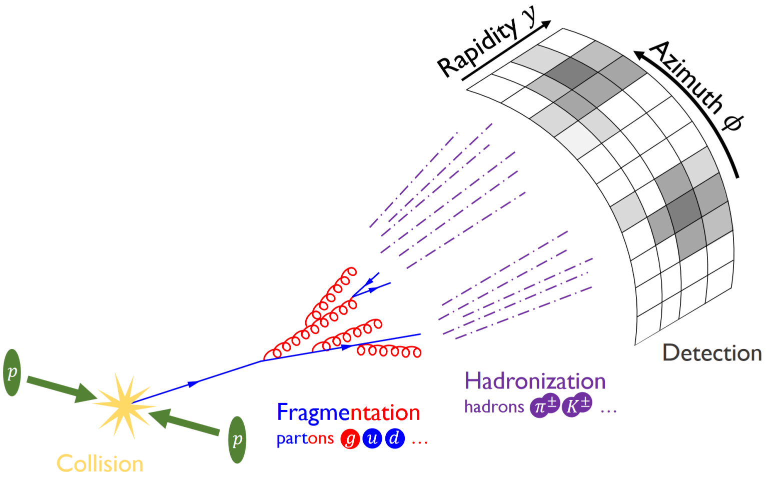


Fig. 1.2: Schematic view of jet production [24, p. 30]: Final state quarks from the original particle reaction after the proton-proton collision undergo fragmentation by emitting gluons and quark-antiquark pairs. Finally, hadron showers are created during hadronization. These showers are observed in the detector.

In the same way as photons can be emitted from charged particles in QED, gluons can be emitted from quarks in QCD [13]. This process was sometimes even called "gluon bremsstrahlung" [25] in analogy to the electromagnetic partner process. The emitted gluons can form additional quark-antiquark

[4] There is an exception for one quark in the hadronization process. Because the lifetime of the t quark is lower than the time scale at which hadronization takes place, it does not hadronize [21].

pairs. Therefore, a quark final state gains complexity because of these gluon emissions even before the hadronization happens. This process of gluon emission is referred to as fragmentation[5]. This fragmentation can lead to the formation of additional separate jets if gluons are emitted at sufficiently large angles from the original final state quark. In fact, observed three-jet events in final states with two quarks led to the discovery of the gluon at DESY in 1979 [25, 26]. Particle jet production is hard to model in general, even within the SM the understanding and simulation of the hadronization process is still an active field of research. Fig. 1.2 shows a schematic view, illustrating the jet production process. It should be noted that the discussed fragmentation and hadronization may happen simultaneously, however, to illustrate the different process steps, they were presented one after another.

1.4 Incompleteness and issues

Although the SM is considered a successful theory, providing an explanation of the results of many particle physics experiments and even predicting the existence of particles before their experimental discovery [4], it has severe deficits considering theoretical aspects and experimental observations that seem not to fit within the model [21]. Some important points should be listed here to motivate the further search for new phenomena and theories in particle physics.

Gravity is the fourth known fundamental force in the current understanding of physics. However, this interaction is not included in the SM, since the formulation of a theory of gravitation (like the theory of General Relativity accomplished for a non-quantum-theory) as a quantum theory did not succeed in the past, because of multiple conceptual and mathematical problems [21]. While this is a severe issue from a theory standpoint, it is commonly assumed that gravitational interaction, in its current understanding, would be neglectable in interactions of elementary particles because of its very small strength compared to the other fundamental interactions [4]. From a theoretical point of view, it is also desired to unify the strong and the electroweak interaction as it was accomplished with QED and the weak interaction. Multiple theories have been proposed, but the question has not been resolved yet [4, 11]. Various other problems exist from a theoretical perspective, including the so-called hierarchy problem raising questions regarding the validity of the SM because of inconsistencies related to the Higgs boson mass [4, 27]. As already stated, experimental results strongly indicate the existence of neutrino oscillations, which would imply that neutrinos are massive particles [12]. Since the minimum SM assumes neutrinos to be massless, the theory has to be expanded to incorporate the neutrino mass [4, 11]. Apart from this, cosmological observations suggest that the visible matter, which the SM attempts to describe, only makes up about 5 % of the components of the universe. Therefore a very large fraction of the components of the universe, namely about 26 % dark matter and about 69 % dark energy, are beyond the scope of the SM [4, 11].

From this selection of issues, it can already be concluded that the search for new physics phenomena and theories is very much required to improve the understanding of the physics of elementary particles and our cosmos in general.

[5] The terms hadronization and fragmentation are not used consistently in literature. Sometimes they are even used interchangeably.

2 Experimental Setup

2.1 The Large Hadron Collider

2.1.1 Introduction and history

The Large Hadron Collider (LHC) is the most powerful particle accelerator in the world, currently reaching center-of-mass energies of up to $\sqrt{s} = 13.6\,\mathrm{TeV}$. Design values are even higher with an energy of $\sqrt{s} = 14\,\mathrm{TeV}$ and a maximum luminosity of $\mathcal{L} = 10^{-34}\,\mathrm{cm}^{-1}\mathrm{s}^{-1}$ [28]. It is located in a tunnel on average 100 m underground in the Franco-Swiss border area, at CERN, the European Organization for Nuclear Research [29]. The tunnel was previously used by the Large Electron-Positron Collider (LEP). After its operation had ended in 2000, the LHC was constructed in the same tunnel and declared operational in 2008 [30]. The LHC is a synchrotron collider with two particle beams accelerated in opposing directions along a circumference of 27 km. Particles travel along the ring in two beam pipes, which incorporate an ultra-high vacuum with pressures of only 10^{-13} bar. To keep them on the circular track[1], 1232 dipole magnets with a magnetic field strength of 8.33 T are used [27, 31]. Additionally, quadrupole and octopole magnets are used to focus the particle beam along its track. The superconducting coils of the magnets are cooled down to as low as 1.5 K to create these strong magnetic fields [27]. Collision experiments with both lead ions and protons are conducted at the LHC. The particles are not accelerated by the LHC immediately, instead, a series of pre-accelerators is used before the protons are induced in the large accelerator [29]. LHC operation and the data-taking periods are divided into multiple parts. The first data-taking period (Run 1) ranged from 2009 to 2013 at $\sqrt{s} = 7 - 8\,\mathrm{TeV}$, the second one (Run 2) from 2015 to 2018 at $\sqrt{s} = 13\,\mathrm{TeV}$, and during the writing of this thesis the third data taking period (Run 3) is still ongoing, having started in 2022 at $\sqrt{s} = 13.6\,\mathrm{TeV}$ [32, 33]. To date, the observation of the Higgs boson is considered the biggest accomplishment of the LHC [5, 6]. Fig. 2.1 shows an overview of the CERN accelerator complex with the LHC and many other accelerators and experiments.

2.1.2 Proton acceleration

Since this thesis only uses recorded data from 2018 (part of Run 2) with proton-proton collisions, only the proton operation at the LHC for Run 2 should be briefly described in this section. The proton source for LHC is hydrogen atoms from a gas bottle, which are first ionized and accelerated to 50 MeV in the Linear accelerator 2 (LINAC 2)[2] [35]. The next stage is the Proton Synchrotron Booster (BOOSTER), where the protons are accelerated to 1.4 GeV[3] [36] and injected in the Proton Synchrotron (PS). Here, the proton energy is further increased to 25 GeV [31, 37]. The last pre-accelerator is the Super Proton Synchrotron (SPS) which brings the proton energy up to 450 GeV when they are finally injected into the LHC [31, 38]. In the large accelerator, the protons experience acceleration up to the design energy maximum of 7 TeV before being brought to collision (for $\sqrt{s} = 13\,\mathrm{TeV}$ in Run 2, the maximum energy was 6.5 TeV) [32]. It should be pointed out that protons in

[1] In fact, the LHC tunnel does not emerge as a perfect circle, instead there are multiple sections of the tunnel, some of which are unbowed and others curved [31].

[2] Since Run 3, the Linear accelerator 4 (LINAC 4) is used, accelerating negative hydrogen ions instead of protons for the first acceleration stage. However, for the previous LHC runs, the LINAC 2 was used.

[3] This value and the following values belong to LHC Run 2, they might deviate for the current Run 3.

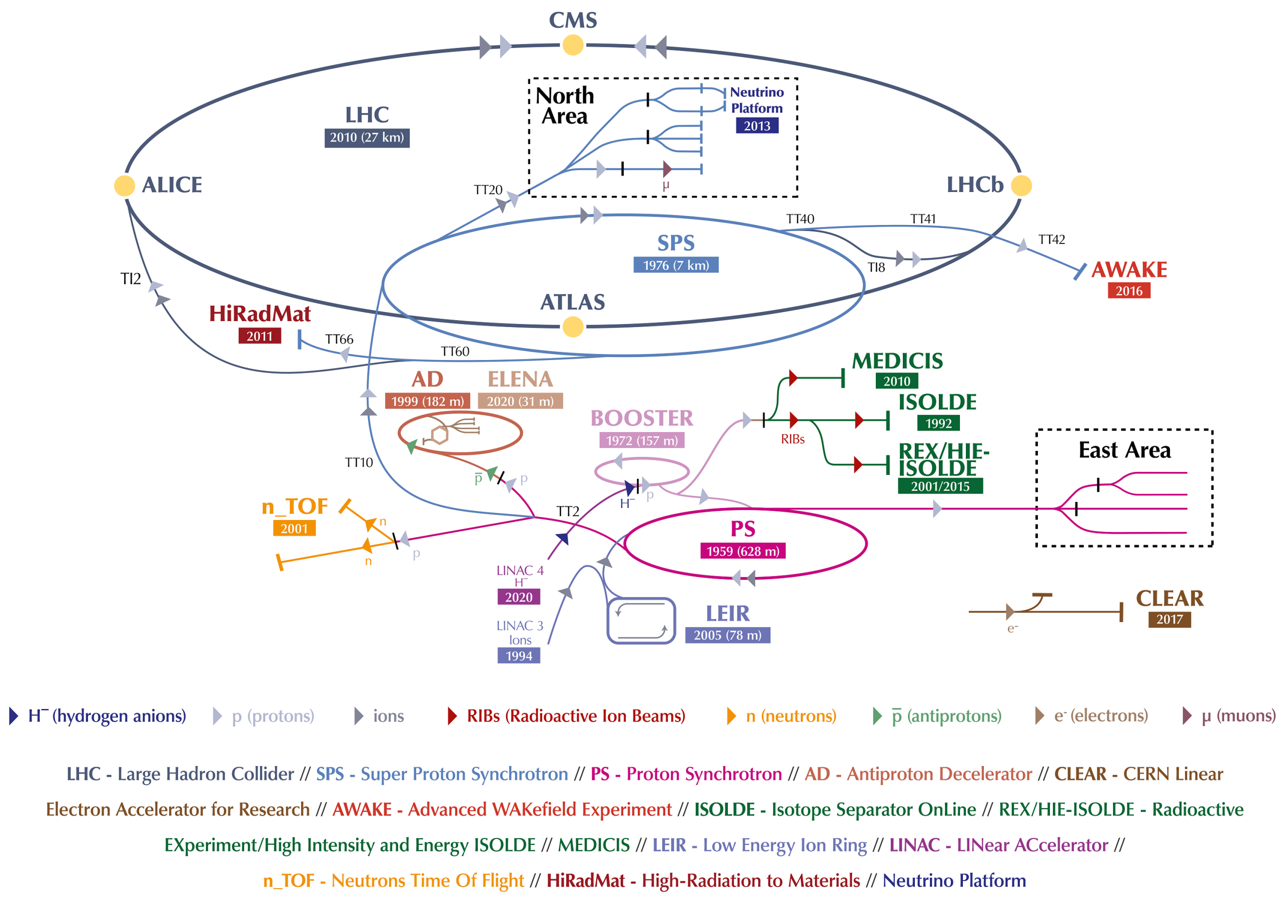

CMS
North Area
Neutrino Platform
2013
μ
LHC
2010 (27 km)
ALICE
LHCb
TT20
TT40
TT41
TT42
AWAKE
2016
SPS
1976 (7 km)
TI8
TI2
HiRadMat
2011
ATLAS
TT66
TT60
MEDICIS
2010
ISOLDE
1992
RIBs
REX/HIE-ISOLDE
2001/2015
East Area
AD
1999 (182 m)
ELENA
2020 (31 m)
BOOSTER
1972 (157 m)
TT10
p̄
p
H⁻
n_TOF
2001
n
TT2
PS
1959 (628 m)
LINAC 4
H⁻
2020
LINAC 3
Ions
1994
LEIR
2005 (78 m)
CLEAR
2017
e⁻
H⁻ (hydrogen anions)
p (protons)
ions
RIBs (Radioactive Ion Beams)
n (neutrons)
p̄ (antiprotons)
e⁻ (electrons)
μ (muons)
LHC - Large Hadron Collider // SPS - Super Proton Synchrotron // PS - Proton Synchrotron // AD - Antiproton Decelerator // CLEAR - CERN Linear Electron Accelerator for Research // AWAKE - Advanced WAKefield Experiment // ISOLDE - Isotope Separator OnLine // REX/HIE-ISOLDE - Radioactive EXperiment/High Intensity and Energy ISOLDE // MEDICIS // LEIR - Low Energy Ion Ring // LINAC - LINear ACcelerator // n_TOF - Neutrons Time Of Flight // HiRadMat - High-Radiation to Materials // Neutrino Platform

Fig. 2.1: Schematic image of the current installations at the CERN accelerator complex, the large LHC ring (dark blue) can be seen at the top with the crossing points (yellow) and the names of the four large detectors [34, modified]. Since pre-acceleration was switched from LINAC 2 to LINAC 4 for Run 3, the old linear accelerator used in Run 2 cannot be found in the image.

the LHC are accelerated and collided in about 2000 bunches of approximately 10^{11} protons, which are distributed along the length of the accelerator [39].

2.1.3 Particle detectors

At four points of the LHC, the two particle beams cross each other, resulting in collisions. To observe these collisions and the created particles as a result, the LHC features seven[4] different experiments. Three of these are smaller experiments, which observe particles in the forward region of the collision points or simply the beamline particles [40]:

- Total, elastic and diffractive cross-section measurement (TOTEM): Measuring slightly deflected protons in the forward regions.
- Large Hadron Collider forward (LHCf) experiment: Simulation of cosmic rays under laboratory conditions.
- Monopole and Exotics Detector at the LHC (MoEDAL): Search for magnetic monopoles and other exotic particles.

At the crossing points, the four large LHC detectors are positioned [32, 40]:

- A Toroidal LHC Apparatus (ATLAS): Largest general-purpose detector at LHC.

[4] This is correct for Run 2. For the current Run 3, two additional experiments, the Forward Search Experiment (FASER) and the Scattering and Neutrino Detector (SND@LHC) experiment, were added [40].

- Compact Muon Solenoid (CMS): General-purpose detector (details in sec. 2.2).
- Large Hadron Collider beauty (LHCb): Investigation of final states with b quarks.
- A Large Ion Collider Experiment (ALICE): Detector for ion collisions at the LHC.

Since this thesis uses CMS data, only this detector is introduced in more detail in the next section 2.2. Before the details are discussed, one important relation in particle physics should be mentioned. The expected event rate $\dot{N}$ for a process with cross section σ at an experiment with given instantaneous luminosity $\mathcal{L}$ is given as [4]:

$$\dot{N} = \sigma \cdot \mathcal{L}. \tag{2.1}$$

2.2 The Compact Muon Solenoid Detector

2.2.1 Introduction

The CMS detector is a large general-purpose particle detector at the LHC. It measures 28.7 m in length and 15 m in diameter [41]. ATLAS and CMS have the same purpose of searching for new physics in proton collisions, however, their designs are different [32] and CMS is also used for heavy ion collisions. After introducing the coordinate frame used by CMS, the subsystems of CMS are briefly described, starting with the innermost systems. Generally, the detector can be divided into the barrel section and the two endcaps right and left to the barrel. Fig. 2.2 shows a rendering of the CMS detector with the different detector subsystems. A slice view of the detector can be found in fig. 2.3, a few pages below.

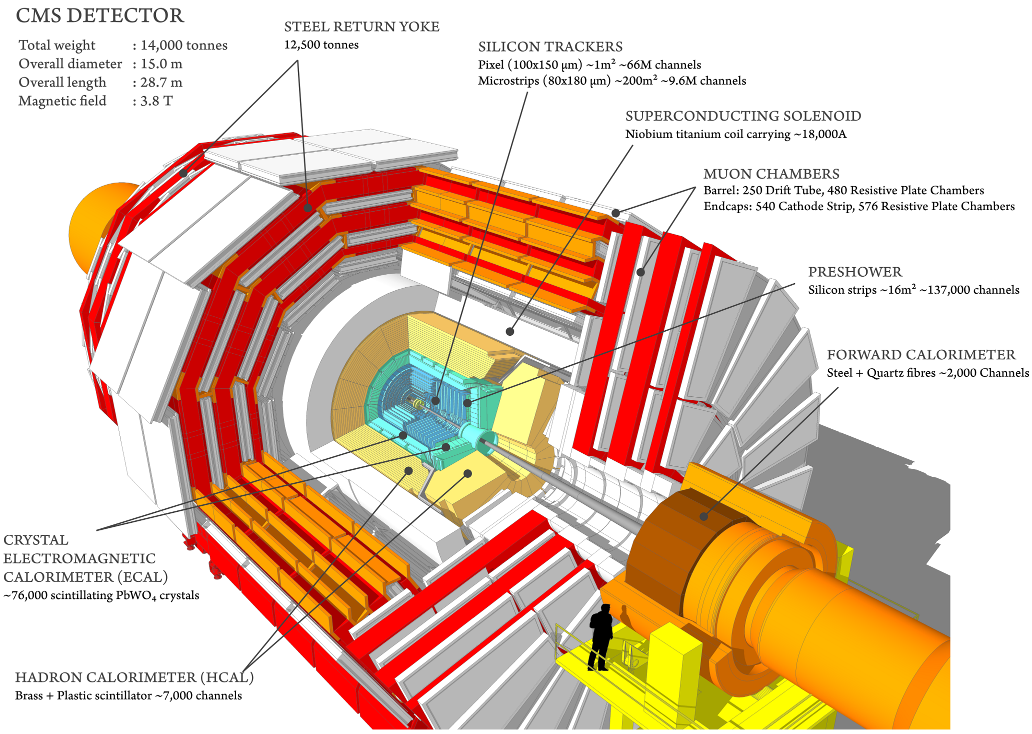

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~1m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Fig. 2.2: Rendering of the CMS detector with all its subsystems and a human to scale [41].

2.2.2 Coordinate system

The center of the CMS coordinate system lies at the nominal interaction point, where the two proton beams are expected to cross. In cartesian coordinates, the x-axis points towards the center of the LHC, the y-axis points upwards, and the z-axis points along the beam line[5]. Adapted polar coordinates are also used in CMS. Radial distance r and azimuthal angle ϕ (in the x-y-plane) correspond to the common spherical coordinates. However, instead of the polar angle θ, the pseudorapidity η is used, defined by the relation [4, 10]:

$$\eta = -\ln \tan \frac{\theta}{2}. \tag{2.2}$$

This has the advantage that in the high-energy limit $E \gg m$, the pseudorapidity η becomes the particle rapidity $y = \operatorname{artanh}(v_z)$[6], which is a Lorentz invariant quantity [4, 27]. The pseudorapidity is also used to conveniently define the angular distance of two objects with angular coordinates (η_1, ϕ_1) and (η_2, ϕ_2) as [27]:

$$\Delta R = \sqrt{(\eta_2 - \eta_1)^2 + (\phi_2 - \phi_1)^2}. \tag{2.3}$$

2.2.3 Detector subsystems

2.2.3.1 Inner tracking system

Particles emerging from collisions exit the beam pipe vacuum and first travel through the tracker. This subsystem detects the tracks of all charged particles at multiple points with an accuracy of 10 µm, enabling precise reconstruction of the particle path. This information is relevant to calculate the particle momentum from the deflection in the magnetic field as well as for reconstructing short-lived particles from their tracked decay products [10, 32]. In CMS, the pixel detector, consisting of four cylindrical barrel sensors and barrel disks, is the first layer of the tracker, with a pixel size of only $150 \times 100\,\mu\text{m}^2$ and a total number of 66 million pixels. It covers an angular range of $|\eta| < 2.5$. Silicon strip detectors are used for the second level of the tracker. These silicon strips have a thickness of 320 µm and a pitch between 80 µm and 141 µm [4, 42].

2.2.3.2 Electromagnetic calorimeter

Calorimeters are generally used to measure particle energies. In calorimeters, the particles deposit energy, depending on the particle type as well as their energy, through various interaction processes. Multiple calorimeter systems are used at CMS. The innermost is the Electromagnetic Calorimeter (ECAL), primarily used to measure energies of photons and electrons [32]. It consists of 61,200 lead tungstate ($PbWO_4$) crystals in the barrel part and 7324 crystals in the endcaps. The material is an inorganic scintillator, therefore the deposited energy is measured by detecting the scintillation photons in the crystals with Avalanche photodiodes and vacuum phototriodes. The ECAL system measures particle energies with angles of $|\eta| < 3.0$ [42].

2.2.3.3 Hadronic calorimeter

The Hadronic Calorimeter (HCAL) is the second calorimeter in CMS, primarily focused on measuring energies of hadrons, e.g. as constituents of particle jets. Hadrons might lose fractions of their energy already in the ECAL, therefore the purpose of the HCAL is to measure all remaining hadron energy. It consists of the Hadron Barrel (HB) and the Hadron Endcap (HE) located one layer above the ECAL as well as the Hadron Outer (HO) which is located outside of the solenoid. Together, the HCAL features more than 70,000 scintillator tiles which are read out with hybrid photodiodes. It covers the same angular range of $|\eta| < 3.0$ as the ECAL[7] [4, 42].

[5] More specifically, the z-axis points from LHC octant 5 to 4 [4]. These octant numbers refer to different tunnel sections.

[6] In natural units, which are used in this thesis, the velocity v_z (in the z-direction) has no dimension since the speed of light $c = 1$ was set to a dimensionless number.

[7] Note that CMS features some forward detectors which can observe particles at higher $|\eta|$ values, which are not listed in the brief overview given in this chapter [42].

2.2.3.4 Superconducting solenoid

The magnetic field needed to measure the momenta of charged particles from the deflected paths is generated by a cylindrical, superconducting 4-layer coil made out of niobium–titanium (NbTi). It is cooled to a temperature of 4.5 K and generates a magnetic field with a flux density of up to 3.8 T [4, 32]. An iron yoke massing about 10,000 t, made out of five wheels and two endcaps, is used to return the magnetic flux and also serves as a support structure for the detectors [42].

2.2.3.5 Muon system

Muon detection is of central importance in the CMS experiment, it is even featured in the experiment name. The muon system is the outermost layer of the detector since muons are heavy particles that can penetrate all inner detector components. It features various gaseous detectors to measure the particle tracks, later allowing calculation of the muon momentum. All gaseous detectors rely on the ionization of gas inside the detector by the passing muons [4]. The choice for these detectors was motivated by the very large detection area of 25,000 m^2 required, while simultaneously restricting the costs [42]. In the barrel region ($|\eta| < 1.2$), mainly Drift Tubes (DT) are used, while the endcap region ($0.9 < |\eta| < 2.4$) features Cathode Strip Chambers (CSC) [4]. Additionally, Resitive Plate Chambers (RPC) are used as a redundant detector type throughout the whole angular range up to $|\eta| < 1.9$ [43]. For high momenta ($p_\mu > 1\,\mathrm{TeV}$), the muon momentum can be measured with a resolution as low as 5 % [42].

2.2.4 Trigger and data aquisition

Proton-proton collisions at the LHC happen at a rate of up to approximately 40 million events per second, since this large number of collisions can not all be computed and stored, it is necessary to reduce the event rate. This task is given to the CMS trigger system. The trigger system incorporates multiple trigger levels. The Level 1 Trigger (L1 Trigger) reduces the forwarded event rate to less than 100 kHz with custom-built electronics, using only parts of the recorded data at lower resolutions. The final selection is performed by the High-Level Trigger (HLT), which has access to all recorded data for the event and essentially consists of about 1000 commercial computer processors [42]. Accepted events by the HLT are then stored for future analysis. CERN operates the Worldwide LHC Computing Grid (WLCG), allowing access and processing of the data for the scientists analyzing the recorded data from all around the world [32].

2.2.5 Object reconstruction

2.2.5.1 Introduction

Usually, when performing an analysis, particle candidates have to be reconstructed from the recorded data. In CMS, the reconstruction is divided into multiple steps. First, the different subdetectors reconstruct energy deposits and tracks separately. This information is combined later, correlating the different "basic elements" [45] of the event. This reconstruction method at CMS is collectively referred to as Particle Flow (PF) [45, 46]. Since jets are of special importance to this thesis, jet reconstruction will be addressed in a separate section (sec. 2.2.5.2), while this section will only provide a brief introduction to the reconstruction of other objects, without aspiration to completeness.

The measured tracks from the tracker system are used to determine the vertices of the event. Muons are reconstructed from the recorded tracks in the muon chambers linked to the tracks of the silicon tracker. For electrons, the particle track is linked with an energy cluster from the ECAL. Since photons carry no electric charge, they are not detected by the tracker system, therefore energy clusters in the ECAL with no corresponding track are classified as photons. Analogously to electrons and photons,

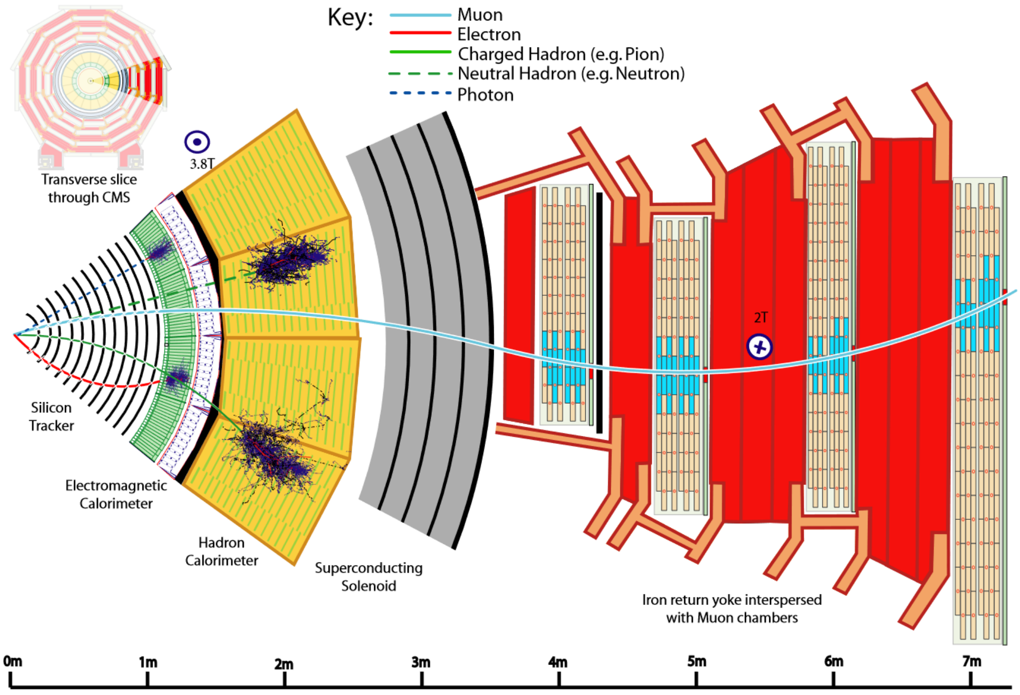

Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
Transverse slice through CMS
3.8T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
0m
1m
2m
3m
4m
5m
6m
7m

Fig. 2.3: Example particle tracks in a slice view of the CMS detector [44, fig. 1]. The various detector subsystems are highlighted in different colors. Characteristic signatures of different particle types are illustrated, which the Particle Flow (PF) algorithm uses to identify particles from the detector outputs.

electrically charged and neutral hadrons are reconstructed [45]. Higher observables, for example Missing Transverse Energy (MET), are then calculated from the reconstructed objects. Fig. 2.3 illustrates the detector signatures for different particles in the CMS detector.

2.2.5.2 Jet algorithms

The reconstruction of jets requires an algorithm that merges the shower particles originating from the quark hadronization to single jet objects. Generally, there are two approaches: Cone-type algorithms simply combine all hadrons within a defined radius, while successive recombination algorithms combine particles iteratively. Common cone-type algorithms include the SISCone algorithm [47], while successive recombination algorithms include the Cambride/Aachen [48], k_T [49] and anti-k_T [50] algorithms [51, 52]. Any jet algorithm is desired to be infrared-safe and collinear-safe since these two properties ensure the validity of perturbation theory in theoretical cross section calculations. Infrared-safe implies that the reconstruction of the jets is independent of the presence of soft gluon emissions, while collinear-safe describes the independence of collinear splitting of particles [23]. The listed algorithms satisfy these requirements.

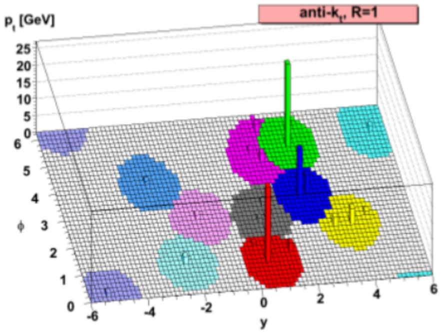


Fig. 2.4: Example of jet reconstruction with the anti-k_T algorithm [51, p. 12]. The algorithm yields circular jets if there are no overlaps within $2R$ with other hard particles.

Since CMS uses the anti-k_T algorithm, only this algorithm will be introduced in this section. For this algorithm, the distances between the objects are defined as [50]:

$$d_{ij} = \min\left\{p_\mathrm{T}(i)^{-2}, p_\mathrm{T}(j)^{-2}\right\} \cdot \frac{\Delta R^2}{R^2} \qquad \text{and} \qquad d_{iB} = p_\mathrm{T}(i)^{-2}, \tag{2.4}$$

where i and j are the indices of the objects to be merged, B stands for the beam, p_T is the corresponding transverse momentum, ΔR is the angular distance between the objects i and j similar as in eq. 2.3, and R is the jet cone radius (in the η-ϕ-plane), which is a constant parameter. These distances are calculated for all objects, then the objects with the smallest distance are merged. This process is repeated unless d_{iB} is the smallest distance. Then, object i is labeled as a jet and is removed from the set with objects still to merge [4, 50]. Reviewing this procedure, the successive nature of the anti-k_T algorithm becomes obvious. Fig. 2.4 shows an example of jet reconstruction with the anti-k_T algorithm.

3 Model Unspecific Search in CMS

3.1 Introduction

Recorded data by CMS and various other experiments is used to verify the SM and search for new physics phenomena, commonly referred to as Beyond the Standard Model (BSM) physics. There exist numerous dedicated analyses, which search for discrepancies from the SM and compare data to simulated signals from BSM models. Usually, a selection of few or only one final state is used in such analyses, which are related to the studied physics process. For example, hypothetical dijet and trijet resonances are studied only in the respective final states [53, 54]. Another example would be various decay modes of particles, e.g. the Higgs boson decay into two photons which is studied in the corresponding final state [55]. However, computing resources and personpower are limited, making it impossible to analyze every final state in reasonable time periods. Besides that, it is also possible that BSM physics reveals itself in different signals as would be expected from existing extending theories. These still unknown signal shapes might not be considered in dedicated analyses. An alternative approach from dedicated analyses is to systematically scan all recorded events for deviations from the SM by comparing it to simulations [10, 32]. These analyses are regarded as model-independent since they do not specify the final state nor the signal shape[1] [4]. This type of analysis dates back to a first study conducted at the LEP in 1998, where L3 data was compared to SM simulation [56]. It should be pointed out, that model-independent analyses are usually less sensitive than fine-tuned dedicated analyses. If a model-independent study should reveal significant deviations from SM simulations, this would trigger a dedicated investigation, studying the cause of the deviation with a desired higher sensitivity [10].

The Model Unspecific Search in CMS (MUSiC) analysis is such a model-unspecific search conducted with CMS data. The analysis considers multiple distributions for hundreds of accessible final states and scans them for possible deviations from the simulation. No complex filters or rigorous kinematic restrictions are set to maximize the analyzed phase space by the analysis [10]. The MUSiC analysis at CMS was introduced in 2008 [57] and is continuously updated and improved to use the most recent recorded data and extend the search [58, 59].

This thesis does not perform a complete MUSiC analysis of a dataset, instead, it aims at exploring a possible extension of the event classes in MUSiC. However, the whole analysis procedure should be briefly presented in the next subsections to lay out the general idea and working principle of MUSiC.

3.2 Analysis procedure

3.2.1 Dataset

The MUSiC analysis requires a full set of recorded data at CMS as well as a complete set of simulated samples, which cover all possible interactions described within the SM. The simulated Monte Carlo (MC) samples are produced by different event generators. Usually, the simulation samples are generated by the CMS collaboration and made public in the central data system, the Data Aggregation

[1] However even model-independent analyses are restricted to the detector limitations and very much rely on the quality of the simulated dataset.

System (DAS), where they can be accessed with a specific DAS key. All generated MC samples as well as the data samples are stored in Analysis Object Data (AOD), MiniAOD, or NanoAOD files. Past MUSiC analyses used MiniAOD files, but current and future analyses will use the NanoAOD file format. The samples used in this thesis are discussed in sec. 4.3, since this chapter should only give an overview over the MUSiC workflow.

3.2.2 Skimming

Not all information from the sample files is relevant for the analysis, therefore some preprocessing is done before the classification step, also called skimming. Some prefiltering is applied to only use data from good CMS runs and luminosity blocks for the analysis. The trigger information, as well as the PF objects for the selected events, is read out and stored in ROOT files[2]. The skimming drastically reduces the file size of each sample by extracting only the relevant information, which consequently shortens the computing time for the following MUSiC analysis steps. Note that some simulation samples feature process overlaps with others, which have to be cleared by specific filters, removing the events from overlapping physics processes from one dataset. The skimming step is computationally intensive, therefore it is run in parallel on the worldwide CMS computing grid using the CMS Remote Analysis Builder (CRAB).

3.2.3 Classification

3.2.3.1 Trigger readout

During the skimming process, the different HLT trigger seeds are read out. During classification, the events satisfying certain trigger requirements are then selected to include them in the further analysis. Past MUSiC analyses in most cases required at least one lepton (e, μ) or one photon. Therefore a variety of single and double muon, electron, and photon triggers was used (see e.g. [10, 32]). This thesis introduces jet triggers to MUSiC, which will be discussed in sec. 4.4 in details.

3.2.3.2 Object selection

For each accepted event, the object content has to be determined. Since the MUSiC analysis relies fully on the quality of MC simulation of SM processes, object selection requirements are applied, to ensure a low object misidentification and fake rate, according to the CMS recommendation. These requirements are commonly referred to as Tight Identification (Tight ID). Also, kinematic cuts are applied for the different objects, which are listed in tab. 3.1.

Object	p_T [GeV]	Pseudorapidity
Muon	> 25	$\|\eta\| < 2.4$
Electron	> 25	$0 < \|\eta\| < 1.442$ or $1.566 < \|\eta\| < 2.5$
Photon	> 25	$\|\eta\| < 1.442$
Jet, bJet	> 50	$\|\eta\| < 2.4$
Missing transverse momentum	> 100	−

Tab. 3.1: Object selection requirements in MUSiC, taken from the most recent iteration of the analysis [32].

Note that only photons from the low-eta barrel region are used, to reduce the number of fake objects. Additionally, the barrel-endcap transition region is excluded for electrons. Jets are reconstructed with the anti-k_T algorithm (see sec. 2.2.5.2) with a distance parameter of $R = 0.4$ (`AK4PF` jet). B-tagging of jets is also performed in the analysis, using the tight working point of the DeepJet algorithm [60]. All recommended object corrections in CMS are applied in MUSiC, these are introduced in the following

[2] The file format was changed recently, legacy MUSiC analyses stored skimmed data in pxlio files.

section. Note that tau leptons (τ) are currently not considered in the MUSiC analysis. In MUSiC, Missing Transverse Energy (MET) is essentially treated as a separate physics object if the missing transverse momentum exceeds the selection threshold.

3.2.3.3 Object corrections

The selected objects are corrected, following the recommended procedure by CMS. Scale factors are used for most of the reconstructed objects to account for differences in the reconstruction efficiency for objects in data and MC. Additionally, there exist multiple corrections for object momenta, which are applied following official recommendations, and were obtained from official calibration measurements. These corrections usually depend on the kinematic variables of the object, e.g. its energy or angular variables. Note that because of the large number of protons in one bunch, multiple collisions may happen during the crossing of the two proton bunches. This effect is called Pileup (PU) and is of special importance at high luminosity accelerators [4]. In fact, during the 2018 data taking with the CMS detector, more than 30 proton-proton collisions per bunch crossing were measured on average [61]. In CMS, pileup correction sets are used to perform the event corrections and match the pileup distributions for data and MC. This thesis uses the PU corrections recommended by CMS [62].
Since this thesis focuses on jets, the jet correction workflow recommended by CMS should be discussed briefly. The two most important corrections include Jet Energy Correction (JEC) and Jet Energy Resolution (JER). The JEC correction includes multiple steps. The first correction is due to PU, which was already introduced, and usually leads to increased jet momenta since particles from other vertices within the bunch collision add their energy to the reconstructed jets. Additionally, the simulated detector response used for simulated samples shows deviations from ideal detector behavior and therefore has to be corrected. With a correction set dependent on the pseudorapidity and the transverse momentum of the jet[3], a correction factor is applied on the reconstructed jet four vector [63]. The second correction is JER. It accounts for the fact that the jet energy resolution in data is different than in MC. Therefore, smearing or correction of the four-vector is performed for MC. This analysis follows the official recommendation and uses the ”hybrid” method for JER, which is described in the CMS prescription [64]. Note that when correcting the jet energy, the obtained missing transverse momentum has to be corrected accordingly to the jet energy corrections to ensure transverse momentum balancing.
Since protons are composite particles made of quarks and gluons, the simulation of proton-proton collisions, as they are performed at the LHC, is challenging. First, it is unclear which of the particles in the proton participate in the primary collision. The particle content of the proton is described by Parton Distribution Functions (PDFs). PDFs are functions that describe the probability of finding a given quark with a given longitudinal momentum fraction in the hadron [23]. Usually, a set of PDFs for all possible quarks, antiquarks, and gluons is used to describe the full content of a hadron. PDFs can not be calculated perturbatively, but are usually extracted from data [23]. The recommended NNPDF 3.1 [65] PDF set is used in the analysis.

3.2.3.4 Object cleaning

Overlaps of reconstructed objects cannot be excluded. Therefore, the objects are cleared against each other when they have an angular distance ΔR (as defined in eq. 2.3) below a certain threshold. The distance thresholds as well as the order of clearing are presented in tab. 3.2. Note that the object cleaning is of special importance in jet reconstruction since the PF algorithm potentially reconstructs the same detector signature as a jet and as a different object.

[3] PU correction is also binned in energy density and jet area [63].

Reference object	Object to clean	Distance threshold
Muon	Electron	0.4
Electron, muon	Photon	0.4
Jet, bJet	Photon	0.5
Jet, bJet	Electron	0.5
Jet, bJet	Muon	0.5

Tab. 3.2: Object cleaning requirements in MUSiC in chronological order. The objects to clean are cleared against the reference objects when their distance ΔR deceeds the threshold.

3.2.3.5 MC weighting

Since the number of generated events for the MC simulation samples are generally independent of the expected event rates of a real measurement, the MC events have to be reweighted. According to eq. 2.1, the process cross section σ (returned by the MC generator and possibly corrected to higher order calculations with the k-factor) and the luminosity $\mathcal{L}$ determine the event rate. For the MC event weight to match with the expected counts in data and therefore enable a direct comparison between data and MC, reweighting is necessary. The weight related to cross section and luminosity w_{lumi} is obtained as [59]:

$$w_{\text{lumi}} = \frac{k \cdot \sigma \cdot \int \mathcal{L} \, \mathrm{d}t}{N_{\text{MC}}}. \tag{3.1}$$

As already mentioned in sec. 3.2.3.3, PU of multiple proton-proton collisions per bunch is very likely observed. Initially, the MC pileup distributions do not match with the distributions of data. An additional weight factor w_{PU} is used to introduce a correction for this [4]. Finally, MC generators possibly pass a generator weight w_{gen} per event (float number), which has to be taken into account. The main reason for introducing a generator weight in the calculation is the possibility of negative weights being passed by the MC generator. Because of this, N_{MC} in eq. 3.1 does not describe the integer number of generated MC events, but instead the sum of all generator weights w_{gen}, to achieve reasonable normalization for event weights. The total MC weight w_{MC} is then obtained by combining the three weight parameters [4]:

$$w_{\text{MC}} = w_{\text{lumi}} \cdot w_{\text{PU}} \cdot w_{\text{gen}}. \tag{3.2}$$

3.2.3.6 Event classes

The essential part of MUSiC is the model unspecific aspect of the analysis. This is realized by sorting the events to event classes, according to the physics object content instead of restricting the analysis to specific final states. MUSiC differentiates three different types of event classes, which will be briefly introduced here [59]:

- **Exclusive classes**: Exclusive event classes only include the events that have exactly matching physics objects with the ones required by the class. Therefore, each event automatically becomes a member of one exclusive class.

- **Inclusive classes**: Inclusive classes contain events with all physics objects required in the class name or a higher number of physics objects. Therefore one event can be a member of multiple inclusive classes. Inclusive classes are sometimes denoted by a +X suffix in their class name.

- **Jet-inclusive classes**: These classes behave like exclusive classes, but also allow higher jet multiplicities in the events. Jet-inclusive classes can somewhat remedy the effect of increasing jet multiplicities because of possible gluon emissions (as described in sec. 1.3.2.2). Jet inclusive classes are sometimes denoted with a +Njets suffix.

Note that previous MUSiC analyses limit the jet multiplicity to five. Events with a higher jet count are only considered for inclusive and jet-inclusive classes [59]. Fig. 3.1 illustrates the filling of the event classes with an example event containing $\{1\mu,\ 2\text{jets}\}$.

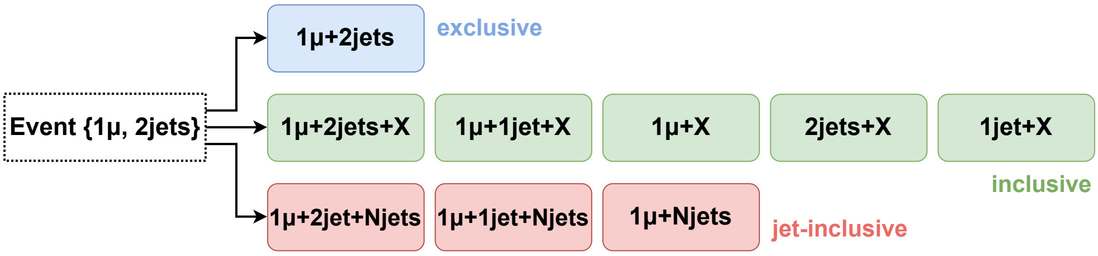


Fig. 3.1: Example of MUSiC classification for one event containing $\{1\mu,\ 2\text{jets}\}$.

3.2.4 Scan

3.2.4.1 Distributions of interest

MUSiC is designed to search for deviations in three[4] kinematic distributions per event class. For each class, the respective quantities of the distributions are calculated from the event [59]. These quantities include:

- **Sum of transverse momenta** S_T: This quantity is obtained by summing the transverse momenta of all physics objects in the class (indexed with i), hence:

$$S_\mathrm{T} = \textstyle\sum_i |\vec{p_{\mathrm{T},i}}| \,. \tag{3.3}$$

  If MET is included in the class, then $|\vec{p_{\mathrm{T,miss}}}|$ is also considered for the sum. The quantity of S_T represents the total energy of the interaction process. Many BSM models expect new particles with high masses, therefore signs of this could be visible in the high-energy tails of this distribution [59].

- **Invariant mass** m_inv or **transverse mass** m_T: The invariant mass is calculated from the energies E_i and momenta $\vec{p_i}$ of all physics objects in the class as [59]:

$$m_\mathrm{inv} = \sqrt{\left(\textstyle\sum_i E_i\right)^2 - \left(\textstyle\sum_i \vec{p_i}\right)^2}. \tag{3.4}$$

  For classes containing MET, instead of calculating the invariant mass, the transverse mass is calculated from transverse energy $E_{\mathrm{T},i}$ and transverse momentum $\vec{p_{\mathrm{T},i}}$ of the objects, according to the following formula [59]:

$$m_\mathrm{T} = \sqrt{\left(\textstyle\sum_i E_{\mathrm{T},i}\right)^2 - \left(\textstyle\sum_i \vec{p_{\mathrm{T},i}}\right)^2}. \tag{3.5}$$

  This distinction is made, because only the transverse component of the missing energy (MET) is known [66]. The mass quantities are of interest when allegedly massive BSM particles should be observed in resonances, decaying to the final state objects in the event class. They are only calculated when there are at least two objects in the class [59].

- **Missing transverse momentum** $p_\mathrm{T,miss}$: Event classes that include MET introduce the missing transverse momentum as a third kinematic variable. It is calculated as the negative sum of

[4] For classes without MET, only two distributions are analyzed, since in this case $p_\mathrm{T,miss}$ is not present. Classes with only one physics object also do not consider the mass variable, further reducing the number of distributions by one.

the transverse momenta of all objects identified by the PF algorithm indexed by j:

$$p_{\mathrm{T,miss}} = |\vec{p_{\mathrm{T,miss}}}| = \left| -\sum_j \vec{p_{\mathrm{T},j}} \right| . \tag{3.6}$$

Note that corrections of the reconstructed objects have to be taken into account and MET has to be corrected accordingly. The MET distribution is of special interest when it comes to identifying non-interacting particles, like neutrinos or hypothetical "dark" BSM particles [59].

It should be emphasized that for an event in an inclusive or jet-inclusive class, only the physics objects that are explicitly named in the name of the inclusive class are considered when calculating these quantities. The physics objects are sorted after their p_T, starting with the highest momentum. This means that for example an event containing $\{2\mu\}$ would only contribute as $S_\mathrm{T} = p_{\mathrm{T},\mu 1}$ to the $1\mu + \mathrm{X}$ class, but as $S_\mathrm{T} = p_{\mathrm{T},\mu 1} + p_{\mathrm{T},\mu 2}$ to the 2μ class. However, there is an exception for MET which is always calculated using kinematic information of all objects, since it is essentially treated like a separate object.

All distributions are analyzed in the form of histograms, whose bin width is not initially set. While small bin sizes would theoretically enable higher resolutions and thus higher sensitivity for narrow BSM signals, wider bins would decrease the computation time. Additionally, very narrow bins might emphasize random fluctuations of data or MC, possibly hiding the main deviations that the MUSiC analysis should reveal. Therefore it was decided to automatically calculate the bin width in MUSiC according to the typical overall detector resolution of the objects considered for the corresponding event classes as integer multiples of 10 GeV [59].

3.2.4.2 p-value and Region of Interest scan

To quantify the significance of the deviation between data and MC, the analysis calculates p-values for each histogram, which can be regarded as a measure of this significance. In MUSiC, a hybrid Bayesian-frequentist approach is used for this [59, 67]. To calculate a p-value, the event count[5] N_MC with its total systematic uncertainty σ_MC as well as the measured data count N_data is needed. First, assume that the count from the SM expectation N_SM would be known. Then, statistical fluctuations between data and the SM expectation are modeled with a Poisson distribution around the SM event count [4, 10]. Since the MC underlies systematic uncertainties, it would be incorrect to assume the MC count N_MC as the SM expectation N_SM directly. Instead, the systematic uncertainty on the MC simulation is taken into account with a truncated Gaussian distribution [4]. Combining these assumptions, the p-value can be calculated as [59]:

$$p_\mathrm{data} = \begin{cases} \displaystyle\sum_{i=N_\mathrm{data}}^{\infty} \alpha \cdot \int_0^\infty \mathrm{d}x \, \exp\left(-\frac{(x-N_\mathrm{MC})^2}{2\cdot\sigma_\mathrm{MC}}\right) \cdot \frac{x^i \cdot e^{-x}}{i!}, & N_\mathrm{data} \geq N_\mathrm{MC} \\ \displaystyle\sum_{i=0}^{N_\mathrm{data}} \alpha \cdot \int_0^\infty \mathrm{d}x \, \exp\left(-\frac{(x-N_\mathrm{MC})^2}{2\cdot\sigma_\mathrm{MC}}\right) \cdot \frac{x^i \cdot e^{-x}}{i!}, & N_\mathrm{data} < N_\mathrm{MC} \end{cases} , \tag{3.7}$$

where α is a normalization factor. As can be seen in eq. 3.7, the Poisson distributions are summed up from 0 to N_data or from N_data to ∞. This is done for the p-value to reflect the significance of a deviation at least as large as between data and the SM expectation. The smaller the calculated p-value, the more significant the deviation between data and SM expectation.

In MUSiC exists a dedicated algorithm, the Region of Interest (RoI) finder, which scans the distributions for the region with the most significant deviation. All contiguous combinations of bins are considered as regions, therefore a p-value is calculated for all of these. The region with the smallest

[5] As already mentioned, the MC event count is in fact no integer count, but consists of the sum of the MC event weights w_MC. Their calculation has already been described in sec. 3.2.3.5.

p-value $p_{\text{data,min}}$ defines the selected RoI, as illustrated in fig. 3.2. Note that, to reduce the sensitivity to statistical fluctuations, the minimum width of the scanned regions is set to three for the S_{T} and $p_{\text{T,miss}}$ distributions. Since mass resonances of hypothetical new particles might be relatively narrow, the minimum bin limit is set to one for the mass distributions. Also, there exist some requirements for quality control of the regions, which should not be explained here in detail [59].

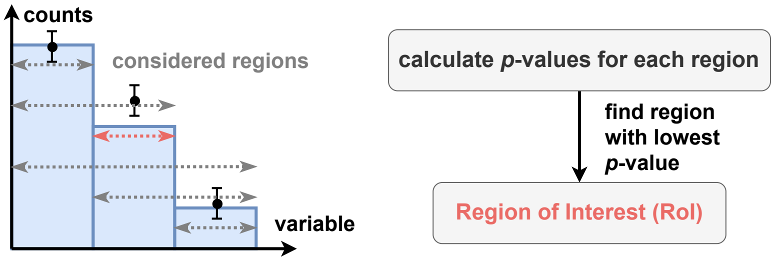

counts
considered regions
variable
calculate *p*-values for each region
find region with lowest *p*-value
Region of Interest (RoI)

Fig. 3.2: Illustration of the MUSiC RoI scan, in this example the minimum width of the regions is one bin. The illustration is based on [59, fig. 2].

3.2.4.3 Look-Elsewhere Effect and global comparison

The p-value calculated for each region only quantifies the significance of the local deviation in the respective region. However, obtaining a global measure of the deviation for each distribution is desired, since only then the deviations can be compared between different distributions. The $\tilde{p}$-value is introduced, which describes the probability to observe a deviation at least as large as present in any of the considered regions throughout the distribution, and is, therefore, this desired global measure of deviation. When transitioning from the local significance p to the global significance $\tilde{p}$, the Look-Elsewhere Effect (LEE) has to be taken into account [59]. This effect occurs when different regions in a distribution are considered for the deviation [4]. For MUSiC, the calculation of the $\tilde{p}$-value is not performed analytically but using pseudo experiments. In each pseudo experiment, the SM expectation is varied in a random manner, accounting for the expectation and the associated uncertainties of the simulation. More precisely, the bin counts for the MC expectation are varied with diced shifts that represent the different systematic uncertainties. Formally, the mean count N_n of bin n is shifted according to the following scheme for each pseudo experiment [59]:

$$N_{n,\text{shifted}} = N_n + \textstyle\sum_i \kappa_i \cdot \Delta_{i,n}, \tag{3.8}$$

where κ_i is a random number following a standard normal distribution $\mathcal{N}(0,1)$ and $\Delta_{i,n}$ is the width of the symmetrized 68 % confidence interval for the systematic uncertainty i and the respective bin n. Additionally, the bin counts $N_{n,\text{shifted}}$ is smeared with a Poisson distribution to consider the statistical uncertainty. Of course, this shifting procedure can only be applied to bins that feature a nonzero MC count and uncertainty [59].

Up to 10.000 pseudo experiments are performed, this number is chosen as a tradeoff between computing time and sensitivity [10]. The RoI scan is performed for each of these experiment results and the smallest p-value p_{min} is stored. With this, the $\tilde{p}$-value is then obtained as the fraction of pseudo experiments that lead to a more significant deviation than the smallest data p-value $p_{\text{data,min}}$ [59]:

$$\tilde{p} = \frac{N_{\text{pseudo}}(p_{\text{min}} < p_{\text{data,min}})}{N_{\text{pseudo}}}, \tag{3.9}$$

where N_{pseudo} is the number of pseudo experiments. A lower bound for the $\tilde{p}$-value is set by the number of pseudo experiments as $\tilde{p} \geq {}^{1}/_{N_{\text{pseudo}}}$, since the probability for a deviation can never be exactly zero [10].

To obtain an overview of the RoI scan and simplify the search for deviations, MUSiC plots the number of classes of one type and for one distribution against $\tilde{p}$ bins in a logarithmic plot. In these plots, the median $\tilde{p}$ of the SM expectation, as well as the uncertainty intervals, obtained from pseudo experiments (SM versus shifted SM) are plotted together with the $\tilde{p}$ values from the data-SM comparison. With this plotting scheme, deviations as well as their significance occurring in a large number of classes are visible together in one plot. An example of the result of a MUSiC analysis is presented in fig. 3.3. It is taken from the 2021 MUSiC paper [59], which presented a full MUSiC analysis with 2016 data.

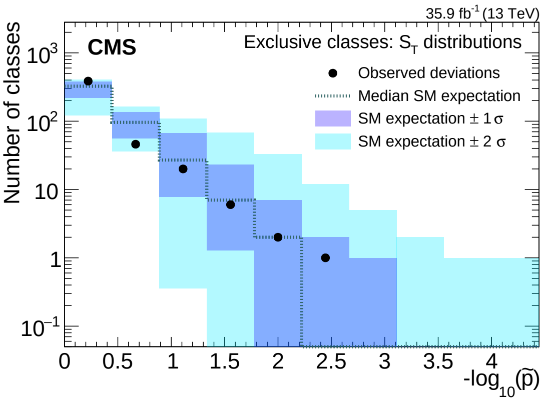


Fig. 3.3: Example result of a MUSiC scan of S_{T} distributions of exclusive classes, taken from the 2021 MUSiC paper [59, fig. 11]. The black dotted line shows the SM median expectation with the uncertainty intervals in blue. The black dots represent the $\tilde{p}$ values for data, which seems to be mostly compatible with the simulated SM prediction within the uncertainty bands [59].

3.3 Exploring the extension to jet-only final states

Apart from continuing to improve the existing MUSiC framework and pursuing the analysis of the full Run 2 and Run 3 data, there is a desire to extend the analysis to a broader range of final states. Allowing more final states would increase the number of analyzed classes and therefore the sensitivity of MUSiC to potential BSM signals that are only perceptible in classes currently excluded from the analysis. However, such extensions of the analyzed final states are only viable if MC can reasonably model all relevant SM processes for these classes. Therefore, studies have to be performed on the behavior of the possible class extensions, and possible corrections for the distributions have to be determined before these classes can be included in future MUSiC analyses.

One of the desired extensions to MUSiC are jet-only final states and event classes. Currently, MUSiC requires at least one lepton (e, μ) or photon, as already described, and only includes jets as complementary objects which may or may not be included in an event triggered with leptons or photons. If jet-only classes should be included, this trigger strategy has to be changed by introducing a jet trigger. Additionally, the quality of the MC simulation for these classes is unexplored, therefore studies have to be performed on this topic and potential corrections have to be introduced to improve the simulation quality.

This thesis aims to contribute to the possible extension of MUSiC with jet-only final states by conducting the first studies on this topic since 2015[6]. Initially, studies on including different jet triggers in the analysis are performed (sec. 4). Then, object selection and object merging strategies from other dedicated CMS analyses with jets are investigated and adapted for the potential use in MUSiC (sec. 5). Finally, the remaining disagreements between data and simulation are discussed and addressed (sec. 6). Because of the large scope of the topic and the limited working time for a bachelor thesis, the inclusion of jet-only event classes to MUSiC could not be completed, however, steps were taken in this direction and potential issues with these event classes were uncovered.

To restrict the scope of the thesis, the decision was made in the beginning to only use the recorded dataset of CMS from 2018. This decision was driven by the limited working time for the thesis. Therefore, the analysis focuses on jet-only final states, excluding most other physics objects in the larger part, this will be concretized in the next sections. A full MUSiC analysis, including a complete classification and scan, is also beyond the scope of this thesis.

[6] In 2015, an analysis with jet triggers in MUSiC was performed by A. Albert (RWTH Aachen University) [68], which, although a different dataset was used, shares similarities with this thesis.

4 Jet-triggered events

4.1 Previous studies on jet-only final states

4.1.1 Dedicated CMS searches

There exist numerous dedicated analyses in CMS which study jet-only final states to search for hints of BSM physics. Most prominently, there exist dijet resonance analyses, like [53, 69], which search for mass resonances of hypothetical particles in dijet final states. The results of such analyses are lower mass limits for the hypothetical particles in common BSM theories. Because of space limitations, these BSM theories will not be explained here in detail. Still, some limits should be briefly listed here since they will be used in the discussion further below. A recent Run 2 dijet resonance analysis published in 2020 [53] puts the minimum mass for string resonances, scalar diquarks, axigluons and colorons, and excited quarks to at least 6.3 TeV, and for color-octet scalars, W' and Z' bosons to at least 2.9 GeV at 95 % confidence level. Graviton lower limits are set at 2.6 TeV and dark matter mediators at 2.8 TeV for narrow resonances. Wide resonances have a lower mass limit of 4.8 TeV. Therefore, at least for common BSM theories, no significant contribution would be expected for low-mass dijet events, which could be defined as $m_{\mathrm{inv}} < 2.6\,\mathrm{TeV}$ from these results. Besides dijet analyses, there exist dedicated searches for some other jet-only final states, e.g. trijet events [54].

4.1.2 Previous studies in MUSiC

As already mentioned, a study on including jet triggers in MUSiC was already performed in 2015 by A. Albert[1] in a master thesis [68]. In the thesis, the 2012 dataset recorded by CMS at $\sqrt{s} = 8\,\mathrm{TeV}$ was analyzed. A jet trigger for the leading jet $p_{\mathrm{T}} > 320\,\mathrm{GeV}$[2] was used with PYTHIA 6 [70] QCD samples and the rest of the full MUSiC MC sample set at the time. To ensure a high trigger efficiency, a transverse momentum cut was applied to the leading jet of $p_{\mathrm{T}} > 400\,\mathrm{GeV}$. Mostly the same MUSiC selection criteria as in tab. 3.1 were used[3]. However, since a different dataset at lower center-of-mass energy was used, a comparison of the results from the master thesis with the results from this analysis, which will be presented below, is not directly possible.

Resulting plots for the jet-only classes from the master thesis are presented in the appendix in sec. B.1. These classes are found to be dominated by QCD MC samples. Data and MC mostly seem to agree within the uncertainties. Note that the uncertainties are in the order of 50 % because a 50 % cross section uncertainty was applied on the dominating LO QCD samples, according to the MUSiC prescription (see discussion of systematics below in sec. 4.6). A systematic decrease of the data/MC for increasing energies in the energy-like distribution is observed. Locally different behavior of the data/MC ratio in the low-energy region below the main trigger efficiency turn-on peak can be seen. After the results of this analysis are presented in sec. 4.10, it will become apparent, that this analysis shares some similarities with the master thesis.

[1] RWTH Aachen University

[2] The HLT trigger path for the used trigger is `HLT_PFJet320`.

[3] Muons were only selected if $|\eta| < 2.1$ and MET was selected at a lower threshold of $p_{\mathrm{T,miss}} > 50\,\mathrm{GeV}$. For the full sample list, the selection requirements, and applied corrections, see the descriptions in the master thesis [68].

4.2 Analysis framework

This thesis explores event classes including only jets, commonly referred to as jet-only classes. For this purpose, a framework different from the regular MUSiC framework was used, allowing the analysis of an increased number of distributions and allowing a more flexible trigger and object selection. This framework is referred to as the ”validation” framework since it enables to validate agreement of simulation and data for a large variety of distributions and will also be used for this purpose in future MUSiC analysis. Therefore, apart from the results of the analysis performed in this thesis, the extension of the MUSiC validation framework, namely the development of a new, universal plotting tool, can be regarded as a contribution to the MUSiC analysis. Fig. 4.1 compares the workflow of a full MUSiC analysis with the workflow in this thesis. Only the skimmer is shared, then this thesis continues the analysis in modified and extended versions of the ”validation” framework.

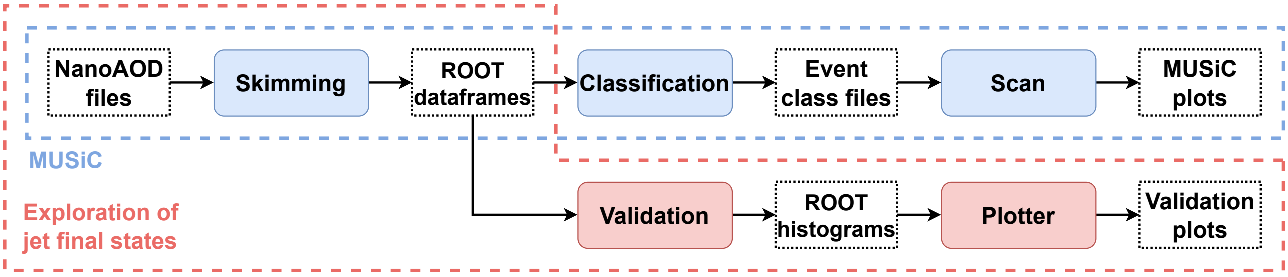


Fig. 4.1: Comparison of MUSiC workflow (blue) with the analysis in this thesis (red). Both analyses only share the skimming part, after that, a different framework was used. The illustration scheme was adapted from [32, fig. 3.1].

4.3 Dataset

As already stated, this thesis uses the 2018 dataset recorded by CMS, which is part of LHC Run 2 at $\sqrt{s} = 13\,\mathrm{TeV}$ and has an integrated luminosity of $\int \mathrm{d}t\ \mathcal{L} = 59.8\,\mathrm{fb}^{-1}$. In particular, the `Jet_HT` dataset[4] is used. The MC samples are generated by different event generators, including PYTHIA 8 [71], MADGRAPH 5 aMC@NLO 2 [72], POWHEG V2 [73–84] and SHERPA 2 [85]. All processes predicted by the SM and accessible with MUSiC should be covered, therefore a large number of samples is necessary. The sample selection is similar (apart from the year) to the most recent MUSiC analysis of Run 2 of 2016 data, which was published in a 2021 CMS paper [59]. The simulated samples cover the following process groups: Drell-Yan, Gamma (γ), Higgs (H^0), Multi-Boson, QCD, Top (t), TTbar ($t\bar{t}$), and W. It should be noted that all MC samples used in this analysis are generated by the CMS collaboration and acquired from DAS and no private MC samples are used. All samples used are from the UltraLegacy reconstruction campaign in CMS.
Since the jet-only classes are mostly dominated by events from QCD multijet samples, two different datasets for these processes are considered for future use. One QCD dataset was generated with PYTHIA 8 and is binned in the transverse momentum p_T of the leading jet[5]. The p_T bins range from $15\,\mathrm{GeV}$ to infinity without overlap. The other dataset in question was generated with MADGRAPH 5 aMC@NLO 2, is binned in transverse momenta sum of the jets $\sum_\mathrm{jets} p_\mathrm{T} = H_\mathrm{T}$ (the H stands for hadronic)[6]. Note that therefore, the transverse momenta sum of the event S_T is not the same as H_T, which only refers to the jets in the event. The H_T bins range from $300\,\mathrm{GeV}$ to infinity without overlap.

[4] DAS name: `/JetHT/Run2018*-UL2018_MiniAODv2_NanoAODv9-v2/NANOAOD` for runs `A-D`.

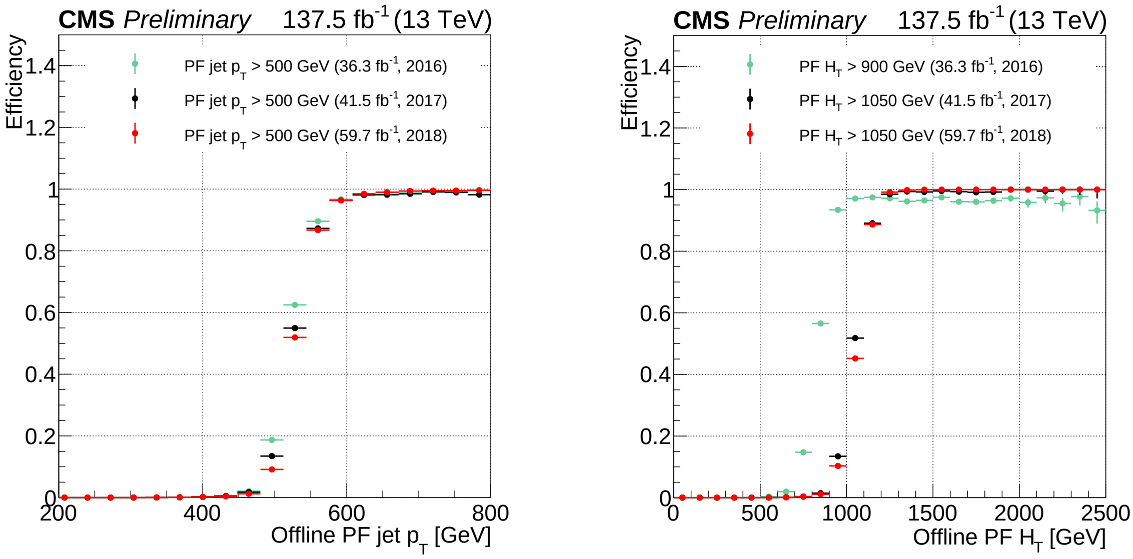
[5] DAS name: `/QCD_Pt_*to*_TuneCP5_13TeV_pythia8/RunIISummer20UL18NanoAODv9-106X_upgrade2018_realistic_v16_L1v1-v1/NANOAODSIM`.

[6] DAS name: `/QCD_HT*to*_TuneCP5_PSWeights_13TeV-madgraph-pythia8/RunIISummer20UL18NanoAODv9-106X_upgrade2018_realistic_v16_L1v1-v1/NANOAODSIM`.

Both QCD multijet datasets are simulated in Leading Order (LO).
The recommended cross sections by CMS specified for the respective samples are used in this analysis. k-factors to correct the cross sections to higher-order calculations are applied for some processes. A full list of all MC samples can be found in the appendix in tab. A.1. In this table, the cross section, generator order, and k-factor are also listed. All samples are available in the NanoAOD file format, and as for a full MUSiC analysis, the important information has to be extracted from the files for further analysis. As already shown in fig. 4.1, the analysis in this thesis shares the skimming part with MUSiC, which was explained in sec. 3.2.2. Therefore, the skimming process is not explained here anymore.

4.4 Trigger strategy

To analyze jet-only final states, the trigger strategy from previous MUSiC analyses has to be changed. Events with jets should be selected by the trigger, thus a jet trigger has to be introduced to the analysis. As for the QCD samples, there exist both p_{T} and H_{T} triggers. In the beginning, it is unclear which samples and triggers should be chosen, therefore in this chapter studies are conducted with different sample-trigger combinations. The two HLT trigger paths selected for this analysis are the `HLT_PFJet 500` trigger, which fires when the event has at least one jet with $p_{\mathrm{T}} > 500\,\mathrm{GeV}$, and the `HLT_PFHT1050` trigger, which fires when the event contains hadrons (components of jets) with $H_{\mathrm{T}} > 1050\,\mathrm{GeV}$. This selection is based on the fact that these triggers are the ones with the lowest firing thresholds from the unprescaled[7] triggers in the respective category [86, 87]. Additionally, these two trigger paths are commonly used by dedicated CMS analyses, including dijet (e.g. [53]) and trijet (e.g. [54]) and Higgs (e.g. [88]) analyses.



(a) `HLT_PFJet500` trigger [86, p. 4] **(b)** `HLT_PFHT1050` trigger [86, p. 5]

Fig. 4.2: Measured trigger efficiencies by CMS in 2018 for the two selected triggers (red data points).

Generally, trigger efficiencies have to be accounted for in any analysis. The strategy to mitigate event loss because of low trigger efficiencies is to introduce cuts on the respective kinematic variables that ensure a high trigger efficiency. The trigger efficiency for 2018 was measured by the CMS collaboration [86], therefore from these results the thresholds can be set. Fig. 4.2 shows the results of the trigger

[7] Trigger prescaling effectively means that many events that satisfy the trigger requirements are rejected. Usually, this is done to reduce the frequency at which data has to be analyzed and stored in the data acquisition system. However, for this analysis unprescaled triggers are preferable.

efficiency measurements for both triggers.
A trigger efficiency value of 0.95 is regarded as high enough to neglect the correction of the efficiency. Therefore, the thresholds are selected as $p_{\mathrm{T,thres}} = 600\,\mathrm{GeV}$ for the leading jet in case of the `HLT_PFJet500` trigger and as $H_{\mathrm{T,thres}} = 1400\,\mathrm{GeV}$ for the jets in the event in case of the `HLT_PFHT1050` trigger. These cuts are applied not until after the object selection. By applying these large thresholds, it is ensured that the number of misidentified objects is low. The triggers are regarded as fully efficient above the selected thresholds and no trigger efficiency scale factors are applied.

4.5 Object selection

4.5.1 Selection criteria and corrections

The object selection step is similar to MUSiC. Of course, this has the reason that to study potential additions to MUSiC, similar selection criteria should be used to be able to make meaningful observations. Jet reconstruction is again performed with the anti-k_{T} algorithm (see sec. 2.2.5.2) with the distance parameter $R = 0.4$ (`AK4PF` jet). All objects are selected from the accepted events according to the requirements that were already listed in tab. 3.1 and with the Tight ID requirements. The object corrections are performed according to the CMS recommendations, as for MUSiC. Note that jet energies are also corrected according to the CMS standard. All corrections were briefly introduced in sec. 3.2.3.3. Object cleaning is also performed similarly to MUSiC, shown in tab. 3.2. It should be pointed out that, unlike in MUSiC, b-tagging is not performed in this analysis since it adds additional complexity to the jet-only classes, and the primary goal of this analysis is to explore the general behavior of these jet-only classes. Still, the future goal would be to introduce b-tagging into jet-only classes, at the latest if these classes should be introduced to the MUSiC analysis. As stated in the last section, the selected trigger efficiency thresholds are applied after the object selection and correction.

4.5.2 Lepton, photon and MET veto

As stated, this thesis aims to explore jet-only event classes. This can be motivated by the fact that all events containing at least one lepton (e, μ) would already be included in the full MUSiC analysis. To enforce jet-only classes, a muon, electron, and photon veto is applied after reconstructing the objects. It should be noted that tau leptons (τ) were not yet included in the MUSiC analysis and therefore also not in this study on jet-only final states. Therefore taus are not reconstructed or selected and consequently also not vetoed.
Events with MET (according to the selection criteria) are also vetoed since the event classes should only contain jets and in MUSiC, MET is considered as a separate physics object.

4.6 Systematic uncertainties

Different sources induce systematic uncertainty into the analysis. These sources have to be properly treated. The treatment of systematic uncertainties in this analysis follows the prescription for a full MUSiC analysis, the latest being [59] published in 2021. Tab. 4.1 lists all considered systematic uncertainties in the analysis. The considered uncertainties will also be briefly introduced in the following paragraphs.
Statistical uncertainties are considered. The statistical error is calculated as the square root of the sum of squared weights for each sample in the respective bin. The systematic uncertainty for the integrated luminosity in 2018 was determined as 2.5 % in the official CMS measurement [89]. The performed event weight pileup corrections (sec. 3.2.3.5) have associated uncertainties, which are considered according to the official recommendation. Since MUSiC fully relies on MC modeling and its corresponding

cross sections, it was decided to introduce cross section uncertainties. Following the past MUSiC analysis, for LO samples, a generous 50 % uncertainty is applied. Higher-order cross sections are assumed precise, which is generally untrue. However, since the distributions of the jet-only classes are dominated by LO QCD events (see next sections and chapters), the uncertainties of other, higher-order samples are neglectable. PDF uncertainties are considered, the uncertainties are applied following the official PDF4LHC recommendations [90]. In combination with the PDF uncertainties, systematic uncertainties on the measured value of the strong interaction constant $\alpha_\mathrm{s} = 0.118 \pm 0.0015$ are applied. Uncertainties on the JEC and JER jet corrections (sec. 3.2.3.3) are considered, following the official recommendations [63, 64]. Finally, there is an uncertainty associated with a prefiring issue of the Level 1 trigger in CMS, which is corrected with a separate event weight factor, as for the last full MUSiC analysis [59]. This issue is related to the degradation of the ECAL crystals, which leads to timing delays. Therefore, it is possible that the previous bunch crossing is recorded instead of the current one, resulting in a decreased efficiency to record interesting events [91].

Systematic source	Prescription	Typical relative uncertainty
Statistical uncertainty	Square root of the sum of squared weights for each sample	$< 1\,\%$[a]
Integrated luminosity	Official CMS recommendation [89]	2.5 %
Pileup correction	Official CMS recommendation	$\approx 2\,\%$[b]
Cross sections	50 % on all events of MC samples produced at LO, as for past MUSiC analyses	$\approx 45 - 50\,\%$[c]
Parton Distribution Functions and α_S uncertainty	Official PDF4LHC recommendation [90] and $\alpha_\mathrm{s} = 0.118 \pm 0.0015$ added in quadrature	$\approx 0.4 - 1\,\%$
Jet Energy Correction	Official JEC recommendation [63]	$\approx 5 - 15\%$[d]
Jet Energy Resolution	Official JER recommendation [64]	$\approx 1\,\%$[e]
Prefiring correction	Official CMS recommendation	$< 0.1\,\%$
Combined		$\approx 45 - 50\,\%$[f]

Tab. 4.1: List of systematic sources and their related uncertainties. The typical relative uncertainties refer to the typical relative error for the jet-only event class distributions with all samples combined, and not the values per sample, since errors can vary much between different samples.

[a] Statistical uncertainty is very small for typical event classes with event counts in the order of 10^6. However, for low-statistics classes, the statistical error exceeds this typical value and becomes significant.

[b] Most classes have value around 2 %, however up to ≈ 20 % are observed for very high jet multiplicity classes that have very low statistics.

[c] Depends on the fraction of LO samples in the class. For jet-only classes, this uncertainty is about 50 % because of the large contributions of the QCD samples, which are produced at LO. Very high jet multiplicities and the 1jet class have lower QCD contributions and therefore lower cross section uncertainties.

[d] This uncertainty increases with the jet multiplicity and exceeds the typical value for a large number of jets.

[e] For large jet multiplicities, this value is larger.

[f] Since most of the jet-only classes are dominated by QCD samples (generated at LO), the cross section uncertainty is dominant and therefore the combined relative uncertainty is very high, around 50 %. For classes with lower QCD contributions, this value is smaller. If other systematics have significant contributions, the value can even exceed 50 %.

The uncertainties usually consist of two shifts, "up" and "down", with possibly asymmetric uncertainty intervals. Before the systematics are propagated, both shifts are symmetrized, analogous to the MUSiC analysis. When combining the samples, they are treated fully correlated within one systematic source, except for the cross section and statistical uncertainty[8]. When combining the uncertainties, different

[8] Following the MUSiC instructions, the cross section uncertainty is treated fully correlated between samples within one process group and of the same order, and uncorrelated between orders and groups. The statistical uncertainty of all samples is treated uncorrelated [59].

systematic sources are treated as uncorrelated, following the MUSiC prescription. Whenever bins are combined, the errors of different bins for one sample are treated fully correlated, as in the full analysis [59].

A full MUSiC analysis considers even more systematic uncertainties than this analysis, such as the uncertainties on the lepton and photon corrections, reconstruction efficiencies and misidentification uncertainties [59]. However, these sources are excluded from this study with jet-only classes for multiple reasons. Lepton and photon uncertainties are neglected to decrease the complexity of the analysis since these objects are vetoed. Still, they might have a small impact on the vetoing of events. MET is also vetoed, however since MET systematics are related to jet systematics, these systematics on MET are considered.

Reconstruction efficiencies and correct identification probability are generally very high for jets and therefore systematics on reconstruction and misidentification are neglected. The same can be said for the trigger efficiency correction and respective uncertainty, which are not accounted for since kinematic thresholds have been selected to ensure high efficiency.

4.7 Filling of jet-only classes

As shown in fig. 4.1, the studies conducted in this thesis use a different analysis framework than MUSiC after the skimming step. However, to analyze jet-only event classes, an analysis step similar to the classification had to be implemented in the "validation" framework used in this analysis. Event classes including only jets are considered, all other MUSiC objects (e, μ, γ, MET) are vetoed (see sec. 4.5.2). Only jet-inclusive and exclusive classes are considered here since the differentiation between jet-inclusive and inclusive classes is ambiguous when the object veto is applied. It should also be noted that the limit on the jet multiplicity for the classification (which is set to five in MUSiC, see sec. 3.2.3.6) is lifted in this analysis, so higher jet multiplicities can also be investigated in this first study. Apart from these changes, the classification of events happens analogously to MUSiC, which was already illustrated in fig. 3.1. The event class distributions are stored as ROOT histograms.

4.8 Analyzed distributions

As already stated, one of the reasons to use a different framework than MUSiC in this analysis is to include a larger variety of distributions. Besides the MUSiC distributions, namely S_T, m_inv or m_T and $p_\mathrm{T,miss}$ (MET) explained in sec. 3.2.4.1, the "validation" framework features the following additional distributions:

- Object multiplicities (N_jet, N_e, N_μ, N_γ): Used to validate the vetoing and classification and to observe possible dependence of the agreement of data and MC on jet multiplicity.
- Leading and subleading jet kinematic variables (p_T, η, ϕ): Used to explore the behavior of these distributions, to better understand the S_T distribution and to validate applied cuts and selection criteria.
- Differential angular distributions between leading and subleading jet (ΔR, $\Delta\eta$, $\Delta\phi$): Used to understand the layout of jets in the event and to observe possible dependence of the agreement of data and MC on kinematic variables.
- ϕ_MET angular distribution of $p_\mathrm{T,miss}$ (MET).

4.9 Plotting

A completely new plotting tool was developed for the analysis, which stacks all different MC samples for each class and sorts the samples into process groups, denoted with different colors in the plots. The categories are sorted by their contribution. Also, all systematic errors are included and combined by the plotting script according to the prescription in sec. 4.6. The MC errors are represented by gray dashed areas. All data samples are combined and plotted with statistical errors as circles with errorbars in the plot. Additionally, a data/MC plot is created which describes the agreement between data and MC per bin. To allow better readability, the bins in the data/MC plot are merged in regions with low statistics[9]. The axis limits are set automatically to match the first and last data point, therefore regions with only MC counts but no data are not plotted. For the data/MC plot, the y-limits can be fixed to allow better readability. If some bins show larger deviations, the data points are not shown within the y-limits, then an arrow indicates this overflow[10].

The plotting tool can also visualize the total event counts for different event classes simultaneously. In this case, the integrated MC counts are compared to the integrated data counts in a bar plot, with different classes on the x-axis. For this plot, a data/MC plot is also created.

4.10 Results

In the following, the results of the analysis described in the last sections are presented for different sample-trigger configurations. In this section only the results for the p_T-binned QCD samples and H_T trigger are shown. Two more configurations were considered, which are only presented in the appendix in sec. B.3 (H_T-binned QCD samples and H_T trigger) and sec. B.4 (p_T-binned QCD samples and p_T trigger). It was decided to only present the results for the other configurations in the appendix because they show mostly similar behaving distributions. For all configurations, plots for many more event classes as presented were produced and analyzed which can not be included in this thesis because of space limitations. Note again that just the jet-only classes of the form $n$jets[+Njets] ($n \geq 1$) were considered for this analysis.

4.10.1 p_T-binned QCD samples and H_T trigger

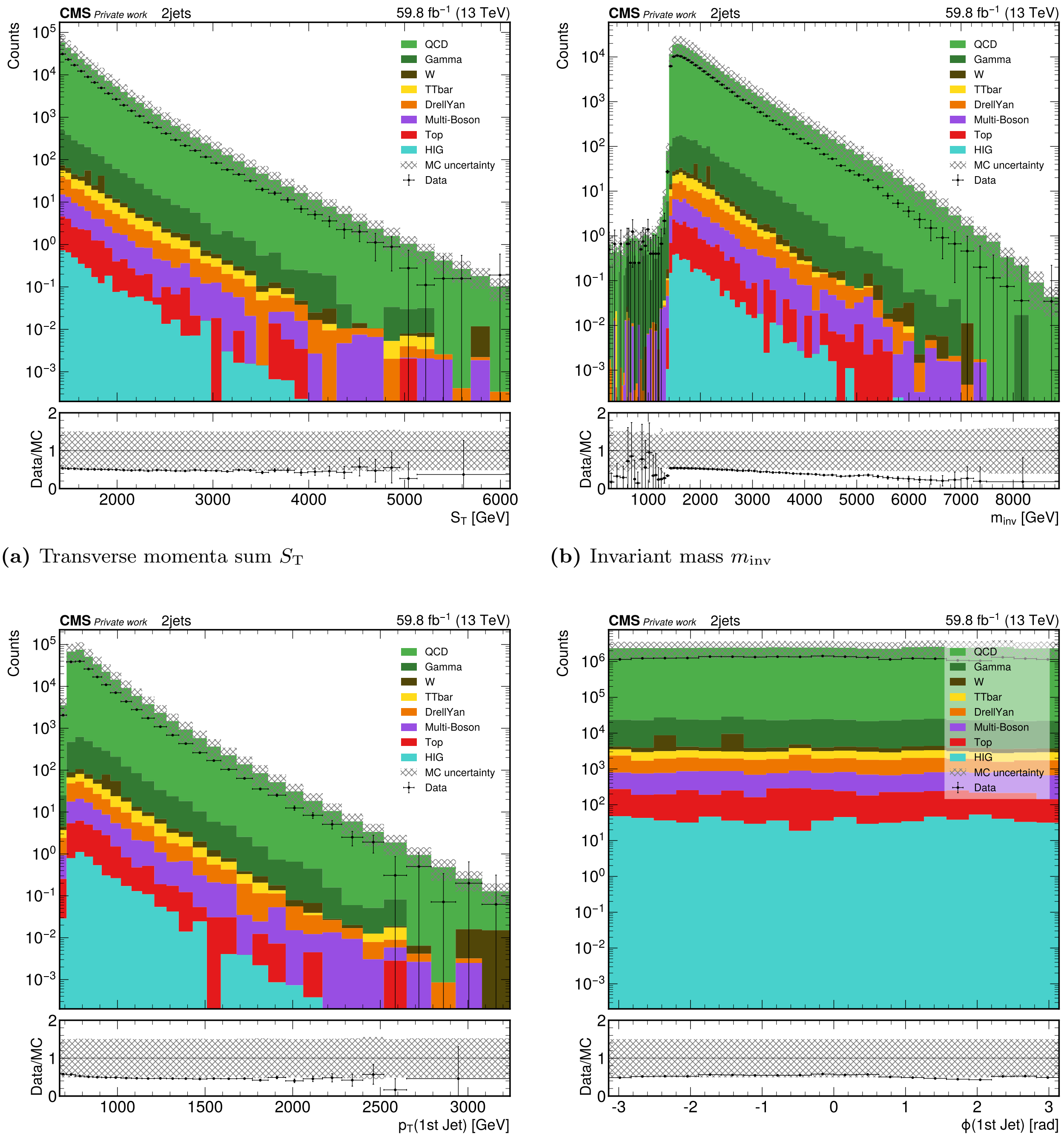
This sample-trigger-configuration implements the trigger efficiency cut of $H_\mathrm{T} > 1400\,\mathrm{GeV}$, since the H_T trigger is used (see sec. 4.4). The p_T-binned QCD samples are used in this configuration. There are 15 exclusive and 17 jet-inclusive classes found in data[11] and 17 exclusive and 17 jet-inclusive classes in MC.

First, the distributions from the 2jets exclusive class are reviewed. Fig. 4.3 shows four different distributions from this class. The first observation is that data is lower than MC by a factor of approximately ≈ 2. This deviation can be found in all distributions for the class, in energy-like distributions (e.g. S_T, m_inv and $p_\mathrm{T,leading}$) as well as in angular distributions (e.g. ϕ_leading). For the S_T plot (fig. 4.3a), the deviation is constant in the lower energy regime but for increasing energies the data/MC ratio decreases slightly. The applied trigger cut at $H_\mathrm{T} = 1400\,\mathrm{GeV}$ can be seen in the plot, the distribution only starts at the threshold value. The m_inv distribution (fig. 4.3b) shows a significantly larger energy dependence of the data/MC ratio. In the high energy regime (in this

[9] More precisely, bins are merged until their combined MC count is greater than one, the last merged bin has a nonzero MC count, and the data count in the combined region is greater than zero. This merging is similar, but not identical to the legacy MUSiC plotter.

[10] If the data errorbar exceeds the y-limit, this is marked with an arrowhead (black triangle). If the data point lies out of the plotting range, instead of an arrowhead, a full arrow with a bold tail is used as a marker. Therefore, if only the errorbar is in the frame, it can be read out in which direction the datapoint would be found.

[11] Since the 15jets and 16jets exclusive classes have no data point, but the 17jets exclusive class has, only the jet-inclusive classes with the jet multiplicities 15 and 16 are filled, which explains the deviation in the number of classes.



(a) Transverse momenta sum S_{T}

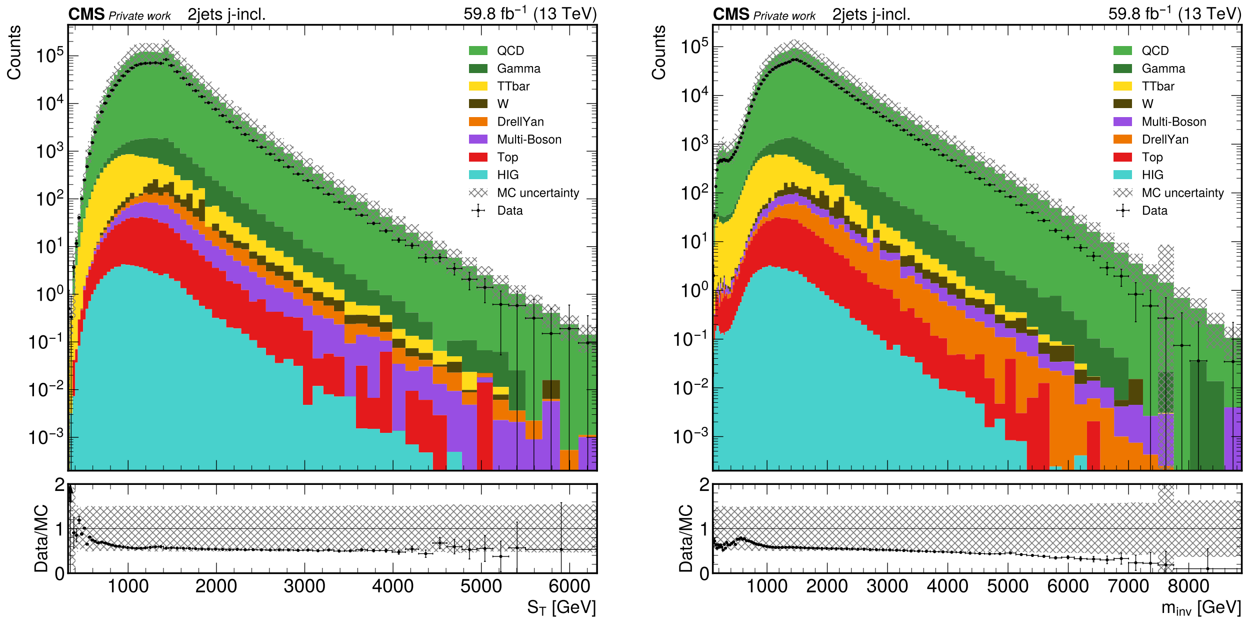
(b) Invariant mass m_{inv}

(c) Transverse momentum of leading jet $p_{\mathrm{T,leading}}$

(d) Phi of leading jet ϕ_{leading}

Fig. 4.3: Distributions for the 2jets exclusive class with p_{T}-binned QCD samples and H_{T} trigger.

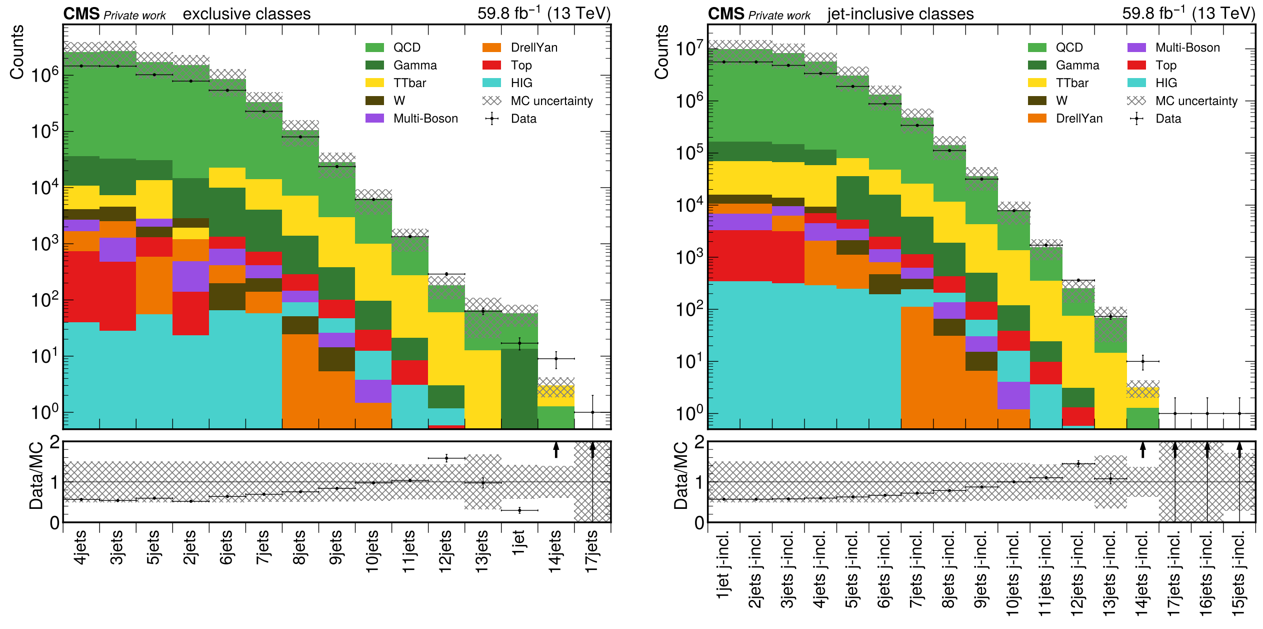
case $> 5\,\mathrm{TeV}$), the ratio decreases from the initial ≈ 0.5 to only ≈ 0.2. Note that in this high-energy region, the statistics are also rather low. The effect of the H_{T} cut can also be seen in the m_{inv} distribution, however only implicitly at around $1400\,\mathrm{GeV}$. Apparently, a few events with lower invariant mass are leaking through the cut. In the $p_{\mathrm{T,leading}}$ distribution (fig. 4.3c) an implicit cut can be seen at approximately $700\,\mathrm{GeV}$. This corresponds to half of the H_{T} trigger efficiency cut, which seems suggestive since the leading jet in the 2jets exclusive class has to carry at least the momentum of half of the momenta sum. The data/MC ratio also shows a slight decrease with increasing energy, as was already observed in other distributions. The noticed deviation between data and MC can also be seen in angular distributions, as already mentioned. An example of this can be seen in the polar angle of the leading jet ϕ_{leading} distribution (fig. 4.3d). A uniform distribution is expected and also roughly followed by data and MC, however an approximately constant offset between data and MC can be seen, with a data/MC ratio of ≈ 0.5. Additionally, it should be pointed out, that the QCD samples mainly dominate this jet-only event class, with about 95 % or more contribution to the total class yield.



(a) Transverse momenta sum S_{T}

(b) Invariant mass m_{inv}

Fig. 4.4: Distributions for the 2jets jet-inclusive class with p_{T}-binned QCD samples and H_{T} trigger.



(a) Exclusive classes

(b) Jet-inclusive classes

Fig. 4.5: Comparison of the integrated event counts for the classes with p_{T}-binned QCD samples and H_{T} trigger. In the event count plots, the categories are sorted by their contribution for every class separately, while the sorting was homogeneous for all bins in the kinematic distributions previously presented. The sorting of the classes on the x-axis is according to the integrated data counts.

Other classes should also be analyzed, next in this discussion is the 2jets jet-inclusive class, which is related to the already discussed exclusive class. As already stated, this class also includes events with higher jet multiplicities as specified in the class name. Fig. 4.4 shows two distributions from this class. The deviation between data and MC can be observed and is of the same magnitude as for the exclusive class discussed above. For the S_{T} distribution (fig. 4.4a), only a slight decrease of the data/MC ratio can be seen with increasing momentum. Additionally, a local increase in the ratio in the low-energy region is observable. This low-energy region is a distinct feature of the jet-inclusive event classes. Since the H_T trigger efficiency cut is applied, it can not be found in the 2jets exclusive class, as mentioned in the discussion above. However, for the jet-inclusive class, higher jet multiplicity classes contribute. Since higher jet multiplicity classes share the S_T with more than two objects, but the calculated and plotted S_T in fig. 4.4a for the 2jets jet-inclusive event class only considers the two leading jets during its evaluation, the low energy tail (below the trigger cut at 1400 GeV) is created. The m_inv distribution (fig. 4.4b) also shows a tail at lower energies, which can be interpreted with the contributions of higher multiplicity classes, as already explained. At low energies, a local increase of the data/MC ratio is also visible. Apart from this, the ratio is around ≈ 0.5 and decreases with increasing energies, as already observed for the exclusive class. Note also that in the high energy region, one bin shows an exceptionally high MC uncertainty. This can be most probably attributed to a single event that induces a high uncertainty on one of its corrections, e.g. PDFs weight correction factor.

Two distributions for the 4jets exclusive class are shown in the appendix in sec. B.2 in fig. B.3. Similar behavior is observed as was already described for the 2jets classes.

Apart from the investigation of kinematic distributions, the dependence of the data/MC agreement from the jet multiplicity should be analyzed. For this purpose, the integrated event counts of the event classes are compared in figure 4.5. All event classes with data are shown in the plot. First, the exclusive classes (fig. 4.5a) are discussed. Of all jet-only exclusive classes, the 4jets class is the most inhabited one. The deviation between data and MC, which was already observed in the kinematic distributions, can also be seen in the comparison of the integrated counts. While the data/MC ratio is around ≈ 0.5 for lower jet multiplicities $(2-5)$, for higher jet multiplicities the ratio increases up to ≈ 1. Therefore, a significant dependence of the ratio on the jet multiplicity is observed. Note that jet-only event classes with very high jet multiplicities were found in the 2018 data, the most exotic jet-only event identified in the data has 17 jets. For these very high jet multiplicities, the MC prediction is observed to deviate even more significantly from data as already seen for low multiplicities, however, the statistics are also very low for these high-multiplicity classes[12]. Also, the single jet exclusive class shows larger deviations from MC simulation as the ratio of ≈ 0.5 observed for multiplicities of $2-5$. As already noticed in the discussion above, most of the jet-only classes are dominated by the QCD MC samples. With increasing jet multiplicity, the $t\bar{t}$ process group contribution increases, nonetheless, up to the 14jets class, the MC prediction stays QCD-dominated. The 1jet class shows significant contributions from the γ process group. Since the jet-inclusive classes simply accumulate all classes with higher jet multiplicities than given in their class name, the plot for these classes (fig. 4.5b) duplicates the behavior seen for the exclusive classes. In particular, it can be seen that the 1jet exclusive class does not contribute significantly to the corresponding jet-inclusive class because of its low event count, as already observed in the overview of the exclusive classes. The increasing data/MC ratio towards 1 with increasing jet multiplicity (except for very high multiplicities) is of course also observed in the plot for the jet-inclusive classes.

The improved agreement of data and MC (attributed to the jet multiplicity dependence of the ratio) for higher jet multiplicities can be well illustrated with the distributions for the 10jet exclusive class in the appendix in sec. B.2 in fig. B.4.

4.10.2 Global observations

The resulting distributions for the jet-only classes of three different sample-trigger-configurations were presented in the last section (sec. 4.10.1) as well as in the appendix (sec. B.2–B.4). Since most of

[12] While the MC event weight can take any floating point number, only 0 and 1 is possible for a data weight. Therefore comparison at low statistics is difficult.

the observed characteristics are similar, these global observations should be summed up briefly in this section.

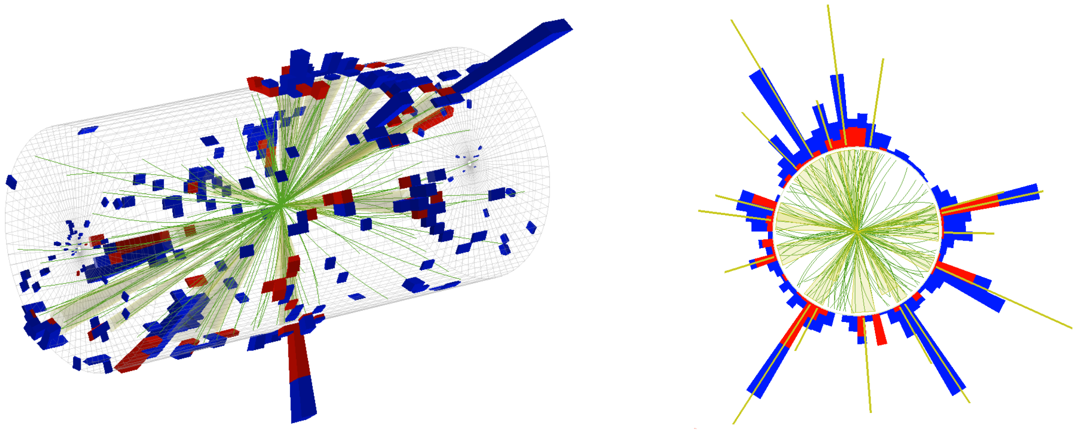
Most importantly, a significant deviation between data and MC has been observed for all sample-trigger combinations in all distributions, no matter whether energy-like or angular, for most of the classes. For jet multiplicities in the range $2-5$, the data/MC ratio is approximately 0.5. The ratio is observed to have a dependence on the kinematic properties of the jets, for increasing S_T and m_inv, a decrease of the ratio can be seen. This is most relevant for the mass distributions. Additionally, a strong dependence on the jet multiplicity of the event was observed for all sample-trigger combinations. As already stated, for medium multiplicities of $2-5$ jets, the ratio remains approximately constant at ≈ 0.5, however an increase of the ratio towards 1 is observed with increasing jet multiplicities up to 11 jets[13]. Very high jet multiplicities (> 11 jets) as well as single jet events do not seem to be well modeled by MC, since the deviation for these classes is even higher.

The exclusive event classes show the applied trigger efficiency thresholds as sharp cuts, while the jet-inclusive classes show a turn-on peak, with events below the trigger efficiency cut threshold. These events originate from the higher jet multiplicity classes that are included in the plots for the jet-inclusive class distributions. Also, the energy-like distributions, meaning transverse momenta sum, and the masses are generally observed to have decreasing data/MC ratio for increasing energies[14].

In the low energy region (energies below the trigger efficiency cuts), a local increase of the data/MC ratio is visible. It appears as a bump-like feature and can be found for all jet-inclusive classes, most prominently in the mass distributions.

Another observation in all sample-trigger combinations is that all jet-only classes except for very high jet multiplicities are dominated by QCD MC samples. For jet multiplicities $2-5$, the QCD fraction of the total MC background is found to be $> 95\,\%$. This is also the main reason for the large combined MC uncertainties of $\approx 50\,\%$ since the QCD samples are produced at LO and MUSiC (as well as this analysis) applies a generous $50\,\%$ uncertainty on all LO cross sections, as discussed in sec. 4.6.

4.10.3 Details on the 17 jet event



(a) 3D view **(b)** View from the beam axis

Fig. 4.6: Event display of the 17 jet event with calorimeter response (ECAL: red, HCAL: blue), tracks (green) and reconstructed jet cones (yellow).

With the H_T trigger and the described selection requirements, one event containing 17 jets was found

[13] The energy dependence of the ratio (described above) is not depicted in the total event count comparison since essentially the integral of the kinematic histograms is calculated for this comparison.

[14] Since natural units are used, the energy-like quantities transverse momenta sum and invariant mass have in fact the unit of energy. Because of this, the thesis from now on uses the undifferentiated term energy dependence when referring to the dependence of the data/MC ratio on the energy-like distributions transverse momenta sum and invariant mass.

in the 2018 CMS dataset. Such high jet multiplicities are very rare. Efforts were made to produce an event display for this specific event, it is shown in fig. 4.6. Additional views are provided in the appendix in fig. B.13 and the event information is provided in tab. B.1. The jets in this event have transverse momenta in the range of $p_\mathrm{T} \in [52, 173]\,\mathrm{GeV}$, this explains why the event was only found with the H_T trigger and not with the p_T trigger, where $p_\mathrm{T} > 600\,\mathrm{GeV}$ is required for the leading jet. The finding of this event underlines the potential of jet triggers to discover exotic events and therefore emphasizes the idea to extend MUSiC with such jet-only final states.

4.11 Discussion

The jet-only classes found in the dataset were laid out and selected distributions were presented. The most striking observation from this is the large deviation between data and MC. The deviation shows dependencies on different parameters, however, it seems to have a significant constant contribution since the data/MC is systematically lower than one over the whole distribution for small and medium jet multiplicities. The two observed dependencies of the ratio are the energy (or an energy-like quantity as S_T or m_inv) and the jet multiplicity in the class. The source for this deviation is not clear initially. However multiple potential sources could be thought of.

Generally, deviations between data and MC are a potential sign for new physics. Therefore the first thing that has to be discussed is whether the observed deviations potentially are caused by BSM phenomena. The deviation shows a significant constant fraction over all different distributions, both energy-like and angular, and also for a large number of jet multiplicities. This makes it relatively unlikely that the effect is caused by new physics phenomena.

The lower energy region has been accessible by particle physics experiments for many years now. Therefore it would be unlikely, that hints for BSM physics in this lower energy range with this order of magnitude would not have been found by now. This claim is supported by most of the dedicated CMS searches, which exclude affection of the most common BSM theories to the lower energy range by at least 95 % [92]. For example, a recent dijet resonances analysis [53] excludes dijet mass resonances with masses lower than 2.6 TeV, for details see sec. 4.1.1. Similar arguments can be made for the 3jets class with trijet resonance searches [54]. Therefore, it is concluded, that, at least for the common BSM theories, no significant contribution would be expected in the low energy region of the mass distributions for the 2jets (3jets) exclusive class according to these results. However, a large constant fraction of the deviation is observed over the whole distribution range.

The significant deviation in data and MC was also observed with independent analysis code produced and tested by other members of the MUSiC team. Also, all steps of the analysis presented here were carefully checked separately, making a bug in the analysis as a reason for the observed deviation improbable.

The results of this analysis should be briefly compared to the previous study with jet triggers in MUSiC in a master thesis by A. Albert [68], which was presented in sec. 4.1.2 and in the appendix in sec. B.1. Note that a full comparison is not possible, since the old study uses the 2012 CMS dataset recorded at a lower center-of-mass energy, and the object selection and triggers are different. Still, the distributions should be viewed against each other. Compared to this analysis, the configuration with the p_T trigger and p_T QCD samples (sec. B.4) is the closest to the work in the master thesis. Apart from the fact that this analysis shows a significant constant deviation between data and MC and the master thesis does not, similar systematic behavior can be observed. The data/MC ratio is observed to decrease with increasing energy in both analyses. Additionally, both analyses show a local disagreement in the low-energy region. The trigger efficiency cut features can be observed in a similar manner in the plot (except for the exact position of the turn-on peak, which is of course different because of the different triggers used), and at first glance, the distributions seem to have matching shapes. Since the energy dependency of the data/MC ratio was also observed in the master thesis, it is assumed that this effect is independent of the general constant offset between data and MC observed in this analysis.

The energy and jet multiplicity-dependent fractions of the deviation might in fact be related to problems in the MC modeling, which is not unexpected for such complex QCD processes.

A wrong calculation of MC weights could explain the observed deviations between data and MC or at least the constant part of the observed deviation. Many different factors contribute to the MC weight, as discussed in sec. 3.2.3.5. A constant offset could easily be produced by a wrong factor in the weight formula (eq. 3.2). Since the cross sections for the QCD samples are very high in comparison to the other samples, even small deviations in the cross section values can have a large impact on the MC weight of these samples. The QCD samples dominate the jet-only classes, therefore this effect has the potential to explain at least the constant fraction of the deviation. It should be pointed out, that QCD processes are very complex physics processes, therefore misestimations are not unlikely, still, the question is whether a deviation of this order of magnitude could be accounted for by the cross sections.

Since the integrated luminosity is fixed by the dataset and the performed object corrections (sec. 3.2.3.3) are obtained and applied as recommended by CMS, these are found to be an unlikely cause of the observed deviation.

This thesis aims to explore the possible extension of MUSiC with the presented jet-only classes. In this chapter, these classes have shown significant, mostly constant deviations. From the discussion above it is concluded, that a deviation of the observed significance, with a large constant offset, which is visible in all distributions, is most likely not caused predominantly by new physics phenomena. If it is assumed that this deviation would not be related to new physics, but is caused by another source, the deviation should be reduced to include these classes in the MUSiC framework. The main problem when including the classes without further corrections and processing would be, that the MUSiC RoI scanner would detect severe deviations, and therefore the $\tilde{p}$-value for the classes would be lower than would be expected if the deviation is not related to new physics. Therefore, the next parts of this thesis introduce and explore additional analysis steps to weaken the observed deviation. The difficulty with this is, that the corrections should ideally be model-independent, to minimize the signal bias of the whole analysis.

The next two chapters address different properties of the observed deviations. Chapter 5 explores the effect of jet merging algorithms that are commonly used by dedicated jet analyses in CMS, aiming at reducing the dependence of the data/MC deviation from the energy and jet multiplicity. Since the results presented in this chapter seemed mostly similar for all sample-trigger combinations, only the dataset with p_{T}-binned QCD samples and the S_{T} trigger is used for the continued analysis. A generalized algorithm is developed for the jet-only classes for potential use in MUSiC. After the ratios have been flattened by this algorithm, the constant offset can be addressed. Chapter 6 suggests a normalization scheme for the QCD samples, however with a scaling factor calculated from an independent dataset.

5 Wide jet algorithm

5.1 Introduction

Assuming that the observed overall deviation is not caused by new physics, efforts should be made to correct the MC. Before a normalization could be performed, this thesis suggests an additional step to flatten the data/MC ratio. Specifically, the dependence of the ratio from the energy and the jet multiplicity should ideally be weakened.

It was found that the application of a wide jet algorithm, which is commonly used in various dedicated CMS jet analyses [53, 54, 69, 93–96] could have the desired effect. The next section introduces the algorithm that is used in the referenced CMS analyses and discusses its background. After this, a generalized version of this algorithm, which allows arbitrary jet multiplicities and was developed for this thesis, is presented and applied to the dataset. Finally, the results of the algorithm are discussed.

5.2 Wide jet algorithm in dedicated searches

The wide jet algorithm is used in numerous CMS dijet resonance searches, including [53, 69] with Run 2 data and [93–96] with Run 1 data. These publications have applied such an algorithm to reduce the effect of gluon emission from the final state quarks [53]. This process was already explained in sec. 1.3.2.2 and can potentially lead to an increase in jet multiplicity if a gluon emission occurs. In the following paragraph, the working principle of the algorithm will be introduced using the implementation in one of the most recent analyses [53] that use this approach.

Since the respective analysis focuses on dijet resonances, the result of the algorithm should be two jets. First, the two jets with leading p_{T} of every event are selected as so-called seed jets. After this, all remaining jets, which satisfy the selection requirements and are spatially close to one of the seed jets, are merged sequentially. The merging requirement for the non-seed jet is a distance to a seed of $\Delta R < 1.1$ (where ΔR is defined similarly to eq. 2.3). If this requirement is satisfied by both seed jets, the non-seed jet is merged with the seed closer to it, meaning the seed with the smaller ΔR. The jet merging is performed by adding the four-momenta of the seed jet and the non-seed jet that should be merged. All non-seed jets which are not sufficiently close to a seed are ignored from the further analysis.

Usually, these analyses also require a $|\Delta\eta|$ cut between the resulting two wide jets. This is motivated by the fact that t-channel dijet events are usually expected to have an angular distribution that favors a large pseudorapidity difference between the final state quarks [53]. To reduce the influence of this understood dijet background, other analyses apply a cut by defining a maximum accepted pseudorapidity difference $|\Delta\eta| < \Delta\eta_{\mathrm{thres}}$. Values for this threshold vary for different analyses between $\Delta\eta_{\mathrm{thres}} = 1.1$[1] [53] and $\Delta\eta_{\mathrm{thres}} = 1.3$ [69] for the two referenced analyses with Run 2 data.

CMS analyses focusing on higher jet multiplicity systems have adapted the dijet approach of wide jets, implementing their own merging algorithms. Notably, a trijet analysis [54], which has been accepted by CMS but not yet published at the time of writing this thesis, uses such an approach. Conceptually, the algorithm is mostly similar. One difference however is that the referenced trijet analysis requires

[1] Actually, the referenced dijet analysis differentiates different regions by $|\Delta\eta|$ bins. The cut $|\Delta\eta| < 1.1$ refers to the signal region of the analysis [53].

a minimum seed momentum $p_\mathrm{T} > 100\,\mathrm{GeV}$. The merging happens similarly to the dijet analyses. Another difference is that for the pseudorapidity cut the maximum $|\Delta\eta|$ is demanded to be below a threshold of $\Delta\eta_\mathrm{thres} = 1.6$ since with more than two objects, multiple pseudorapidity differences have to be considered.

5.3 Generalized wide jet algorithm

The next section presents the generalized wide jet algorithm for potential use in MUSiC. To account for all possible jet multiplicities, the algorithm was slightly modified. This generalized wide jet algorithm is illustrated in fig. 5.1. The following section briefly explains the separate steps. Note that the systematics used in the study are the same as presented in sec. 4.6. It should be noted that the developed generalized wide jet algorithm is applied after the object selection but before the filling of the classes.

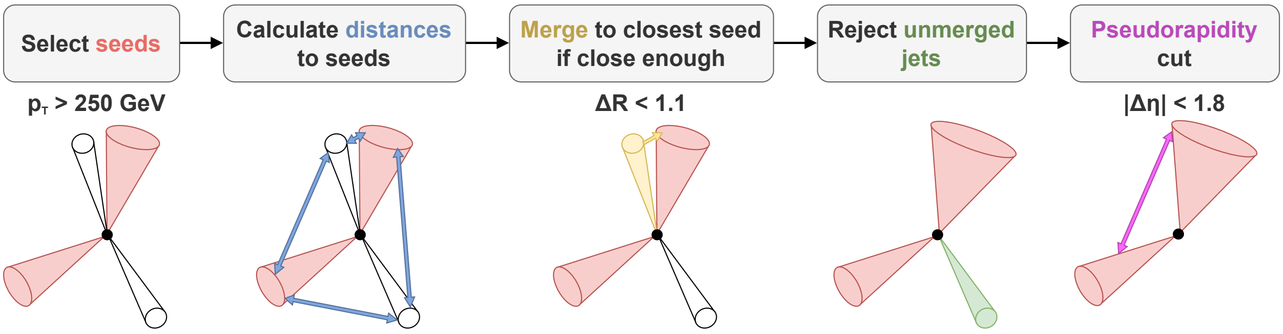


Fig. 5.1: Illustration of the generalized wide jet algorithm.

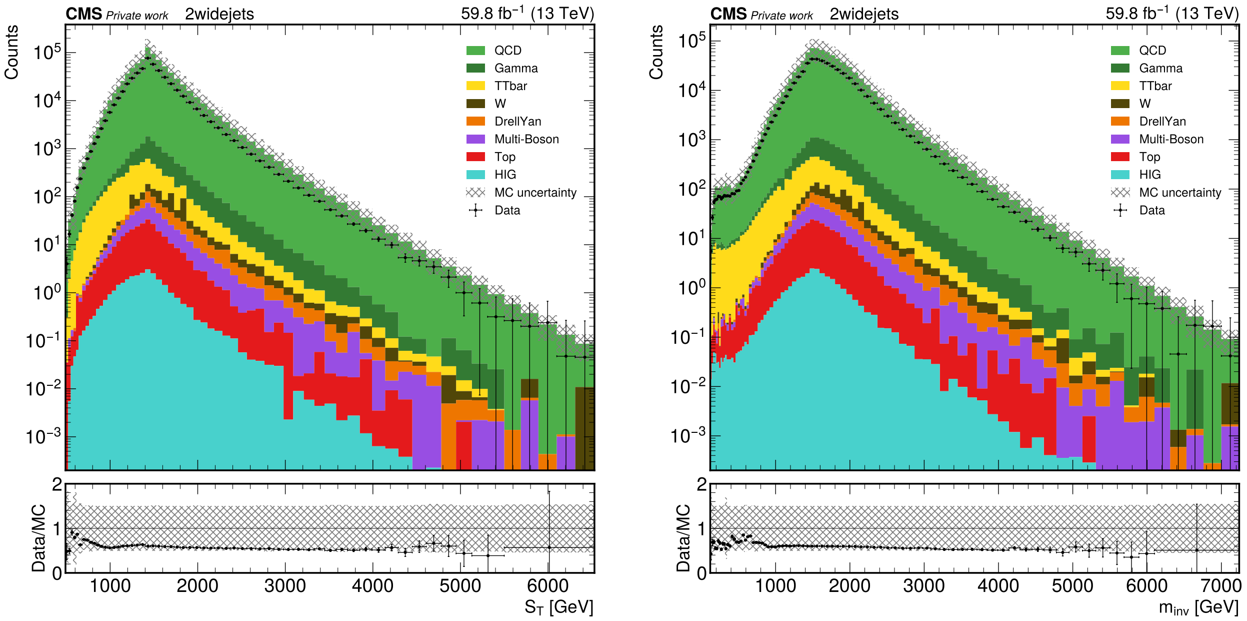
Since the jet multiplicity is a free parameter in MUSiC and only determines the event class the object is sorted into, a seed selection similar to the presented dijet and trijet analysis via the first leading jets is not feasible. Instead, the idea of a minimum transverse momentum requirement for seed jets beyond the object selection threshold for jets is followed, which was already a secondary part of the trijet analysis (sec. 5.2) [54]. The generalized approach requires a jet seed transverse momentum of $p_\mathrm{T} > 250\,\mathrm{GeV}$. All other jets are regarded as non-seed jets. The exact value for this threshold was determined in a coarse study (see sec. 5.5).

The merging algorithm used in the dedicated di- and trijet analyses is already applicable to the generalized approach. Therefore, the same approach as described for the dijet analyses (sec. 5.2) is used: Non-seed jets with $\Delta R < 1.1$ to any seed are merged to the closest seed.

The differential pseudorapidity cut is generally unrelated to the wide jet merging itself, however, such a cut is applied also in the generalized wide jet algorithm since this approach was proven in numerous dedicated analyses, that were already presented. The procedure for the generalized pseudorapidity cut can be modified from the trijet analysis (sec. 5.2) [54]. The maximum pseudorapidity difference considering all wide jets is required to be below a threshold: $\max\{|\Delta\eta|\} < 1.8$. The exact value for this threshold is not adopted from the original analysis but was also determined in a coarse study (see sec. 5.5).

5.4 Results

In the following, the results for the jet-only classes after applying the generalized wide jet algorithm, which was introduced in the last section, are presented. As already stated in sec. 4.11, only the configuration with the p_T-binned QCD samples and the H_T trigger is used for the continued analysis. For this configuration, there are 7 exclusive and 7 widejet-inclusive classes found in data and 9 exclusive and 9 widejet-inclusive classes in MC. To differentiate the plots and class names, in the following, the classes are labeled with "widejets" instead of "jets" when the algorithm was applied.

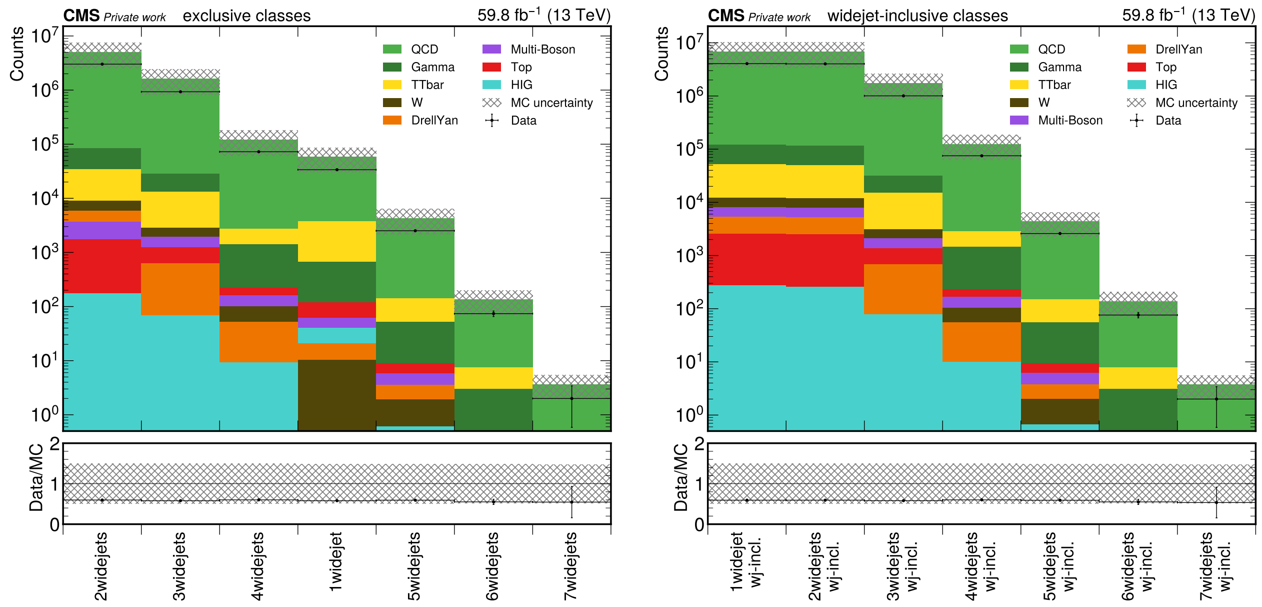


(a) Transverse momenta sum S_{T}

(b) Invariant mass m_{inv}

Fig. 5.2: Distributions for the 2widejets exclusive class.

The 2widejets exclusive class is presented first. Fig. 5.2 shows the transverse momentum sum S_{T} (fig. 5.2a) and the invariant mass m_{inv} (fig. 5.2b) for this event class. It can be observed that the exclusive class distribution shows tails below the trigger efficiency cut of $S_{\mathrm{T}} = 1400\,\mathrm{GeV}$. At first glance, this is unexpected, however, this behavior can be explained with the wide jet algorithm. Since jets that are not selected as seeds and also not spatially close enough to seed jets are rejected, it is possible that the S_{T} of the merged wide jets is lower than the S_{T} of all jets that were selected. Consequently, it is possible that an event passes the S_{T} trigger efficiency cut but does have a lower S_{T} after the merging when the classes are filled. The same explanation can be applied to the tail in the m_{inv} distribution since mass and momentum are correlated. Apart from this, the event statistics in the class have slightly changed after applying the wide jet algorithm. This is further discussed when the integrated class counts plot is presented below. Data and MC still show a significant deviation with a data/MC ratio of still ≈ 0.5, however in comparison to the same class without the wide jet algorithm applied (fig. 4.3), it becomes obvious that the ratio is significantly flattened, especially for the invariant mass plot (fig. 5.2b), which showed a strong decrease with increasing mass before (see fig. 4.3b).
The 2widejets widejet-inclusive and the 4widejets exclusive class are briefly discussed in the appendix in sec. C.1. In the higher multiplicity classes, decreased statistics are observed.
Finally, the integrated class counts should be discussed. The corresponding plot is shown in fig. 5.3. The first observation is a significant reduction in the number of classes compared to the pre-widejet analysis (fig. 4.5). The jet multiplicities seem to decrease drastically by applying the jet merging. The reduction of jet multiplicity might be caused by the fact that jets are merged into seeds or are rejected during the wide jet merging and seed selection. Another interesting observation is that the 1widejet exclusive class became much more inhabited, compared to the 1jet exclusive class before the merging (by about three orders of magnitude more event counts). This can also be understood with the wide jet algorithm, suggesting that a significant number of events with two or more jets underwent either jet merging or the rejection of low energetic jets, which reduced the jet multiplicity to one. While the 1jet class showed significantly worse MC modeling than the other classes in the pre-widejet approach, the data/MC ratio for this class is now comparable to the ratio of the other classes with ≈ 0.5. Although



(a) Exclusive classes **(b)** Widejet-inclusive classes

Fig. 5.3: Comparison of the integrated event counts for the wide jet classes.

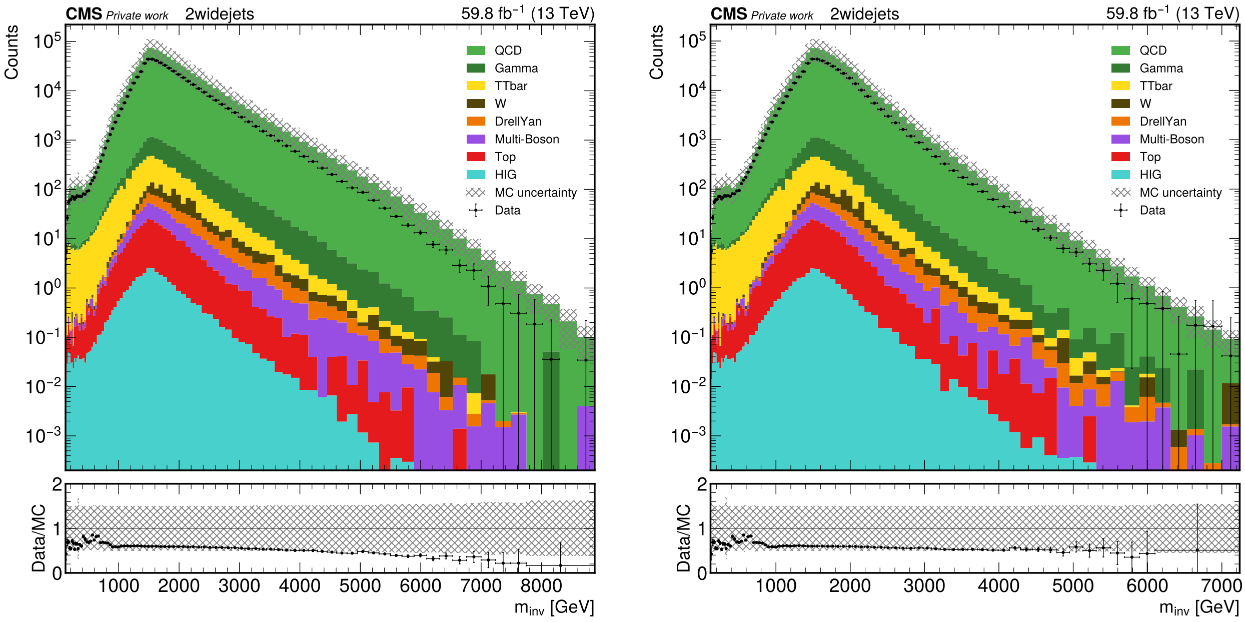
this has already been the case for the jet multiplicities $2-5$ before applying the wide jet algorithm, the ratio now appears almost flat with respect to the jet multiplicity, so the dependence of the ratio on this parameter was decreased, especially for the 1widejet and $\geq$ 6widejets classes. As already said, all other events that filled the higher multiplicity classes were either rejected or their jets were merged, resulting in a decreased multiplicity. The integrated count plot for the widejet-inclusive classes is presented in fig. 5.3b. Since the 1widejet widejet-inclusive class event count equals the total count of selected events in this analysis (because all other physics objects are vetoed), the event count before and after the applied wide jet merging can be compared. Before, there were 5,598,760 events in this class in data (for the corresponding sample-trigger-configuration), after, there were 4,052,644. This means that about 28 % of the events are rejected when applying the generalized wide jet algorithm with respect to the pre-widejet event count. A rejection of events because of the wide jet algorithm happens when no jet has enough momentum ($p_\mathrm{T} > 250\,\mathrm{GeV}$) to qualify as a wide jet seed. Since the MUSiC selection requirement for jets only demands $p_\mathrm{T} > 50\,\mathrm{GeV}$, it is very well possible that jets might have lower energies than they would need to qualify as a seed jet since the events are S_T triggered. This is especially true for high jet multiplicity events.

5.5 Optimization

The parameters $p_{\mathrm{T,thres}} = 250\,\mathrm{GeV}$ for the seed selection and $\Delta\eta_{\mathrm{thres}} = 1.8$ for the pseudorapidity cut were not known initially, but were chosen based on a broad study that was conducted, varying the parameters. Because of the limited working time for the thesis, only a few values per parameter were tested, however, from this, some behaviors already became apparent. Of course, this testing process was conducted prior to running the analysis with the final chosen values, which was already presented in sec. 5.4. Still, the results for the different parameters concerning the energy dependence and jet multiplicity dependence of the data/MC ratio should be briefly presented in this section. The tested configurations include $p_{\mathrm{T,thres}} \in \{150\,\mathrm{GeV},\ 250\,\mathrm{GeV}\}$ and $\Delta\eta_{\mathrm{thres}} \in \{1.4,\ 1.8\}$ or no pseudorapidity threshold.

The energy dependence of the data/MC ratio is observed with the invariant mass distributions. Fig. 5.4 shows a direct comparison between no pseudorapidity cut and a pseudorapidity cut of $\Delta\eta_{\mathrm{thres}} = 1.8$.

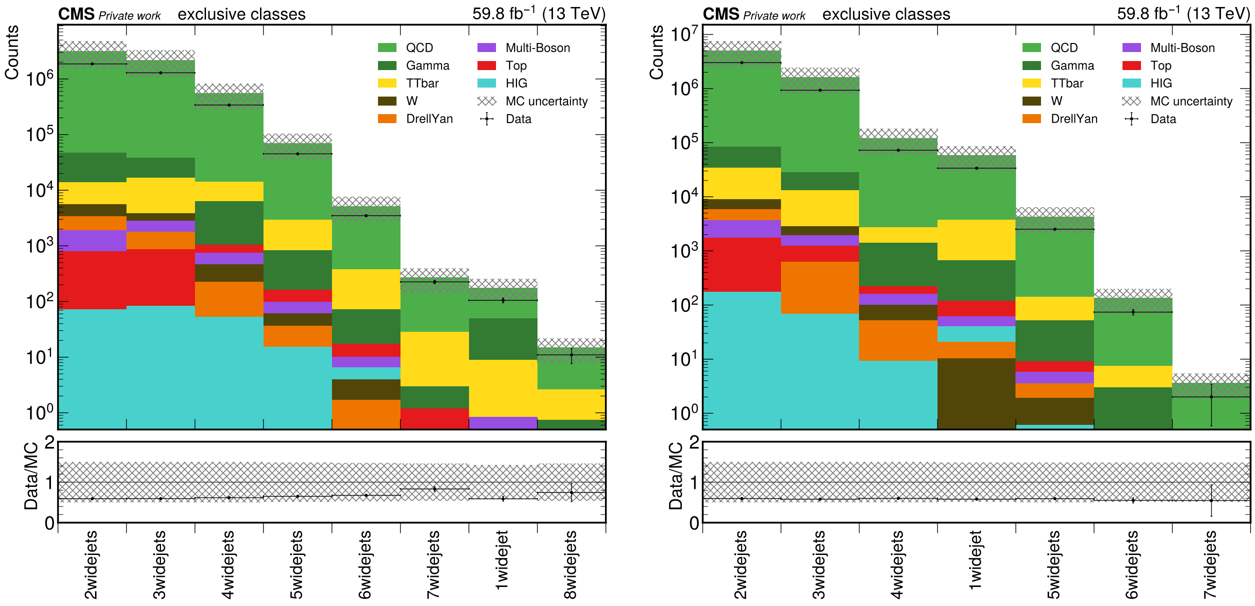
It becomes apparent that this cut influences the decrease of the ratio for increasing energies. When no pseudorapidity cut is applied, the data/MC ratio significantly decreases in the high-energy region. Applying a cut of $\Delta\eta_{\text{thres}} = 1.8$ seems to flatten the ratio and thus decrease the energy dependence. It is concluded that a pseudorapidity cut with $\Delta\eta_{\text{thres}} = 1.8$ should be applied.



(a) $p_{\text{T,thres}} = 250\,\text{GeV}$ and no $|\Delta\eta|$ cut

(b) $p_{\text{T,thres}} = 250\,\text{GeV}$ and $\Delta\eta_{\text{thres}} = 1.8$

Fig. 5.4: Invariant mass distributions m_{inv} for the 2widejets exclusive class for two selected merging and pseudorapidity cut parameters.



(a) $p_{\text{T,thres}} = 150\,\text{GeV}$ and $\Delta\eta_{\text{thres}} = 1.8$

(b) $p_{\text{T,thres}} = 250\,\text{GeV}$ and $\Delta\eta_{\text{thres}} = 1.8$

Fig. 5.5: Integrated event counts for the exclusive widejet classes for two selected merging and pseudorapidity cut parameters.

With the integrated class counts for the exclusive classes, the jet multiplicity dependence of the ratio is monitored. Flatter ratios (with respect to the jet multiplicity) are apparently achieved with the stricter seed selection threshold, therefore $p_{\text{T,thres}} = 250\,\text{GeV}$ is selected. This is illustrated in fig. 5.5

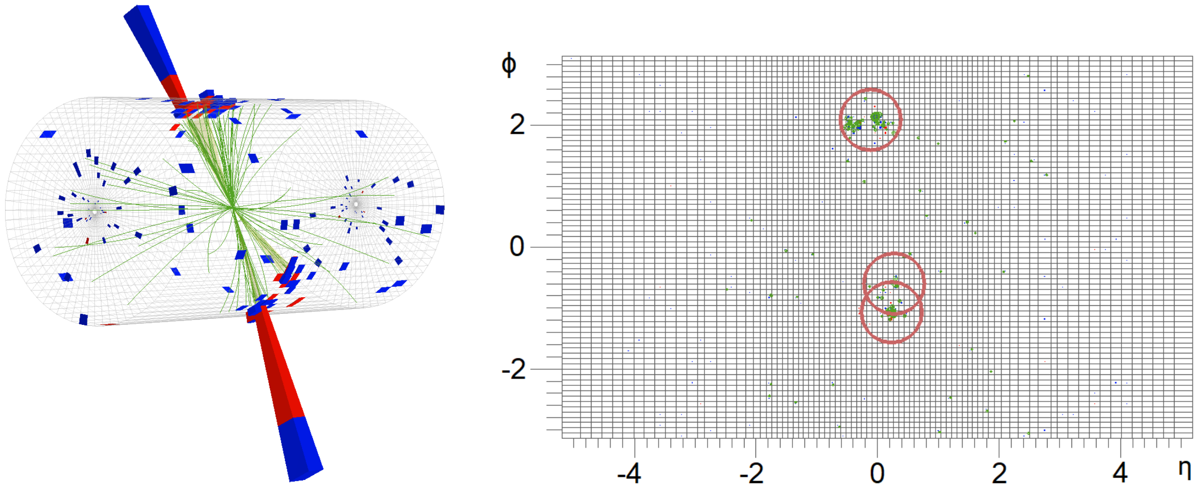
for the two thresholds $p_{\mathrm{T,thres}} = 150\,\mathrm{GeV}$ and $p_{\mathrm{T,thres}} = 250\,\mathrm{GeV}$.

In the appendix, a larger selection of plots is presented in sec. C.2. From these, it becomes apparent, that the pseudorapidity cut has little influence on the multiplicity dependence and the seed threshold has little influence on the energy dependence of the data/MC ratio. Instead, it seems like the cuts are only strongly correlated to one of the dependencies of the ratio, as described above.

The selected value for the merging distance of $\Delta R_{\mathrm{thres}} = 1.1$ could also be contested since it might also influence the event cut. However, this value is used consistently in many publications that made use of a wide jet algorithm [53, 54, 69]. Therefore this parameter is kept at its default value.

With this broad overview of the different parameters, and the thresholds used for the analysis were selected. Note, however, that the test conducted in this section to choose these parameters is very broad and lacks a deeper analysis. Therefore, if a wide jet merging method should be used in the future, a dedicated deeper study on the parameters should be conducted.

5.6 Illustration with event displays



(a) 3D view

(b) Angular position of the jets in $\eta - \phi-$plot

Fig. 5.6: Event display of an example event that undergoes 3jet → 2widejet merging with calorimeter response (ECAL: red, HCAL: blue), tracks (green) and reconstructed jet cones (yellow). Note that there is a fourth small energy deposit in the calorimeter, however, this was not selected as a jet according to the MUSiC object selection criteria.

The wide jet algorithm was explained above in sec. 5.3. The merging process was described in detail, however, it was decided to present one specific example of an event that undergoes the merging. The event that will be presented has three selected jets. Two of the jets are close enough to each other that they can undergo merging, and therefore after applying the algorithm, the widejet multiplicity is only two. Fig. 5.6a shows the event display for the event. Note that one of the subfigures shows the $\eta - \phi-$plot (fig. 5.6b). This plot shows that two of the three jets in fact satisfy the spatial merging requirement $\Delta R < 1.1$ and therefore the two spatially close jets undergo the merging. Although it is not shown here, the jets in the presented event satisfy the other relevant kinematic requirements to qualify as seeds or non-seeds (transverse momenta thresholds). The jet that will be merged with the close seed has a much lower transverse momentum than the seed jet. Therefore it is very likely that in this event, the third jet is in fact created after gluon emission from one of the final state quarks. The event info for the presented event can be found in tab. C.1 in the appendix.

5.7 Discussion

The adapted wide jet approach, which was used in numerous CMS jet analyses, proved to flatten the data/MC ratio. Therefore it could be shown that, in principle, the energy and jet multiplicity dependencies of the ratio can be reduced by applying such an algorithm as well as a pseudorapidity cut. The integral question is whether applying these additional steps and cuts could decrease the sensitivity to potential new physics signals or even bias potential signals, this will be discussed in this section.

It is true that these additional steps lead to a significant decrease in the data count of about 28 %, which is rejected from the further analysis. This is can be regarded as problematic since potential signals in these events can not be found by the search algorithm anymore. Nevertheless, the preprocessing leads to a flattened data/MC ratio, and therefore can be viewed as a step towards acceptable background modeling, apart from the constant offset.

It should also be noted that the thresholds of the wide jet merging and pseudorapidity cut were only determined by a few tests with different parameters. If would be decided to continue with the wide jet approach for including jet classes to MUSiC, a dedicated study entirely focused on the optimization of the parameters would be necessary. Such a study could potentially improve the performance of the wide jet approach even further.

Now a potential signal bias should be discussed. First, it should be noted that signal bias implies that a potential signal in data is treated as background and therefore stays hidden or is significantly changed in its shape or amplitude. In the case of this analysis, exactly similar algorithms are applied to both data and MC. It should also be pointed out that the CMS jet analyses, that use this wide jet algorithm, claim that their analysis is model-independent [53, 69], implying that no significant bias is induced on the signal shape by applying the wide jet algorithm. Therefore it is assumed that applying a similar wide jet algorithm would not imply a significant signal bias. If it would be decided to continue with a wide jet approach in the future, a dedicated signal bias study on this topic could be helpful to definitely exclude this potential issue.

One potential problematic point is the reduced number of classes when applying the algorithm. Since MUSiC aims at scanning deviation in a large number of classes and is mostly sensitive to deviations occurring in these classes, reducing the number of newly added classes to MUSiC might have an impact on the signal sensitivity. This point can not be mitigated since it is a direct consequence of the wide jet algorithm. However it can be argued, that it would be better to analyze fewer jet-only classes, that model the background correctly, than more classes with an uncorrected MC background.

In the last chapter, the hypothesis was raised that the deviation between data and MC can be separated in a significant constant offset (potential cross section or MC weight problem) and an energy and jet multiplicity dependent part (potential problem of MC simulation quality). In this chapter, it could be shown that applying a wide jet algorithm and a pseudorapidity cut, in fact, leads to a significant attenuation of the observed dependencies of the ratio. Since the two applied additional steps target low-energetic jets, potentially originating from gluon emissions, and jets with large pseudorapidity differences, these results suggest that there might exist deficits in MC modeling in these specific cases.

6 Possible renormalization

6.1 Introduction

In previous discussions, it was found that the energy and jet multiplicity dependence of the data/MC deviation could be weakened, leaving a relatively constant offset to correct. This desired correction is of course based on the assumption that the deviation is not caused by new physics phenomena but some other reason, that could not be identified with certainty. One possibility to address the constant offset between data and MC is to rescale the MC according to a scale factor, which is obtained from data. This process is here referred to as normalization. A normalization strategy is proposed and applied to the MC dataset in this chapter.

6.2 Normalization method

The proposed method for normalization is presented in the following sections. Fig. 6.1 presents the workflow of the proposed normalization scheme.

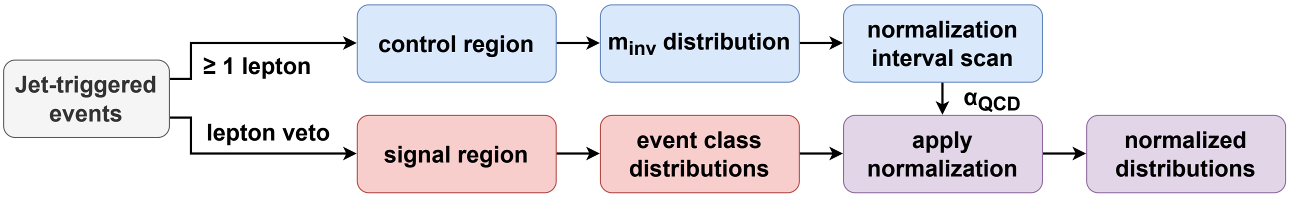


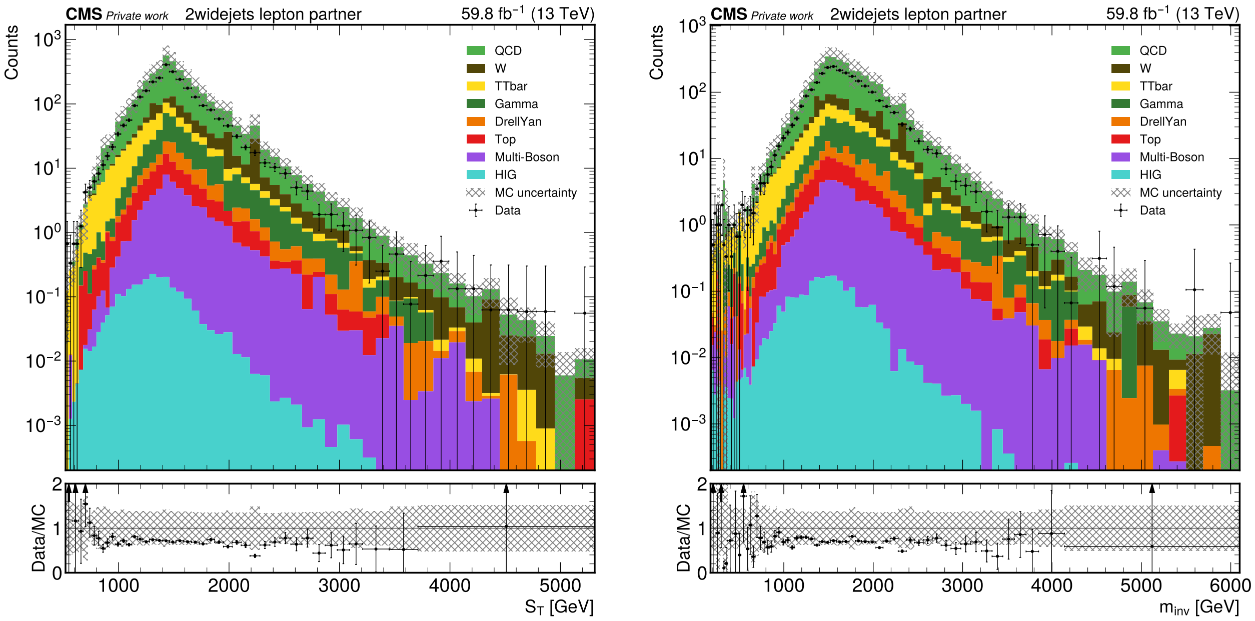
Fig. 6.1: Illustration of the proposed normalization scheme.

The presented method assumes that the QCD MC samples show a constant offset to data, which should be corrected by applying a normalization factor to these samples. The contributions of other non-QCD MC samples to the respective event classes are assumed to be correct. This assumption is based on the fact that for the jet-only event classes, a deviation as high as observed can only originate from these samples because they dominate the event classes. Therefore, only these samples should be rescaled to remedy the difference in the data. Note that normalization means that a constant factor is applied to all associated QCD MC event weights, over all kinematic and angular regions. Since QCD dominates over the full range of all distributions for the jet-only classes, only a very wide signal that is constant over the full range of the distribution is potentially biased when applying the normalization procedure.

6.2.1 Lepton partner classes as independent dataset

Generally, it is favorable to not obtain background estimation from the same dataset that is analyzed but to use a disjoint subset of the dataset to calculate the background model and then apply it to the other subset of the dataset which will be analyzed. This concept of disjoint regions is realized in many dedicated CMS searches, in particular when the background is fully estimated in a data-driven way, e.g. by rescaling data from different regions or with a background fit. The region from which the background is obtained is frequently called the control region, and the region where the final analysis

is performed is called the signal region. For example, the referenced dedicated dijet search in CMS [53] uses one signal region and two control regions defined by different $|\Delta\eta|$ regions.
Usually, dividing the dataset into signal and control regions implies additional cuts. Therefore the signal region has a reduced number of events since it is only a subset of the original dataset. This analysis however aims to use all events of the jet-only classes as the signal region. Because of this, a different control region has to be established. It is proposed to invert the lepton veto for this purpose. The control regions would then be all the partner classes to the jet-only classes, with an inverted lepton veto, meaning that they have to include at least one lepton (e, μ)[1]. Note that only the lepton veto is lifted for the control regions but the photon and MET veto is still applied. The object selection requirements and object corrections are still the same as presented before (see tab. 3.1). Most importantly, still, only the jet trigger and the Jet_HT dataset are used when the lepton veto is inverted. Also, the generalized wide jet algorithm (from sec. 5.3) is applied to the lepton partner classes to ensure comparability.
In the following, the lepton partner classes are briefly presented. As explained above, the generalized wide jet algorithm is continued to be used in this section, therefore there exist no "jet" classes but only "widejet" classes. The analysis identifies 5 exclusive and 5 widejet-inclusive classes in data and 7 exclusive and 7 widejet-inclusive classes in MC.



(a) Transverse momenta sum S_T

(b) Invariant mass m_inv

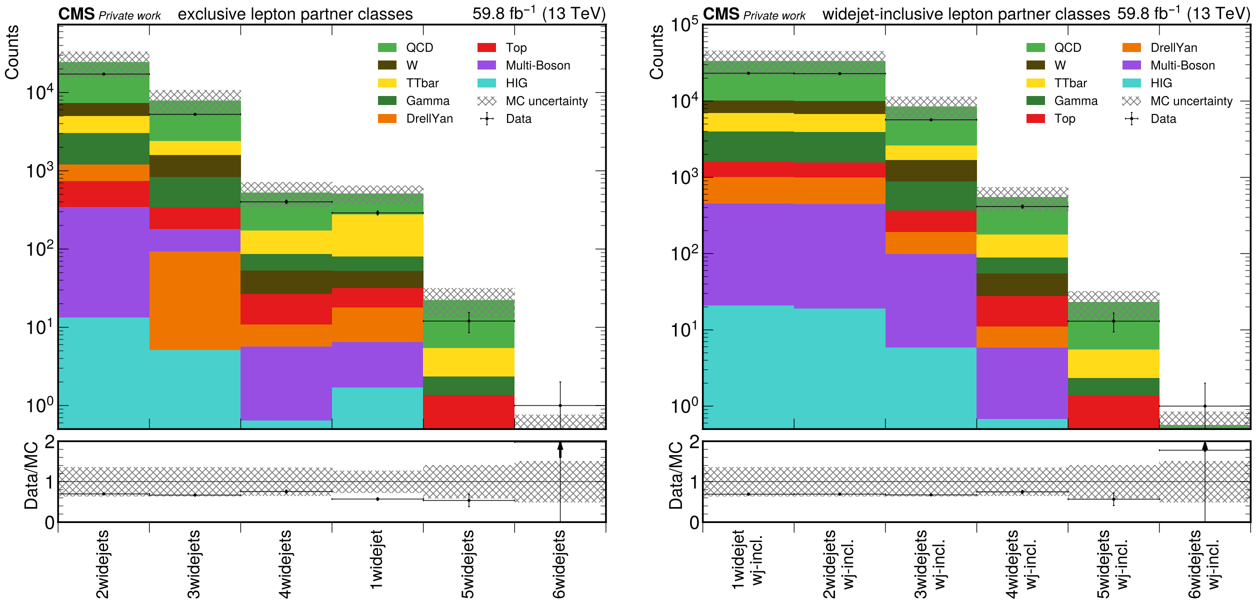
Fig. 6.2: Distributions for the 2widejets exclusive lepton partner class.

Fig. 6.2 shows two distributions for the 2widejets exclusive lepton partner class. The first observation is the decreased statistics of the lepton partner class. The event count is reduced by about two orders of magnitude. Since the QCD cross section error is not applied, the combined uncertainty is smaller as in the sections before. The shape of both S_T and m_inv seems to roughly match with the jet-only classes that were already presented (fig. 5.2). Two bins show spikes induced by the W MC samples. Similar spikes were visible in the jet-only classes in chapter 5, however there they were irrelevant because of the higher statistics of other MC processes. Most likely these spikes are related to low statistics in the generated W MC samples. QCD is still dominating the total MC background, however, with lower contributions than the $> 95\,\%$ contribution that was observed for the jet-only classes in the previous chapters. In the regions with low statistics, namely the left and right tail regions, errors are large and

[1] In terms of class names, the lepton partner classes can be put as $n_1$widejets+n_2e+$n_3\mu$[+Nwidejets] (with $n_1 \geq 1, n_2 + n_3 \geq 1$). However, in the plots, they are simply referred to as $n_1$widejets [wj-incl.] lepton partner.

the data/MC plot shows fluctuating data points. In the central energy region, the ratio seems to be approximately constant, except for the few bins with spiking W contributions The ratio value in this constant region is about $0.6 - 0.7$, which is larger than for the jet-only classes. Note however that the QCD contribution is lower and if it is assumed that the QCD samples mainly induce the offset, this observation would be plausible.

Similar observations can be made with the 3widejets exclusive lepton partner class, which is not presented here. For higher multiplicities, the statistics decrease even more, therefore, due to large statistical errors, no decent observations are possible.



(a) Exclusive classes

(b) Widejet-inclusive classes

Fig. 6.3: Comparison of the integrated event counts for the lepton partner classes.

The integrated class event counts for the lepton partner classes are presented in fig. 6.3. The decreased statistics as well as the QCD domination are easily visible. The higher data/MC ratio of $\approx 0.6 - 0.7$ compared to the jet-only classes, already observed for the 2widejets exclusive class, can also be seen. For the 6widejets class, the data/MC ratio is much higher. However, in this class, only one event is found in the data. Note that no 7widejets exclusive class exists, although this was the case for the jet-only classes (fig. 5.3)

Because of low statistics in the higher multiplicity lepton partner classes, only a selection from these classes is used to calculate the normalization factor. The selected classes are the 1widejet widejet-inclusive, 2widejets exclusive and widejet-inclusive, and the 3widejets exclusive and widejet-inclusive lepton partner classes. These classes have an integrated event count of more than 10^3, as can be seen in fig. 6.3.

6.2.2 Selection of distribution

Because the deviation between data and MC is assumed as mostly constant MC excess over all distributions (see discussion in sec. 4.11), it was decided to only calculate the normalization from a single distribution per class. Ideally, the selected distribution is used in MUSiC anyway since this thesis aims at extending MUSiC classes, and adding another distribution to MUSiC would be a larger change. Therefore, only the invariant mass m_{inv} distributions are used for this purpose.

The obtained normalization factor is generalized and then used for all distributions of the class. With the assumption of a distribution-independent MC excess, this procedure is compatible.

6.2.3 Normalization interval

The normalization factor should not be calculated from the integrated counts of the whole lepton partner class but only from the integrated counts of an interval. This is done to exclude regions where the uncertainty is large. Additionally, the high-energy region should be excluded since a hypothetical BSM signal is usually expected in this region, see also the discussion in sec. 4.11.

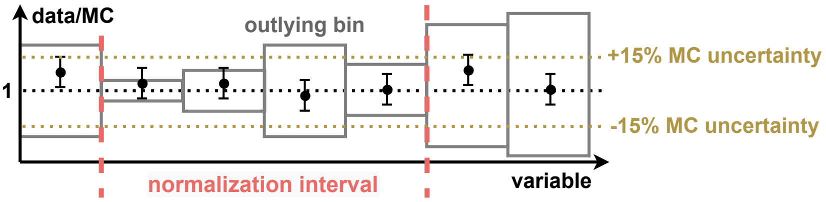


Fig. 6.4: Example interval for the proposed normalization interval finder. The black dots are the data points and the gray areas are the MC uncertainties. As can be seen, the interval starts at the first bin with a relative uncertainty $< 15\,\%$. Bins with a larger uncertainty followed by a bin with an uncertainty $< 15\,\%$ are treated as outliers and included in the interval. The interval length is greater than three bins.

Since this analysis tries to explore a possible extension to MUSiC, it should be model-unspecific. Therefore, an algorithmic approach to finding the normalization interval in the lepton partner class is preferred over a manual selection. The described interval finding approach is presented in fig. 6.4. First, the left edge of the interval should be found. The algorithm selects the first bin from the left that has a relative total MC uncertainty of less than 15 %. The minimal length of the normalization interval is set to three bins to avoid the selection of single bin intervals. To find the right edge of the interval, the algorithm then scans the bin per bin starting at the selected left edge. The first bin which shows a relative total MC uncertainty above the same threshold of 15 % is selected as a candidate for the right edge. To make the algorithm more stable against single outlier bins, it is first confirmed that also the next bin shows an uncertainty of more than 15 % before the bin is selected as the right edge of the interval. If this is not the case, the bin with the higher uncertainty is included in the interval as an outlier and the search for the right edge continues.

6.2.4 Calculating the normalization

In the following, the calculation of the normalization factor should be described. It is assumed that only the QCD MC contributions have a constant offset and all non-QCD contributions are correct, as discussed above. With this assumption, the normalization factor for the QCD contribution to the MC background can be calculated as:

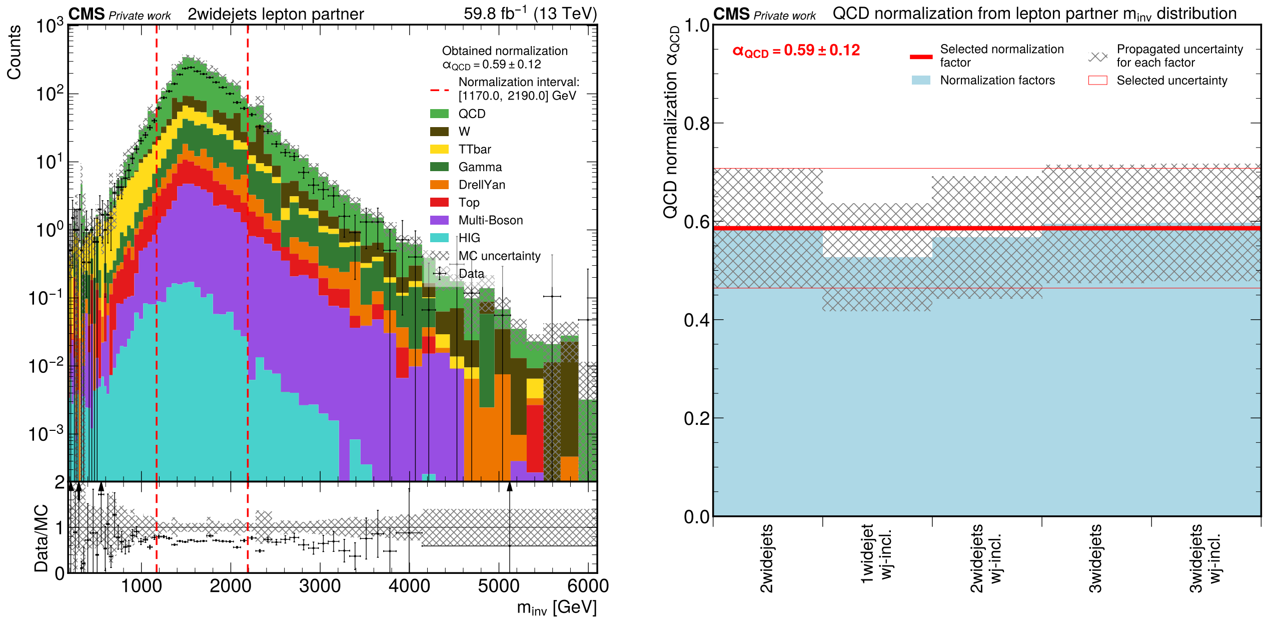
$$\alpha_{\text{QCD}} = \frac{N_{\text{data}} - N_{\text{non-QCD}}}{N_{\text{QCD}}}, \tag{6.1}$$

where N_{data} refers to the integrated data count in the control region (lepton partner class) in the selected normalization interval, N_{QCD} to the integrated QCD MC count, and $N_{\text{non-QCD}}$ to the integrated count of the MC without the QCD samples. As already stated, the fraction of QCD on the total MC for the lepton partner classes might not be the same as for the jet-only classes, however, when assuming that QCD events show a constant offset factor, this normalization factor can still be extracted.

When normalizing the QCD, the uncertainty of 50 % on the QCD LO cross sections, which was initially applied, can be regarded as an overestimation and therefore is not applied[2]. Instead, an uncertainty

[2] Note however, that there exist MC samples produced in LO in different process categories. For these samples, the cross section uncertainty is still applied.

$\sigma_{\alpha,\mathrm{QCD}}$ on the normalization factor should be introduced. There is an intrinsic uncertainty on this factor since the variables, from which α_{QCD} is calculated, have uncertainties themselves. The uncertainty contributions of the different data and MC counts in eq. 6.1 are propagated using Gaussian error propagation. It is assumed that the contributing quantities are uncorrelated. This is an assumption that might not be valid in all cases, nonetheless, to get an estimation of the uncertainty, this approach is used.



(a) Obtained normalization interval for 2jets exclusive lepton partner class in m_{inv} distribution.

(b) Comparison of the obtained normalization factors from the lepton partner classes.

Fig. 6.5: Selected results of the normalization factor calculation with the lepton partner classes.

The results for the normalization interval scan and the calculation of the normalization factors are presented in fig. 6.5. The normalization interval of the 2jets exclusive lepton partner class (fig. 6.5a) is presented. It is marked with the two red vertical lines in the plot. Generally, the selection of the interval seems to have been successful. Not only is the selected interval including the data points with the highest statistics and where the ratio is almost flat, but it is also not located in the high mass region. The second point is important, since usually BSM signals are expected in this region, see the discussion in sec. 4.11. Another interesting feature is that the procedure for outlier skipping seems to work fine. Few bins show a larger uncertainty than the threshold, yet, the normalization interval continues behind this bin.

Fig. 6.5b shows a comparison of the obtained normalization factors for the considered classes. It can be seen that these factors are all of similar order of magnitude and are compatible with each other in their respective uncertainties. Therefore, the obtained value from the 2jets exclusive lepton partner class is selected as a global normalization factor, which is marked by the red line in the plot. Its value is:

$$\alpha_{\mathrm{QCD}} \equiv \alpha_{\mathrm{QCD}}(\text{2jets excl.}) = 0.59 \pm 0.12. \tag{6.2}$$

As explained, the other normalization factors are compatible with this selection. It is remarkable to see that all normalization factors have only slightly deviating values.

6.2.5 Applying the normalization

The obtained normalization factor from the control region is then applied to the signal region, namely the jet-only class distributions. Therefore the QCD contribution of the MC in the jet-only classes should be rescaled with this factor:

$$N'_{\mathrm{QCD}} = \alpha_{\mathrm{QCD}} \cdot N_{\mathrm{QCD}}. \tag{6.3}$$

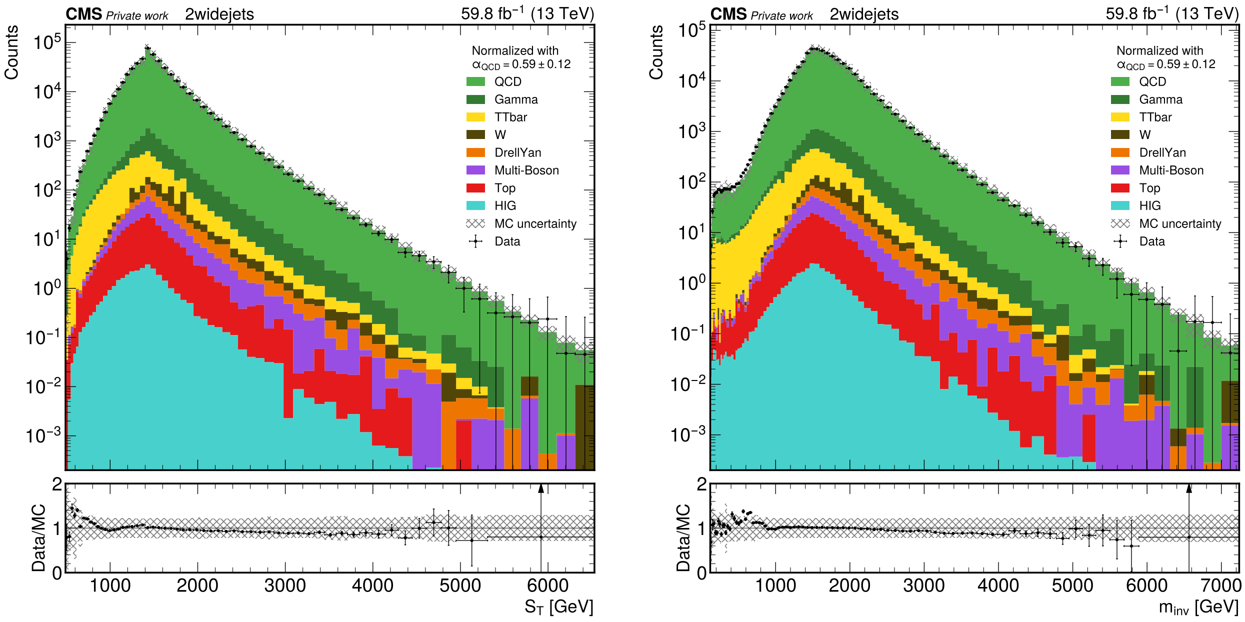
Here, N'_{QCD} describes the new, rescaled QCD counts and N_{QCD} the QCD counts without normalization. Note that, although the normalization was obtained from the invariant mass distribution only, the factor is applied to all distributions, as explained above in sec. 6.2.2.
The uncertainty related to the normalization factor has to be applied to the normalized jet-only classes. It is defined as:

$$\sigma_{\mathrm{norm}} = \sigma_{\alpha,\mathrm{QCD}} \cdot N_{\mathrm{QCD}}, \tag{6.4}$$

where $\sigma_{\alpha,\mathrm{QCD}}$ is the uncertainty on the normalization factor α_{QCD} estimated with the procedure explained above and N_{QCD} is the QCD count before normalization. Note that as for all other uncertainties described in sec. 4.6, this uncertainty is applied bin- and sample-wise and assumed uncorrelated to all other systematics. Except for the cross section uncertainty on the QCD samples, all other systematics are still applied as described in sec. 4.6.

6.3 Results

The obtained normalization factor α_{QCD} is applied to the jet-only classes after the wide jet algorithm was employed (which were presented in sec. 5.4 before the normalization). Fig. 6.6 presents two distributions for the 2widejets exclusive class with the normalized QCD contribution according to eq. 6.3.

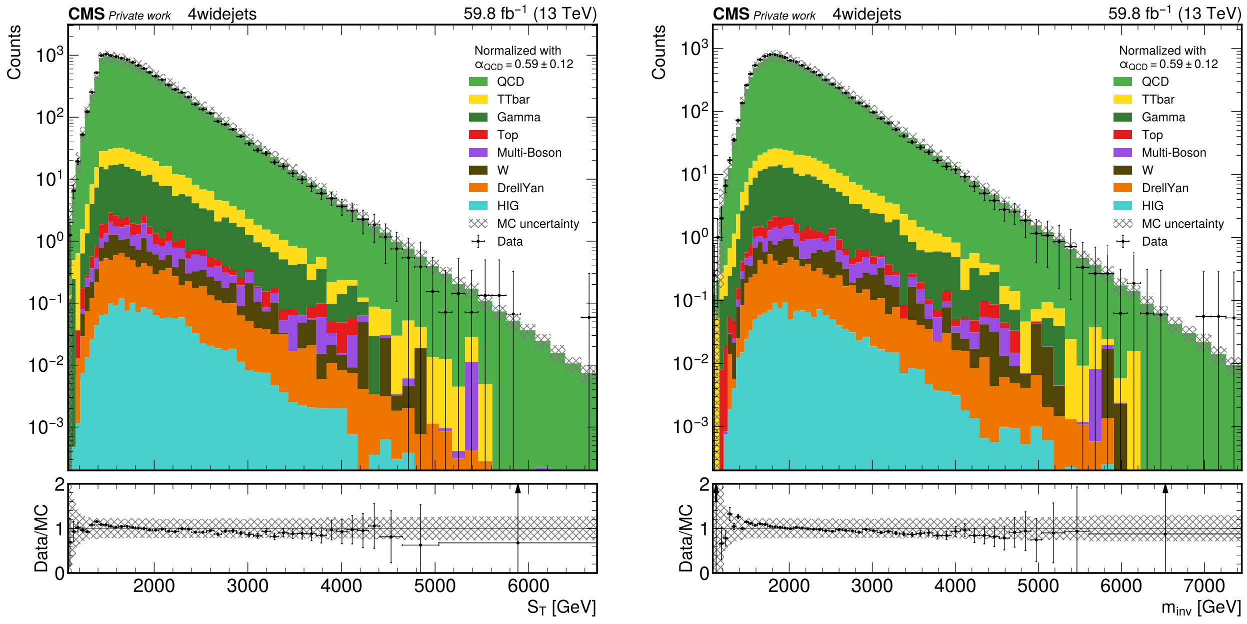


(a) Transverse momenta sum S_{T}

(b) Invariant mass m_{inv}

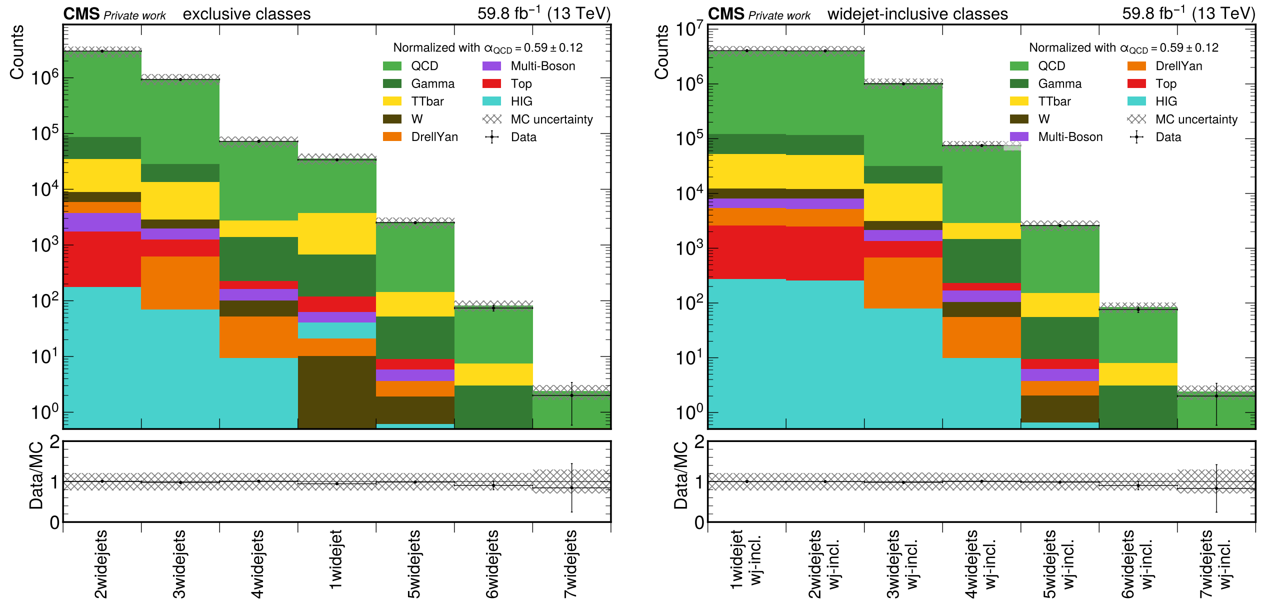
Fig. 6.6: Distributions for the 2jets exclusive class with applied normalization.

Apparently, the obtained normalization factor from the lepton partner classes is in fact able to correct the large constant offset. The data/MC ratio plot shows that the ratio now fluctuates around 1 and the constant offset is not present anymore. A decrease of the ratio with increasing energy is still present, however, this dependency is much smaller than before the wide jet approach was applied, see the discussion in sec. 5.7. Still, the decreasing ratio characteristic did not completely vanish, as well as the fluctuations in the low energy region. Especially the low energy region of the S_{T} distribution (fig. 6.6a) shows a MC excess. Apart from this, it stands out that the uncertainty bars have decreased in size. This is a result of not applying the 50 % uncertainty on the QCD LO samples, but only using the normalization factor uncertainty, which is much smaller compared to this (order of 10 %, see eq. 6.2). The corresponding 2jets exclusive class is presented in the appendix in fig. D.1. Similar characteristics regarding the corrected offset can be seen.



(a) Transverse momenta sum S_T **(b)** Invariant mass m_inv

Fig. 6.7: Distributions for the 4jets exclusive class with applied normalization.



(a) Exclusive classes **(b)** Widejet-inclusive classes

Fig. 6.8: Comparison of the integrated event counts for the wide jet classes with applied normalization.

Fig. 6.7 shows the normalized distributions of the 4jets exclusive class. For this class, the normalization also leads to decent agreement between data and MC. This is remarkable since the normalization factor used was calculated from the 2jets exclusive lepton partner class, as it was described above in sec. 6.2.4. The slightly decreasing characteristic of the ratio for increasing energies is still visible, however, the offset has been corrected successfully.

An overview of the integrated class counts for the jet-only classes after applying the normalization is presented in fig. 6.8. Remarkably, the constant offset of data and MC could be corrected with a single normalization factor that is applied to all classes. Now, the data points in the ratio plot lie around the expected value of 1 and only fluctuate within the error bars. The overall decrease in the relative error because of not applying the QCD cross section uncertainty is also visible in the plot. These observations can be made for both the normalized exclusive and widejet-inclusive classes.

6.4 Discussion

It could be shown, that rescaling the QCD counts with a constant factor allows to correct the constant fraction of the deviation between data and MC. Generally, it was of course expected, that normalizing every distribution separately in fact leads to better agreement of two sets of data. However, it is remarkable, that it could be shown that using the same global normalization factor for the QCD contribution for all classes and distributions leads to good agreement for all event classes, which generally have different fractions of QCD in their MC background.

The potential of signal biasing by a normalization like this should also be addressed again. Generally, the potential of signal biasing is very low, since a constant rescaling factor is applied to all events of the QCD background, no matter what kinematic variable or event class. However, if the normalization factor should be obtained from a region, where a BSM signal is potentially present, it might have a value that would be too high or too low. Therefore, the most severe consequence of this would be a wrong normalization factor, which would lead to a constant data/MC offset. Note that the signal shape is not affected by this, since applying a constant factor to a signal does not change it. Strictly speaking, since the normalization factor is only applied to the QCD samples of the MC background, a more complex signal bias is possible, should the QCD contribution vary strongly over the analyzed energy range. In this case, the normalization of the QCD would not lead to a constant decrease of the MC background. Nonetheless, this risk is small, since the QCD samples dominate the MC background for the analyzed jet-only classes and the strength of this potential signal bias is most likely below the signal sensitivity given by the systematic errors. If the presented approach should be used in MUSiC in the future, a signal study should be performed to address this potential issue. Unfortunately, this thesis was not able to accomplish this due to time constraints as well as space limitations for this document.

7 Conclusion and Outlook

This thesis conducted the first experiments on jet triggers in MUSiC since 2015. The performed work was focused on jet-only event classes. For the analysis, the existing MUSiC "validation" framework was extended and new plotting tools were implemented.

Different QCD datasets and jet triggers were tested and the resulting jet-only classes were analyzed. QCD processes were found to dominate all jet-only classes. The shapes of the different distributions were understood and seemed to roughly match with previously conducted work on this topic [68]. An exotic 17 jet event was found during the analysis of the data with the jet trigger, underlining the possibility to find exotic physics phenomena by including jet triggers in MUSiC. However, significant deviations between data and MC were found in the order of 50 %. The deviation was found to have a large constant component and two relevant dependencies from energy-like quantities and the jet multiplicity. The analysis in this thesis was continued with the p_T-binned QCD sample set and the H_T trigger.

A wide jet merging strategy as well as a differential pseudorapidity cut, both frequently used in dedicated CMS jet analyses, were adapted to a generalized approach to handle arbitrary jet multiplicities and therefore event classes. The wide jet merging and pseudorapidity cut proved to decrease the energy and jet multiplicity dependencies significantly.

To address the remaining roughly constant offset between data and MC, a normalization strategy was proposed. Since QCD samples dominate the event classes, the assumption was made that the constant data/MC offset originates from these samples. Therefore, normalizing the QCD contribution of MC was desired. The region from which the normalization should be calculated was chosen as a disjoint set from the jet-only classes. This was realized by inverting the lepton veto as a control region while keeping all other analysis steps constant. It was decided to use only a region of the invariant mass distribution to obtain the normalization factor, where the event statistics are sufficient. The normalization factor for the 2jets exclusive lepton partner class was generalized to be used for all classes, its value was found to be $\alpha_\mathrm{QCD} = 0.59 \pm 0.12$. All other obtained normalization factors were found to agree with this value in the given uncertainties. Applying the normalization to the jet-only classes was found to successfully correct the constant data/MC offset, a remarkable observation considering the fact that one constant factor was applied to all classes. The relative uncertainties could also be decreased since the LO cross section uncertainty was not applied to QCD, but instead, a normalization error in the order of 10 %.

However, many open questions remain. The additional analysis steps proposed and performed in this thesis definitely showed to improve the data/MC agreement. Deviations of data and MC are still visible, especially in the high and low energy regions, which are not understood. Similar deviations in these regions were also observed in previous work [68].

The whole analysis in this thesis was based on the assumption, that the significant deviation was not created by new physics, which was mainly motivated by the fact that the observed deviation seems to have a significant constant component over the whole phase space. It was speculated that faulty QCD cross sections or other MC weighting problems might be the reason. Yet, the origin of the observed deviation in this thesis could not be found apart from these hypotheses. Future work should be invested in understanding this phenomenon, and with it, the legitimateness of the conducted efforts to achieve better agreement between data and MC. The analysis of the observed deviation could potentially profit from a reassessment of the QCD sample cross sections in the future. This was not possible in the working time of this thesis after all.

If some of the proposed additional analysis steps in this thesis should be considered to counter the strong deviation in the future, it is recommended to conduct a separate signal bias study on these analysis steps. It was argued that the signal bias implied by the proposed analysis steps would be low, referring to dedicated analyses and the claim that the normalization of the dominating MC process with a constant factor does not imply significant bias. Because of time constraints, no full signal bias study could be conducted in this thesis.

Apart from all the remaining challenges, this analysis showed that, in principle, it would be possible to include jet triggers and jet-only classes into MUSiC, which would extend the analyzed phase space region of the MUSiC analysis. However, it was uncovered that additional preprocessing is necessary. If jet-only classes should be included in the MUSiC analysis in the future, this thesis can serve as a basis for these efforts.

8 Bibliography

[1] E. Tiesinga et al. "CODATA recommended values of the fundamental physical constants: 2018". In: *Rev. Mod. Phys.* 93 (2 June 2021), p. 025010. DOI: 10.1103/RevModPhys.93.025010 (Cited on page viii).

[2] J. D. Wells. *Discovery Beyond the Standard Model of Elementary Particle Physics.* 1. Springer Cham, 2020. ISBN: 9783030382049. DOI: https://doi.org/10.1007/978-3-030-38204-9 (Cited on page 1).

[3] CERN. *The Standard Model.* URL: https://home.cern/science/physics/standard-model. Accessed on 08/07/2023. (Cited on pages 1, 2).

[4] D. Duchardt. "MUSiC: A Model Unspecific Search for New Physics Based on CMS Data at $\sqrt{s}$ = 8 TeV". Doctoral Thesis. RWTH Aachen University, 2017 (Cited on pages 1, 5, 8–10, 12, 13, 15, 16, 18, 19).

[5] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. DOI: https://doi.org/10.1016%2Fj.physletb.2012.08.020 (Cited on pages 1, 3, 6).

[6] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.021 (Cited on pages 1, 3, 6).

[7] J. T. Roemer. "Search for Excited Leptons with pp Collisions at $\sqrt{s} = 13$ TeV with the CMS Experiment at the LHC". Doctoral Thesis. RWTH Aachen University, 2020 (Cited on page 1).

[8] Wikimedia Commons. *The Standard Model of Elementary Particles.* URL: https://upload.wikimedia.org/wikipedia/common/0/00/Standard_Model_of_Elementary_Particles.svg. Last accessed on 08/07/2023. (Cited on page 1).

[9] S. P. Martin and J. D. Wells. *Elementary Particles and Their Interactions.* 1. Springer Cham, 2022. DOI: https://doi.org/10.1007/978-3-031-14368-7 (Cited on pages 1–4).

[10] Y. Kaiser. "Sensitivity study of the MUSIC algorithm for LHC run 2". Bachelor Thesis. RWTH Aachen University, 2021 (Cited on pages 2, 9, 13, 14, 18, 19).

[11] J. Lieb. "Discovery Potential of a Model Independent Search for New Physics at the LHC". Master Thesis. RWTH Aachen University, 2017 (Cited on pages 2–5).

[12] S. Westerdale. *Neutrino Mass Problem: Masses and Oscillations.* 2010. URL: http://web.mit.edu/shawest/Public/8.06/termPaperDraft.pdf (Cited on pages 2, 5).

[13] M. K. Gaillard, P. D. Grannis, and F. J. Sciulli. "The standard model of particle physics". In: *Reviews of Modern Physics* 71.2 (Mar. 1999), 96–S111. DOI: 10.1103/RevModPhys.71.S96 (Cited on pages 2–4).

[14] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (9 Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321 (Cited on page 3).

[15] P. W. Higgs. "Broken symmetries, massless particles and gauge fields". In: *Physics Letters* 12.2 (1964), pp. 132–133. ISSN: 0031-9163. DOI: `https://doi.org/10.1016/0031-9163(64)91136-9` (Cited on page 3).

[16] P. W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: `10.1103/PhysRevLett.13.508` (Cited on page 3).

[17] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". In: *Phys. Rev. Lett.* 13 (20 Nov. 1964), pp. 585–587. DOI: `10.1103/PhysRevLett.13.585` (Cited on page 3).

[18] P. W. Higgs. "Spontaneous Symmetry Breakdown without Massless Bosons". In: *Phys. Rev.* 145 (4 May 1966), pp. 1156–1163. DOI: `10.1103/PhysRev.145.1156` (Cited on page 3).

[19] T. W. B. Kibble. "Symmetry Breaking in Non-Abelian Gauge Theories". In: *Phys. Rev.* 155 (5 Mar. 1967), pp. 1554–1561. DOI: `10.1103/PhysRev.155.1554` (Cited on page 3).

[20] F. Rehbein. "Search for sphalerons in the e μ final state at $\sqrt{s} = 13$ TeV with the CMS experiment". Bachelor Thesis. RWTH Aachen University, 2018 (Cited on page 3).

[21] R. Mann. *An Introduction to Particle Physics and the Standard Model*. Taylor and Francis, 2010. ISBN: 9780429141225. DOI: `http://doi.org/10.1201/9781420083002` (Cited on pages 3–5).

[22] G. Altarelli. *Collider Physics within the Standard Model: a Primer*. 2013. DOI: `https://doi.org/10.48550/arXiv.1303.2842` (Cited on page 4).

[23] K. Rabbertz. *Jet Physics at the LHC The Strong Force beyond the TeV Scale*. eng. Springer Tracts in Modern Physics 268. Cham: Springer International Publishing, 2017. ISBN: 9783319421155. DOI: `https://doi.org/10.1007/978-3-319-42115-5` (Cited on pages 4, 11, 15).

[24] E. M. Metodiev. *Data ex Machina: Machine Learning with Jets in CMS Open Data*. 2020. URL: `https://indico.cern.ch/event/809820/contributions/3632591/attachments/1970777/3278357/MLforJets2020_Metodiev.pdf`. Accessed on 19/07/2023. (Cited on page 4).

[25] J. Ellis. "The discovery of the gluon". In: *International Journal of Modern Physics A* 29.31 (Dec. 2014), p. 1430072. DOI: `10.1142/s0217751x14300725` (Cited on pages 4, 5).

[26] DESY. *The discovery of the gluon - a journey back in time to the 70s*. URL: `https://www.desy.de/news/backgrounders/40_years_of_gluon/index_eng.html`. Accessed on 10/07/2023. (Cited on page 5).

[27] D. Brunner. "Suche nach zusätzlichen Raumdimensionen im Massenspektrum von zwei Elektronen bei $\sqrt{s} = 13$ TeV". Bachelor Thesis. RWTH Aachen University, 2017 (Cited on pages 5, 6, 9).

[28] CERN. *LHC, The Guide - CERN-Brochure-2021-004-Eng*. URL: `https://cds.cern.ch/record/2809109/files/CERN-Brochure-2021-004-Eng.pdf`. Accessed on 10/07/2023. (Cited on page 6).

[29] CERN. *Facts and figures about the LHC*. URL: `https://home.cern/resources/faqs/facts-and-figures-about-lhc`. Accessed on 10/07/2023. (Cited on page 6).

[30] S. Myers. "The Large Hadron Collider 2008-2013". In: *Int. J. Mod. Phys. A* 28 (2013), p. 1330035. DOI: `10.1142/S0217751X13300354` (Cited on page 6).

[31] X. C. Vidal and R. C. Manzano. *LHC running - Taking a closer look at LHC*. URL: `https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.lhc_running`. Accessed on 10/07/2023. (Cited on page 6).

[32] A. Andrade. "Machine Learning for a Model Unspecific Search in CMS". Master Thesis. RWTH Aachen University, 2023 (Cited on pages 6–10, 13, 14, 22).

[33] CERN. *The Large Hadron Collider - Timeline.* URL: `https://timeline.web.cern.ch/timeline-header/93`. Accessed on 10/07/2023. (Cited on page 6).

[34] E. Lopienska. *The CERN accelerator complex, layout in 2022.* URL: `https://cds.cern.ch/images/CERN-GRAPHICS-2022-001-1`. Accessed on 11/07/2023. (Cited on page 7).

[35] CERN. *Linear Accelerator 2.* URL: `https://home.cern/science/accelerators/linear-accelerator-2`. Accessed on 10/07/2023. (Cited on page 6).

[36] CERN. *The Proton Synchrotron Booster.* URL: `https://home.cern/science/accelerators/proton-synchrotron-booster`. Accessed on 10/07/2023. (Cited on page 6).

[37] CERN. *The Proton Synchrotron.* URL: `https://home.cern/science/accelerators/proton-synchrotron`. Accessed on 10/07/2023. (Cited on page 6).

[38] CERN. *The Super Proton Synchrotron.* URL: `https://home.cern/science/accelerators/super-proton-synchrotron`. Accessed on 10/07/2023. (Cited on page 6).

[39] B. Salvachua. "Overview of Proton-Proton Physics during Run 2". In: *9th LHC Operations Evian Workshop.* 2019, pp. 7–14. URL: `https://cds.cern.ch/record/2750272` (Cited on page 7).

[40] CERN. *Experiments - LHC experiments.* URL: `https://home.cern/science/experiments`. Accessed on 10/07/2023. (Cited on page 7).

[41] CMS Experiment. *Detector.* URL: `https://cms.cern/detector`. Accessed on 11/07/2023. (Cited on page 8).

[42] CMS Collaboration. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08004. DOI: `10.1088/1748-0221/3/08/S08004` (Cited on pages 9, 10).

[43] CMS Collaboration. "The CMS RPC detector performance and stability during LHC RUN-2". In: *Journal of Instrumentation* 14.11 (Nov. 2019), pp. C11012–C11012. DOI: `10.1088/1748-0221/14/11/c11012` (Cited on page 10).

[44] CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12.10 (2017), P10003. DOI: `10.1088/1748-0221/12/10/P10003` (Cited on page 11).

[45] M. Dordevic. "The CMS Particle Flow Algorithm". In: *EPJ Web Conf.* 191 (2018), p. 02016. DOI: `10.1051/epjconf/201819102016` (Cited on pages 10, 11).

[46] A. Perrotta. *CMS event reconstruction status in Run 2.* Tech. rep. Geneva: CERN, 2018. URL: `https://cds.cern.ch/record/2644443` (Cited on page 10).

[47] G. P. Salam and G. Soyez. "A practical seedless infrared-safe cone jet algorithm". In: *Journal of High Energy Physics* 2007.05 (May 2007), pp. 086–086. DOI: `10.1088/1126-6708/2007/05/086` (Cited on page 11).

[48] CMS Collaboration. *A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging.* Tech. rep. Geneva: CERN, 2009. URL: `https://cds.cern.ch/record/1194489` (Cited on page 11).

[49] S. Catani et al. "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions". In: *Nucl. Phys. B* 406 (1993), pp. 187–224. DOI: `10.1016/0550-3213(93)90166-M` (Cited on page 11).

[50] M. Cacciari, G. P. Salam, and G. Soyez. "The anti-k$_T$ jet clustering algorithm". In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), p. 63. DOI: `10.1088/1126-6708/2008/04/063` (Cited on pages 11, 12).

[51] P. Schieferdecker. *Jet Algorithms.* 2009. URL: https://twiki.cern.ch/twiki/bin/viewfile/Sandbox/Lecture?rev=1;filename=Philipp_Schieferdeckers_Lecture.pdf (Cited on page 11).

[52] S. Catani and Zeppenfeld D. *Jet Algorithms.* 2019. URL: https://s3.cern.ch/inspire-prod-files-6/6904a3576c84c5d1f05a1f171cac3695 (Cited on page 11).

[53] CMS Collaboration. "Search for high mass dijet resonances with a new background prediction method in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2020.5 (May 2020). DOI: 10.1007/jhep05(2020)033 (Cited on pages 13, 21, 23, 32, 34, 39, 40, 42).

[54] CMS Collaboration. *Search for narrow trijet resonances in proton-proton collisions at $\sqrt{s} =$ 13 TeV.* Tech. rep. Geneva: CERN, 2023. URL: https://cds.cern.ch/record/2859386. PAS accepted, but paper not published yet as of 07/08/2023. (Cited on pages 13, 21, 23, 32, 34, 35, 39).

[55] CMS Collaboration. "Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s}$=13 TeV". In: *Journal of High Energy Physics* 2018.11 (Nov. 2018). DOI: 10.1007/jhep11(2018)185 (Cited on page 13).

[56] T. Hebbeker. *A Global Comparison between L3 Data and Standard Model Monte Carlo - a first attempt.* July 1998 (Cited on page 13).

[57] CMS collaboration. *MUSIC – An Automated Scan for Deviations between Data and Monte Carlo Simulation.* Tech. rep. Geneva: CERN, 2008. URL: http://cds.cern.ch/record/1152572 (Cited on page 13).

[58] CMS collaboration. *MUSiC, a Model Unspecific Search for New Physics, in pp Collisions at $\sqrt{s} = 8$ TeV.* Tech. rep. Geneva: CERN, 2017. URL: http://cds.cern.ch/record/2256653 (Cited on page 13).

[59] CMS Collaboration. "MUSiC: a model-unspecific search for new physics in proton–proton collisions at $\sqrt{s} = 13$ TeV". In: *Eur. Phys. J. C* 81.7 (2021), p. 629. DOI: 10.1140/epjc/s10052-021-09236-z (Cited on pages 13, 16–20, 22, 24–26).

[60] E. Bols et al. "Jet flavour classification using DeepJet". In: *Journal of Instrumentation* 15.12 (Dec. 2020), P12012. DOI: 10.1088/1748-0221/15/12/P12012 (Cited on page 14).

[61] CMS TWiki. *Public CMS Luminosity Information - 2018 proton-proton collisions at 13 TeV.* 2023. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults#2018_proton_proton_collisions_at. Accessed on 16/07/2023. (Cited on page 15).

[62] CMS Collaboration. "Pileup mitigation at CMS in 13 TeV data". In: *Journal of Instrumentation* 15.09 (Sept. 2020), P09018–P09018. DOI: 10.1088/1748-0221/15/09/p09018 (Cited on page 15).

[63] CMS Collaboration. *Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV.* 2021. URL: https://cds.cern.ch/record/2792322 (Cited on pages 15, 25).

[64] CMS TWiki. *Jet Energy Resolution.* 2023. URL: https://twiki.cern.ch/twiki/bin/view/CMS/JetResolution. Accessed on 16/07/2023. (Cited on pages 15, 25).

[65] NNPDF Collaboration. "Parton distributions from high-precision collider data". In: *The European Physical Journal C* 77.10 (Oct. 2017). DOI: 10.1140/epjc/s10052-017-5199-5 (Cited on page 15).

[66] G. Lungu. *Missing Transverse Energy (MET) at CMS.* Jan. 2009. URL: https://indico.cern.ch/event/46651/contributions/1143011/attachments/950879/1349271/Lungu-Jterm3_MET_011309.pdf (Cited on page 17).

[67] L. Lyons. “Open statistical issues in Particle Physics”. In: *The Annals of Applied Statistics* 2.3 (2008), pp. 887–915. DOI: 10.1214/08-AOAS163 (Cited on page 18).

[68] A. A. E. Albert. “Extension of the Model Unspecific Search in CMS to Final States with Jets using 2012 Data”. Master Thesis. RWTH Aachen University, 2018 (Cited on pages 20, 21, 32, 49, 63).

[69] CMS Collaboration. “Search for dijet resonances in proton–proton collisions at $\sqrt{s} = 13$ TeV and constraints on dark matter and other models”. In: *Physics Letters B* 769 (June 2017), pp. 520–542. DOI: 10.1016/j.physletb.2017.02.012 (Cited on pages 21, 34, 39, 40).

[70] T. Sjöstrand, S. Mrenna, and P. Skands. “PYTHIA 6.4 physics and manual”. In: *Journal of High Energy Physics* 2006.05 (May 2006), pp. 026–026. DOI: 10.1088/1126-6708/2006/05/026 (Cited on page 21).

[71] T. Sjöstrand et al. “An introduction to PYTHIA 8.2”. In: *Computer Physics Communications* 191 (June 2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024 (Cited on page 22).

[72] J. Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. In: *Journal of High Energy Physics* 2014.7 (July 2014). DOI: 10.1007/jhep07(2014)079 (Cited on page 22).

[73] P. Nason. “A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms”. In: *Journal of High Energy Physics* 2004.11 (Nov. 2004), pp. 040–040. DOI: 10.1088/1126-6708/2004/11/040 (Cited on page 22).

[74] J. M. Campbell et al. “Top-pair production and decay at NLO matched with parton showers”. In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). DOI: 10.1007/jhep04(2015)114 (Cited on page 22).

[75] E. Bagnaschi et al. “Higgs production via gluon fusion in the POWHEG approach in the SM and in the MSSM”. In: *Journal of High Energy Physics* 2012.2 (Feb. 2012). DOI: 10.1007/jhep02(2012)088 (Cited on page 22).

[76] S. Frixione, P. Nason, and C. Oleari. “Matching NLO QCD computations with parton shower simulations: the POWHEG method”. In: *Journal of High Energy Physics* 2007.11 (Nov. 2007), pp. 070–070. DOI: 10.1088/1126-6708/2007/11/070 (Cited on page 22).

[77] S. Alioli et al. “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”. In: *Journal of High Energy Physics* 2010.6 (June 2010). DOI: 10.1007/jhep06(2010)043 (Cited on page 22).

[78] S. Alioli et al. “NLO vector-boson production matched with shower in POWHEG”. In: *Journal of High Energy Physics* 2008.07 (July 2008), pp. 060–060. DOI: 10.1088/1126-6708/2008/07/060 (Cited on page 22).

[79] E. Re. “Single-top Wt-channel production matched with parton showers using the POWHEG method”. In: *The European Physical Journal C* 71.2 (Feb. 2011). DOI: 10.1140/epjc/s10052-011-1547-z (Cited on page 22).

[80] S. Alioli et al. “NLO single-top production matched with shower in POWHEG: s- and t-channel contributions”. In: *Journal of High Energy Physics* 2009.09 (Sept. 2009), p. 111. DOI: 10.1088/1126-6708/2009/09/111 (Cited on page 22).

[81] S. Alioli et al. “NLO Higgs boson production via gluon fusion matched with shower in POWHEG”. In: *Journal of High Energy Physics* 2009.04 (Apr. 2009), pp. 002–002. DOI: 10.1088/1126-6708/2009/04/002 (Cited on page 22).

[82] P. Nason and C. Oleari. “NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG”. In: *Journal of High Energy Physics* 2010.2 (Feb. 2010). DOI: `10.1007/jhep02(2010)037` (Cited on page 22).

[83] T. Melia et al. “W^+W^-, WZ and ZZ production in the POWHEG BOX”. In: *Journal of High Energy Physics* 2011.11 (Nov. 2011). DOI: `10.1007/jhep11(2011)078` (Cited on page 22).

[84] P. Nason and G. Zanderighi. “W^+W^-, WZ and ZZ production in the POWHEG-BOX-V2”. In: *The European Physical Journal C* 74.1 (Jan. 2014). DOI: `10.1140/epjc/s10052-013-2702-5` (Cited on page 22).

[85] E. Bothmann et al. “Event generation with Sherpa 2.2”. In: *SciPost Physics* 7.3 (Sept. 2019). DOI: `10.21468/scipostphys.7.3.034` (Cited on page 22).

[86] CMS Collaboration. *Performance of JetMET high level trigger algorithms in the CMS experiment using proton-proton collisions data at $\sqrt{s}$ = 13 TeV during Run-2.* 2022. URL: `http://cds.cern.ch/record/2842377` (Cited on page 23).

[87] CMS TWiki. *HLT Paths RunII List.* 2019. URL: `https://twiki.cern.ch/twiki/bin/view/CMS/HLTPathsRunIIList#2018`. Accessed on 18/07/2023. (Cited on page 23).

[88] CMS Collaboration. “Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV”. In: *Journal of High Energy Physics* 2020.12 (Dec. 2020). DOI: `10.1007/jhep12(2020)085` (Cited on page 23).

[89] CMS Collaboration. *CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s}$ = 13 TeV.* Tech. rep. Geneva: CERN, 2019. URL: `https://cds.cern.ch/record/2676164` (Cited on pages 24, 25).

[90] J. Butterworth et al. “PDF4LHC recommendations for LHC Run II”. In: *Journal of Physics G: Nuclear and Particle Physics* 43.2 (Jan. 2016), p. 023001. DOI: `10.1088/0954-3899/43/2/023001` (Cited on page 25).

[91] CMS Collaboration. “Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s}$=13TeV”. In: *Journal of Instrumentation* 15.10 (Oct. 2020), P10017–P10017. DOI: `10.1088/1748-0221/15/10/p10017` (Cited on page 25).

[92] CMS TWiki. *CMS Exotica Summary plots for 13 TeV data.* 2023. URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SummaryPlotsEXO13TeV`. Accessed on 18/07/2023. (Cited on page 32).

[93] CMS Collaboration. “Search for resonances and quantum black holes using dijet mass spectra in proton-proton collisions at $\sqrt{s} = 8$ TeV”. In: *Physical Review D* 91.5 (Mar. 2015). DOI: `10.1103/physrevd.91.052009` (Cited on page 34).

[94] CMS Collaboration. “Search for narrow resonances using the dijet mass spectrum in pp collisions at $\sqrt{s} = 8$ TeV”. In: *Physical Review D* 87.11 (June 2013). DOI: `10.1103/physrevd.87.114015` (Cited on page 34).

[95] CMS Collaboration. “Search for narrow resonances and quantum black holes in inclusive and b-tagged dijet mass spectra from pp collisions at $\sqrt{s} = 7$ TeV”. In: *Journal of High Energy Physics* 2013.1 (Jan. 2013). DOI: `10.1007/jhep01(2013)013` (Cited on page 34).

[96] CMS Collaboration. “Search for resonances in the dijet mass spectrum from 7 TeV pp collisions at CMS”. In: *Physics Letters B* 704.3 (Oct. 2011), pp. 123–142. DOI: `10.1016/j.physletb.2011.09.015` (Cited on page 34).

Appendix

A Full list of MC samples

Process group	DAS name	Cross section σ [pb]	k-factor	Order
DrellYan	DYJetsToLL_M-10to50_TuneCP5_13TeV-amcatnloFXFX-pythia8	1.861×10^{4}	1.0	NLO
	DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8	6.104×10^{3}	0.944	NLO
	DYJetsToLL_LHEFilterPtZ-50To100_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	3.617×10^{2}	1.0	NLO
	DYJetsToLL_LHEFilterPtZ-100To250_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	8.497×10^{1}	1.0	NLO
	DYJetsToLL_LHEFilterPtZ-250To400_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	2.906	1.0	NLO
	DYJetsToLL_LHEFilterPtZ-400To650_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	4.352×10^{-1}	1.0	NLO
	DYJetsToLL_LHEFilterPtZ-650ToInf_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	5.448×10^{-2}	1.0	NLO
	DYToEE_M-120To200_TuneCP5_13TeV-powheg-pythia8	1.932×10^{1}	1.0	NLO
	DYToEE_M-200To400_TuneCP5_13TeV-powheg-pythia8	2.731	1.0	NLO
	DYToEE_M-400To800_TuneCP5_13TeV-powheg-pythia8	2.410×10^{-1}	1.0	NLO
	DYToEE_M-800To1400_TuneCP5_13TeV-powheg-pythia8	1.900×10^{-2}	1.0	NLO
	DYToEE_M-1400To2300_TuneCP5_13TeV-powheg-pythia8	1.390×10^{-3}	1.0	NLO
	DYToEE_M-2300To3500_TuneCP5_13TeV-powheg-pythia8	8.278×10^{-5}	1.0	NLO
	DYToEE_M-3500To4500_TuneCP5_13TeV-powheg-pythia8	4.135×10^{-6}	1.0	NLO
	DYToEE_M-4500To6000_TuneCP5_13TeV-powheg-pythia8	4.560×10^{-7}	1.0	NLO
	DYToEE_M-6000ToInf_TuneCP5_13TeV-powheg-pythia8	2.066×10^{-8}	1.0	NLO
	ZJetsToNuNu_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8	2.805×10^{2}	1.63	NLO
	ZJetsToNuNu_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8	7.836×10^{1}	1.62	NLO
	ZJetsToNuNu_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8	1.094×10^{1}	1.46	NLO
	ZJetsToNuNu_HT-600To800_TuneCP5_13TeV-madgraphMLM-pythia8	2.559	1.391	NLO

Process group	DAS name	Cross section σ [pb]	k-factor	Order
	ZJetsToNuNu_HT-800To1200_TuneCP5_13TeV-madgraphMLM-pythia8	1.179	1.391	NLO
	ZJetsToNuNu_HT-1200To2500_TuneCP5_13TeV-madgraphMLM-pythia8	2.883×10^{-1}	1.391	NLO
	ZJetsToNuNu_HT-2500ToInf_TuneCP5_13TeV-madgraphMLM-pythia8	6.945×10^{-3}	1.391	NLO
	ZToMuMu_M-120To200_TuneCP5_13TeV-powheg-pythia8	1.932×10^{1}	1.0	NLO
	ZToMuMu_M-200To400_TuneCP5_13TeV-powheg-pythia8	2.731	1.0	NLO
	ZToMuMu_M-400To800_TuneCP5_13TeV-powheg-pythia8	2.410×10^{-1}	1.0	NLO
	ZToMuMu_M-800To1400_TuneCP5_13TeV-powheg-pythia8	1.700×10^{-2}	1.0	NLO
	ZToMuMu_M-1400To2300_TuneCP5_13TeV-powheg-pythia8	1.390×10^{-3}	1.0	NLO
	ZToMuMu_M-2300To3500_TuneCP5_13TeV-powheg-pythia8	8.948×10^{-5}	1.0	NLO
	ZToMuMu_M-3500To4500_TuneCP5_13TeV-powheg-pythia8	4.135×10^{-6}	1.0	NLO
	ZToMuMu_M-4500To6000_TuneCP5_13TeV-powheg-pythia8	4.560×10^{-7}	1.0	NLO
	ZToMuMu_M-6000ToInf_TuneCP5_13TeV-powheg-pythia8	2.066×10^{-8}	1.0	NLO
Gamma	GJets_DR-0p4_HT-40To100_TuneCP5_13TeV-madgraphMLM-pythia8	1.575×10^{4}	1.0	LO
	GJets_DR-0p4_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8	5.001×10^{3}	1.0	LO
	GJets_DR-0p4_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8	1.154×10^{3}	1.0	LO
	GJets_DR-0p4_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8	1.272×10^{2}	1.0	LO
	GJets_DR-0p4_HT-600ToInf_TuneCP5_13TeV-madgraphMLM-pythia8	9.346×10^{1}	1.0	LO
HIG	ggZH_HToBB_ZToNuNu_M-125_TuneCP5_13TeV-powheg-pythia8	1.437×10^{-2}	1.0	NNLO
	ggZH_HToBB_ZToLL_M-125_TuneCP5_13TeV-powheg-pythia8	7.842×10^{-3}	1.0	NNLO
	ggZH_HToBB_ZToQQ_M-125_TuneCP5_13TeV-powheg-pythia8	4.996×10^{-2}	1.0	NNLO
	GluGluHToZZTo4L_M125_TuneCP5_13TeV_powheg2_JHUGenV7011_pythia8	1.212×10^{-2}	1.0	NLO
	ttHTobb_M125_TuneCP5_13TeV-powheg-pythia8	2.953×10^{-1}	1.0	N3LO
	ttHToNonbb_M125_TuneCP5_13TeV-powheg-pythia8	2.118×10^{-1}	1.0	N3LO
	VBF_HToZZTo4L_M125_TuneCP5_13TeV_powheg2_JHUGenV7011_pythia8	9.905×10^{-2}	1.0	NNLO
	VBFHToBB_M-125_TuneCP5_13TeV-powheg-pythia8	2.203	1.0	NNLO
	VBFHToGG_M125_TuneCP5_13TeV-amcatnlo-pythia8	8.585×10^{-3}	1.0	NNLO
	VBFHToTauTau_M125_TuneCP5_13TeV-powheg-pythia8	2.372×10^{-1}	1.0	NNLO
	VBFHToWWTo2L2Nu_M-125_TuneCP5_13TeV-powheg-jhugen727-pythia8	8.579×10^{-2}	1.0	NNLO
	VHToNonbb_M125_TuneCP5_13TeV-amcatnloFXFX_madspin_pythia8	9.425×10^{-1}	1.0	NNLO
	WminusH_HToBB_WToQQ_M-125_TuneCP5_13TeV-powheg-pythia8	3.675×10^{-1}	1.0	NLO

Process group	DAS name	Cross section σ [pb]	k-factor	Order
	WminusH_HToBB_WToLNu_M-125_TuneCP5_13TeV-powheg-pythia8	1.011×10^{-1}	1.0	NNLO
	WplusH_HToBB_WToLNu_M-125_TuneCP5_13TeV-powheg-pythia8	1.593×10^{-1}	1.0	NNLO
	WplusH_HToBB_WToQQ_M-125_TuneCP5_13TeV-powheg-pythia8	5.890×10^{-1}	1.0	NLO
	ZH_HToBB_ZToNuNu_M-125_TuneCP5_13TeV-powheg-pythia8	8.912×10^{-2}	1.0	NNLO
	ZH_HToBB_ZToLL_M-125_TuneCP5_13TeV-powheg-pythia8	4.865×10^{-2}	1.0	NNLO
	ZH_HToBB_ZToQQ_M-125_TuneCP5_13TeV-powheg-pythia8	3.099×10^{-1}	1.0	NNLO
Multi-Boson	DiPhotonJetsBox_M40_80-sherpa	2.993×10^{2}	1.0	LO
	DiPhotonJetsBox_MGG-80toInf_13TeV-sherpa	8.836×10^{1}	1.0	LO
	WGToLNuG_01J_5f_TuneCP5_13TeV-amcatnloFXFX-pythia8	4.120×10^{-2}	1.0	LO
	WGToLNuG_01J_5f_PtG_130_TuneCP5_13TeV-amcatnloFXFX-pythia8	1.031	1.0	NLO
	WGToLNuG_01J_5f_PtG_300_TuneCP5_13TeV-amcatnloFXFX-pythia8	7.110×10^{-3}	1.0	NLO
	WGToLNuG_01J_5f_PtG_500_TuneCP5_13TeV-amcatnloFXFX-pythia8	9.830×10^{-4}	1.0	NLO
	WWG_TuneCP5_13TeV-amcatnlo-pythia8	2.147×10^{-1}	1.0	NLO
	WWTo1L1Nu2Q_4f_TuneCP5_13TeV-amcatnloFXFX-pythia8	4.353×10^{1}	1.149	NNLO
	WWTo2L2Nu_MLL_200To600_TuneCP5_13TeV-powheg-pythia8	1.048×10^{1}	1.162	NNLO
	WWTo2L2Nu_MLL_600To1200_TuneCP5_13TeV-powheg-pythia8	1.048×10^{1}	1.162	NNLO
	WWTo2L2Nu_MLL_1200To2500_TuneCP5_13TeV-powheg-pythia8	1.048×10^{1}	1.162	NNLO
	WWTo4Q_4f_TuneCP5_13TeV-amcatnloFXFX-pythia8	4.531×10^{1}	1.0	NLO
	WWW_4F_TuneCP5_13TeV-amcatnlo-pythia8	2.086×10^{-1}	1.0	NLO
	WWZ_4F_TuneCP5_13TeV-amcatnlo-pythia8	1.651×10^{-1}	1.0	NLO
	WZG_TuneCP5_13TeV-amcatnlo-pythia8	4.123×10^{-2}	1.0	NLO
	WZTo1L1Nu2Q_4f_TuneCP5_13TeV-amcatnloFXFX-pythia8	1.071×10^{1}	1.0	NLO
	WZTo1L3Nu_4f_TuneCP5_13TeV-amcatnloFXFX-pythia8	3.033	1.0	NLO
	WZTo2Q2L_mllmin4p0_TuneCP5_13TeV-amcatnloFXFX-pythia8	5.595	1.0	NLO
	WZTo3LNu_TuneCP5_13TeV-amcatnloFXFX-pythia8	4.430	1.0	NLO
	WZZ_TuneCP5_13TeV-amcatnlo-pythia8	5.565×10^{-2}	1.0	NLO
	ZGTo2NuG_TuneCP5_13TeV-amcatnloFXFX-pythia8	2.804×10^{1}	1.0	NLO
	ZGToLLG_01J_5f_TuneCP5_13TeV-amcatnloFXFX-pythia8	1.238×10^{2}	1.0	NLO
	ZZTo2L2Nu_TuneCP5_13TeV_powheg_pythia8	5.644×10^{-1}	1.0	NLO
	ZZTo2Q2L_mllmin4p0_TuneCP5_13TeV-amcatnloFXFX-pythia8	3.220	1.0	NLO

Process group	DAS name	Cross section σ [pb]	k-factor	Order
	ZZTo4L_TuneCP5_13TeV_powheg_pythia8	1.727	1.0	NLO
	ZZZ_TuneCP5_13TeV-amcatnlo-pythia8	1.398×10^{-2}	1.0	NLO
QCD	QCD_HT100to200_TuneCP5_PSWeights_13TeV-madgraph-pythia8	2.754×10^{7}	1.0	LO
	QCD_HT200to300_TuneCP5_PSWeights_13TeV-madgraph-pythia8	1.735×10^{6}	1.0	LO
	QCD_HT300to500_TuneCP5_PSWeights_13TeV-madgraph-pythia8	3.668×10^{5}	1.0	LO
	QCD_HT500to700_TuneCP5_PSWeights_13TeV-madgraph-pythia8	2.937×10^{4}	1.0	LO
	QCD_HT700to1000_TuneCP5_PSWeights_13TeV-madgraph-pythia8	6.524×10^{3}	1.0	LO
	QCD_HT1000to1500_TuneCP5_PSWeights_13TeV-madgraph-pythia8	1.064×10^{3}	1.0	LO
	QCD_HT1500to2000_TuneCP5_PSWeights_13TeV-madgraph-pythia8	1.215×10^{2}	1.0	LO
	QCD_HT2000toInf_TuneCP5_PSWeights_13TeV-madgraph-pythia8	2.542×10^{1}	1.0	LO
	QCD_Pt_15to30_TuneCP5_13TeV_pythia8	1.837×10^{9}	1.0	LO
	QCD_Pt_30to50_TuneCP5_13TeV_pythia8	1.409×10^{8}	1.0	LO
	QCD_Pt_50to80_TuneCP5_13TeV_pythia8	1.920×10^{7}	1.0	LO
	QCD_Pt_80to120_TuneCP5_13TeV_pythia8	2.763×10^{6}	1.0	LO
	QCD_Pt_120to170_TuneCP5_13TeV_pythia8	4.711×10^{5}	1.0	LO
	QCD_Pt_170to300_TuneCP5_13TeV_pythia8	1.033×10^{5}	1.0	LO
	QCD_Pt_300to470_TuneCP5_13TeV_pythia8	7.823×10^{3}	1.0	LO
	QCD_Pt_470to600_TuneCP5_13TeV_pythia8	6.482×10^{2}	1.0	LO
	QCD_Pt_600to800_TuneCP5_13TeV_pythia8	1.869×10^{2}	1.0	LO
	QCD_Pt_800to1000_TuneCP5_13TeV_pythia8	3.229×10^{1}	1.0	LO
	QCD_Pt_1000to1400_TuneCP5_13TeV_pythia8	9.418	1.0	LO
	QCD_Pt_1400to1800_TuneCP5_13TeV_pythia8	8.427×10^{-1}	1.0	LO
	QCD_Pt_1800to2400_TuneCP5_13TeV_pythia8	1.149×10^{-1}	1.0	LO
	QCD_Pt_2400to3200_TuneCP5_13TeV_pythia8	6.830×10^{-3}	1.0	LO
	QCD_Pt_3200toInf_TuneCP5_13TeV_pythia8	1.654×10^{-4}	1.0	LO
Top	ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	1.360×10^{2}	1.0	NLO
	ST_s-channel_4f_hadronicDecays_TuneCP5_13TeV-amcatnlo-pythia8	7.104	1.0	NLO
	ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	8.095×10^{1}	1.0	NLO
	ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-amcatnlo-pythia8	3.360	1.0	NLO

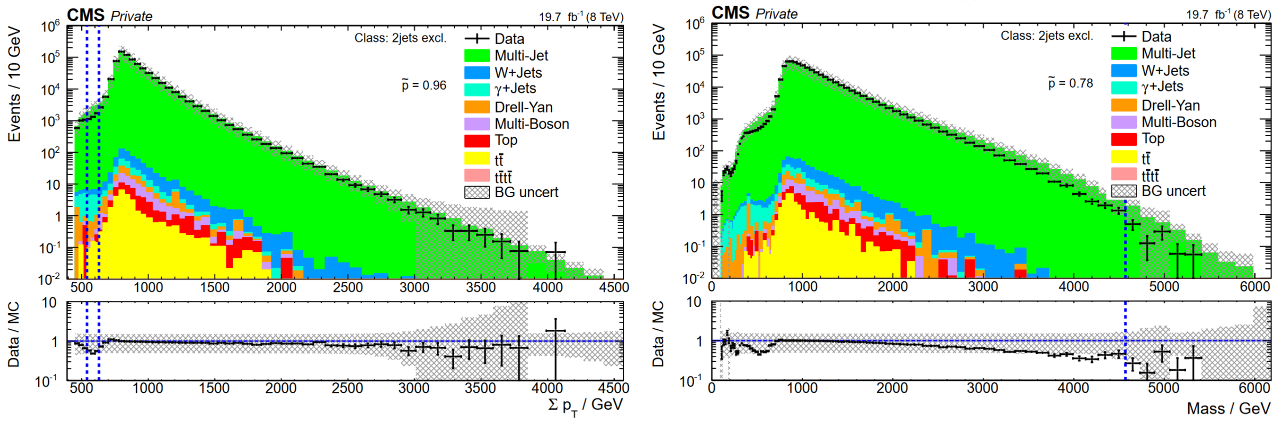
Process group	DAS name	Cross section σ [pb]	k-factor	Order
	ST_tW_top_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	3.809×10^{1}	1.0	NLO
	ST_tW_antitop_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	3.251×10^{1}	1.0	NLO
	TGJets_TuneCP5_13TeV-amcatnlo-madspin-pythia8	2.967	1.0	NLO
	tZq_ll_4f_ckm_NLO_TuneCP5_13TeV-amcatnlo-pythia8	7.580×10^{-2}	1.0	NLO
TTbar	TT_Mtt-700to1000_TuneCP5_13TeV-powheg-pythia8	7.300×10^{2}	1.139	NNLO
	TT_Mtt-1000toInf_TuneCP5_13TeV-powheg-pythia8	7.300×10^{2}	1.139	NNLO
	TTGG_TuneCP5_13TeV-amcatnlo-pythia8	1.696×10^{-2}	1.0	NLO
	TTGJets_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8	3.697	1.0	NLO
	TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8	9.334×10^{1}	1.0	NNLO
	TTToHadronic_TuneCP5_13TeV-powheg-pythia8	3.678×10^{2}	1.0	NNLO
	TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8	3.706×10^{2}	1.0	NNLO
	TTTT_TuneCP5_13TeV-amcatnlo-pythia8	9.103×10^{-3}	1.0	NLO
	TTWJetsToLNu_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8	2.043×10^{-1}	1.0	NLO
	TTWJetsToQQ_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8	4.062×10^{-1}	1.0	NLO
	TTWW_TuneCP5_13TeV-madgraph-pythia8	9.103×10^{-3}	1.0	NLO
	TTZToLL_5f_TuneCP5_13TeV-madgraphMLM-pythia8	5.272×10^{-1}	1.0	LO
	TTZToNuNu_TuneCP5_13TeV-amcatnlo-pythia8	1.476×10^{-1}	1.0	NLO
	TTZToQQ_TuneCP5_13TeV-amcatnlo-pythia8	5.297×10^{-1}	1.0	NLO
	TTZZ_TuneCP5_13TeV-madgraph-pythia8	1.386×10^{-3}	1.0	NLO
W	WJetsToLNu_TuneCP5_13TeV-amcatnloFXFX-pythia8	6.029×10^{4}	1.02	NNLO
	WJetsToLNu_Pt-100To250_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	6.778×10^{2}	1.02	NNLO
	WJetsToLNu_Pt-250To400_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	2.408×10^{1}	1.02	NNLO
	WJetsToLNu_Pt-400To600_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	3.056	1.02	NNLO
	WJetsToLNu_Pt-600ToInf_MatchEWPDG20_TuneCP5_13TeV-amcatnloFXFX-pythia8	4.602×10^{-1}	1.02	NNLO
	WToENu_M-200_TuneCP5_13TeV-pythia8	6.236	1.337	NNLO
	WToENu_M-500_TuneCP5_13TeV-pythia8	2.138×10^{-1}	1.331	NNLO

Process group	DAS name	Cross section σ [pb]	k-factor	Order
	WToENu_M-1000_TuneCP5_13TeV-pythia8	1.281×10^{-2}	1.327	NNLO
	WToENu_M-4000_TuneCP5_13TeV-pythia8	3.030×10^{-6}	0.45	NNLO
	WToENu_M-3000_TuneCP5_13TeV-pythia8	2.904×10^{-5}	1.136	NNLO
	WToENu_M-2000_TuneCP5_13TeV-pythia8	5.560×10^{-4}	1.257	NNLO
	WToMuNu_M-200_TuneCP5_13TeV-pythia8	6.236	1.289	NNLO
	WToMuNu_M-500_TuneCP5_13TeV-pythia8	2.138×10^{-1}	1.273	NNLO
	WToMuNu_M-1000_TuneCP5_13TeV-pythia8	1.281×10^{-2}	1.26	NNLO
	WToMuNu_M-2000_TuneCP5_13TeV-pythia8	5.560×10^{-4}	1.173	NNLO
	WToMuNu_M-3000_TuneCP5_13TeV-pythia8	2.904×10^{-5}	1.038	NNLO
	WToMuNu_M-4000_TuneCP5_13TeV-pythia8	3.030×10^{-6}	0.409	LO
	WToTauNu_M-200_TuneCP5_13TeV-pythia8-tauola	6.370	1.0	LO
	WToTauNu_M-500_TuneCP5_13TeV-pythia8-tauola	2.240×10^{-1}	1.0	LO
	WToTauNu_M-1000_TuneCP5_13TeV-pythia8-tauola	1.370×10^{-2}	1.0	LO
	WToTauNu_M-2000_TuneCP5_13TeV-pythia8-tauola	5.560×10^{-4}	1.0	LO
	WToTauNu_M-3000_TuneCP5_13TeV-pythia8-tauola	3.420×10^{-5}	1.0	LO
	WToTauNu_M-4000_TuneCP5_13TeV-pythia8-tauola	3.030×10^{-6}	1.0	LO

Tab. A.1: Full list of MC samples. Note that, as stated in the text, the S_{T}- and p_{T}-binned QCD samples were never used simultaneously, since they cover the same processes.

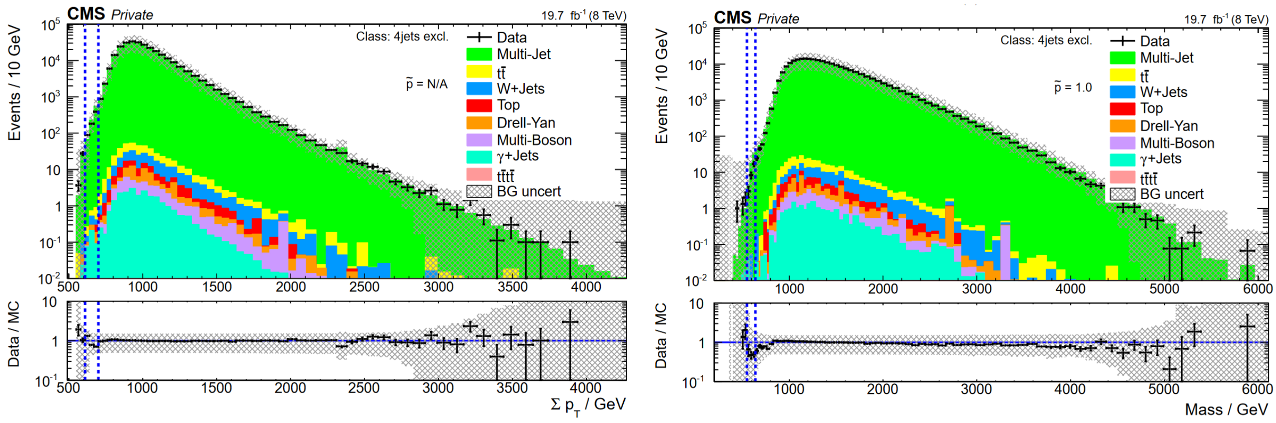
B Additional content for chapter 4

B.1 Previous studies in MUSiC



(a) Transverse momenta sum S_{T} **(b)** Invariant mass m_{inv}

Fig. B.1: Distributions for the 2jets exclusive class taken from A. Albert's master thesis [68, p. 73, fig. 4.20].



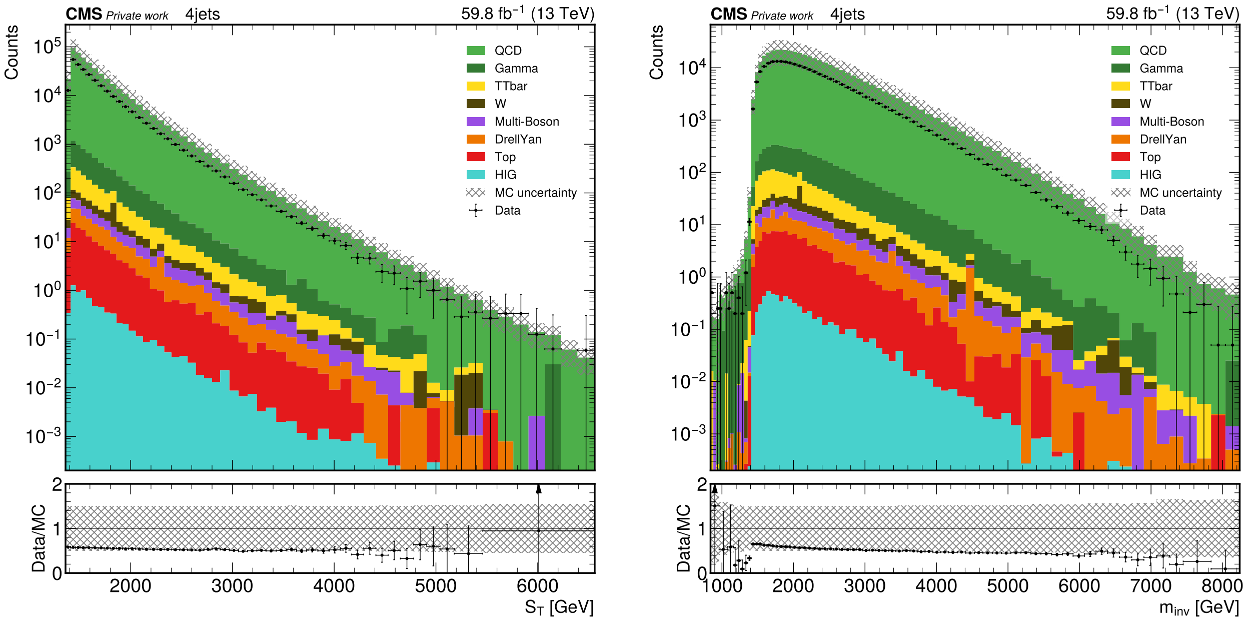
(a) Transverse momenta sum S_{T} **(b)** Invariant mass m_{inv}

Fig. B.2: Distributions for the 4jets exclusive class taken from A. Albert's master thesis [68, p. 74, fig. 4.21].

B.2 p_{T}-binned QCD samples and H_{T} trigger

This section presents additional plots for the sample-trigger configuration with the p_{T}-binned QCD samples and H_{T} trigger, originally presented in sec. 4.10.1.

Two distributions of the 4jets exclusive class are presented in fig. B.3. These distributions show roughly similar behavior as the 2jets exclusive class, presented in fig. 4.3. Again, for the exclusive class, the H_{T} trigger efficiency cut is visible in the respective transverse momenta sum plot (fig. B.3a). A deviation between data and MC, of the same order as already seen (factor ≈ 2), is observed for this class. This is also true for all other distributions for the class which are not shown. The S_{T} plot shows a mostly constant deviation. The m_{inv} distribution (fig. B.3b) also shows behavior already observed for the 2jets exclusive class, where some events were leaking in the m_{inv} below the trigger efficiency cut applied to H_{T}. The ratio is observed to be around ≈ 0.5 for lower energies and decreasing with higher energies. In the leaking region (below $m_{\mathrm{inv}} < 1400\,\mathrm{GeV}$), statistics are low and worse agreement is observed.

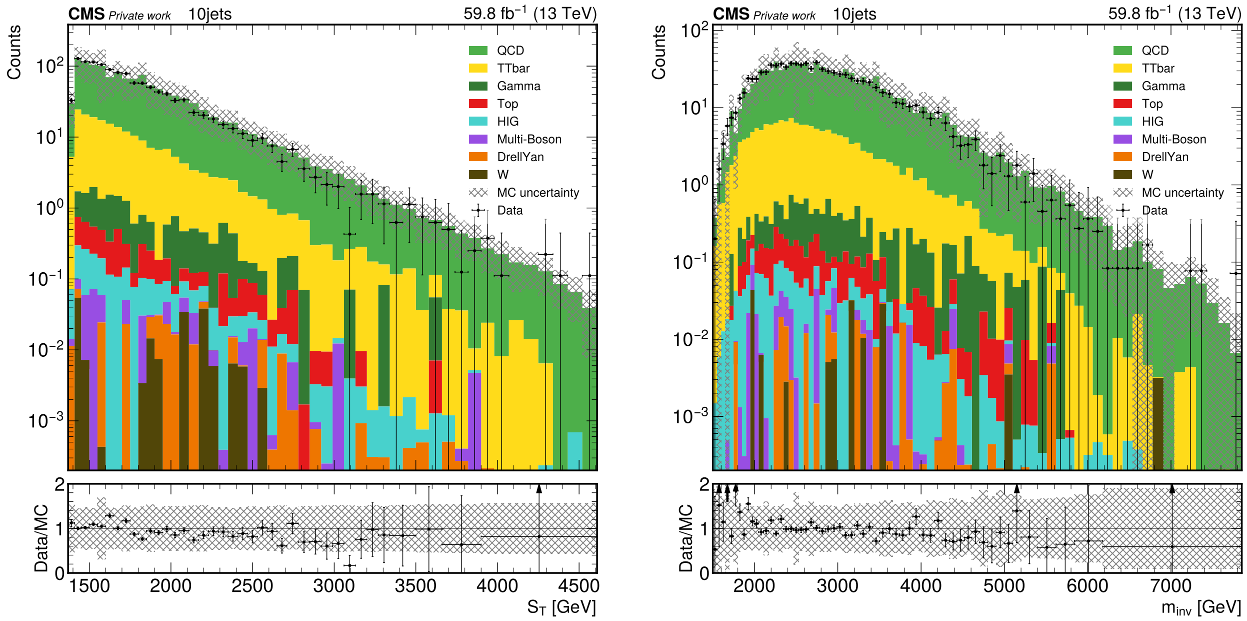


(a) Transverse momenta sum S_{T}

(b) Invariant mass m_{inv}

Fig. B.3: Distributions for the 4jets exclusive class with p_{T}-binned QCD samples and S_{T} trigger.

Fig. B.4 presents the 10jets exclusive class. Both the S_T (fig. B.4a) and m_inv (fig. B.4b) distributions show decent agreement of data and MC with a data/MC ratio around one fully covered in the uncertainties. Apparently, the smaller ratio for lower multiplicities is recovered for this class, which was already suggested by the comparison of the total class event counts in fig. 4.5. The agreement for this class is remarkable, even apart from the fact that agreement in lower classes is not achieved since the statistics are rather low.



(a) Transverse momenta sum S_T

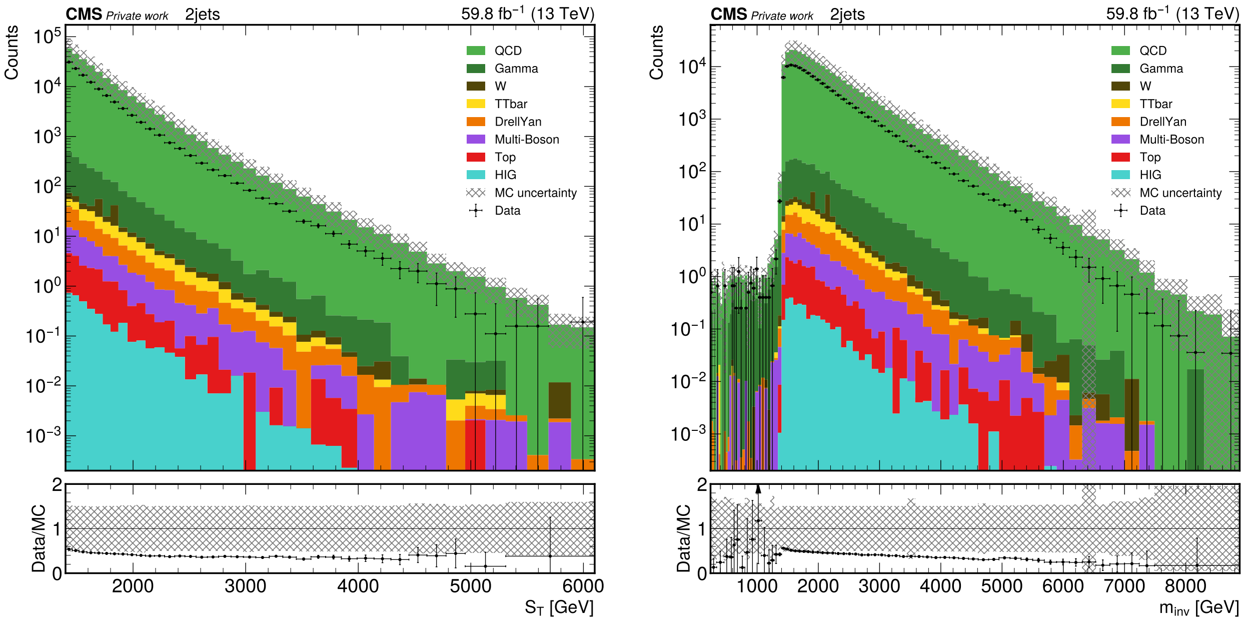
(b) Invariant mass m_inv

Fig. B.4: Distributions for the 10jets exclusive class with p_T-binned QCD samples and S_T trigger.

B.3 H_T-binned QCD samples and H_T trigger

Since the H_T trigger is used, the $H_\mathrm{T} > 1400\,\mathrm{GeV}$ trigger efficiency cut is applied. However, this configuration uses the H_T-binned QCD samples. 15 exclusive and 17 jet-inclusive classes are found in data and 17 exclusive and 17 jet-inclusive classes in MC, as for the other analysis with the H_T-trigger above (sec. 4.10.1).

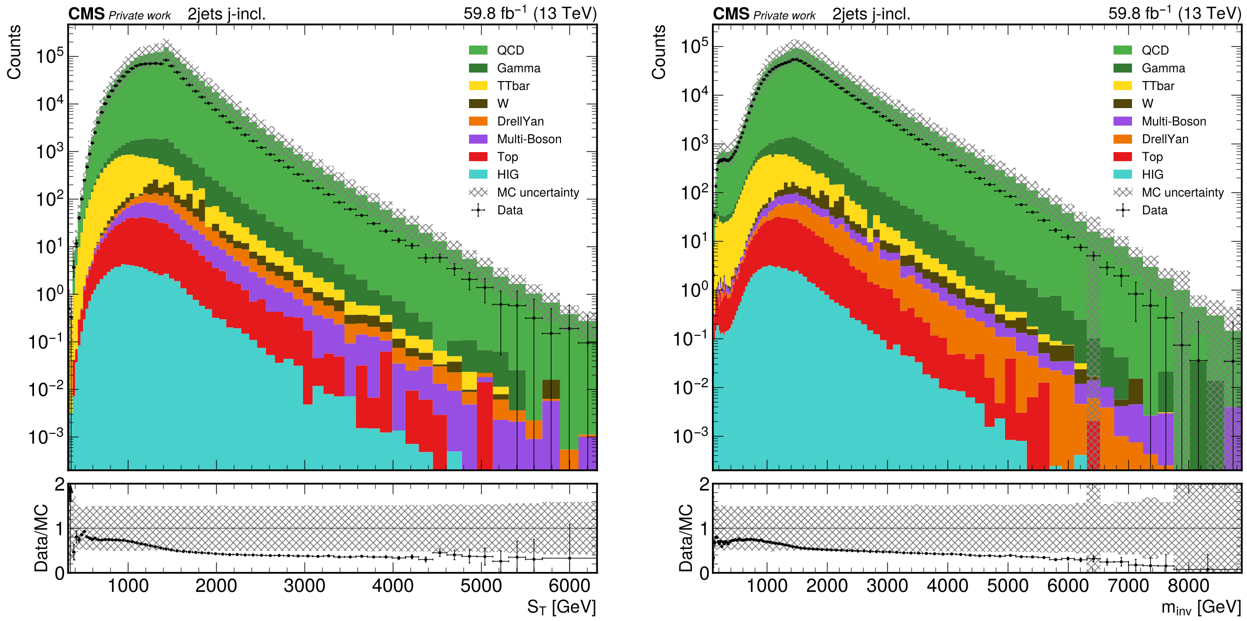
Although a different QCD dataset is used, the distributions look mostly similar. Most prominently, the strong deviation between data and MC is observed in a similar fashion as for the already presented configuration. However, the deviation seems slightly stronger, with a data/MC factor in the order of $\approx 0.4 - 0.5$. The offset does not seem as constant for all distributions as for the already presented configuration, since for this sample-trigger-configuration the data/MC offset varies more between the S_T and m_inv distributions. The distributions also again show a decreasing ratio with increasing energy in the energy-like plots (S_T and m_inv). Distributions for the 2jets exclusive class are shown in fig. B.5. The 2jets jet-inclusive class is presented in fig. B.6 and the 4jets exclusive class in fig. B.7.



(a) Transverse momenta sum S_T

(b) Invariant mass m_inv

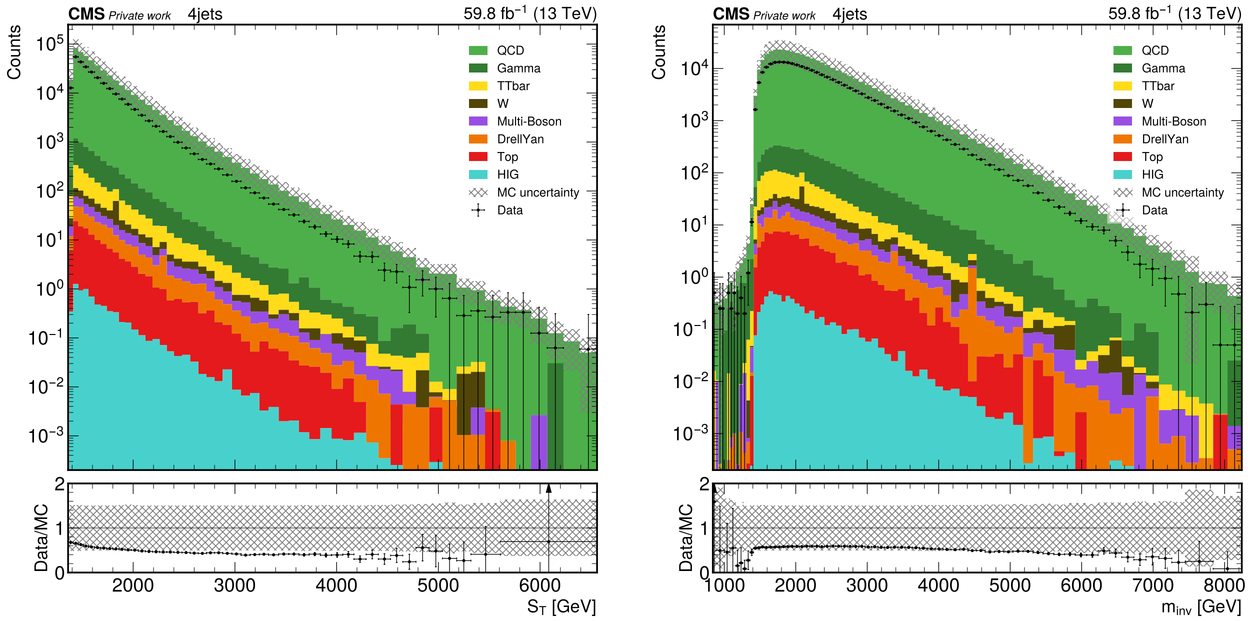
Fig. B.5: Distributions for the 2jets exclusive class with S_T-binned QCD samples and S_T trigger.



(a) Transverse momenta sum S_{T}

(b) Invariant mass m_{inv}

Fig. B.6: Distributions for the 2jets jet-inclusive class with S_{T}-binned QCD samples and S_{T} trigger.

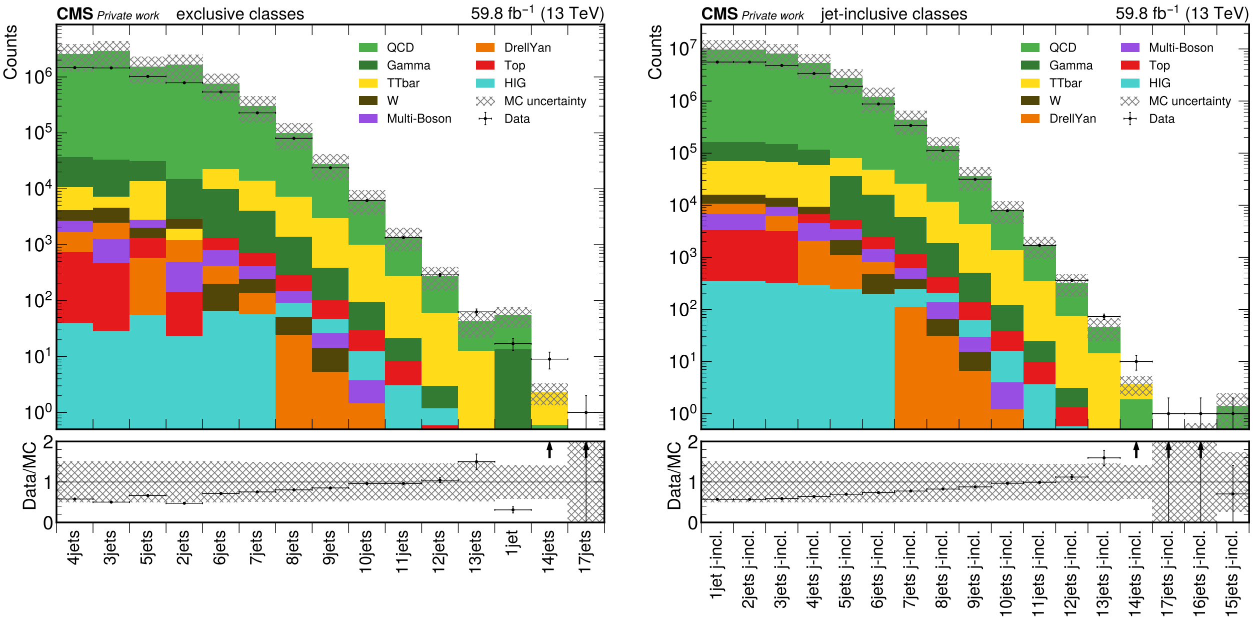


(a) Transverse momenta sum S_{T}

(b) Invariant mass m_{inv}

Fig. B.7: Distributions for the 4jets exclusive class with S_{T}-binned QCD samples and S_{T} trigger.

The integrated class counts are shown in fig. B.8. The data/MC ratio shows a similar increase with the jet multiplicity, as was already observed in the other configuration. Also, for jet multiplicities $2-5$, the ratio seems almost constant at ≈ 0.5, as observed for the previous configuration.



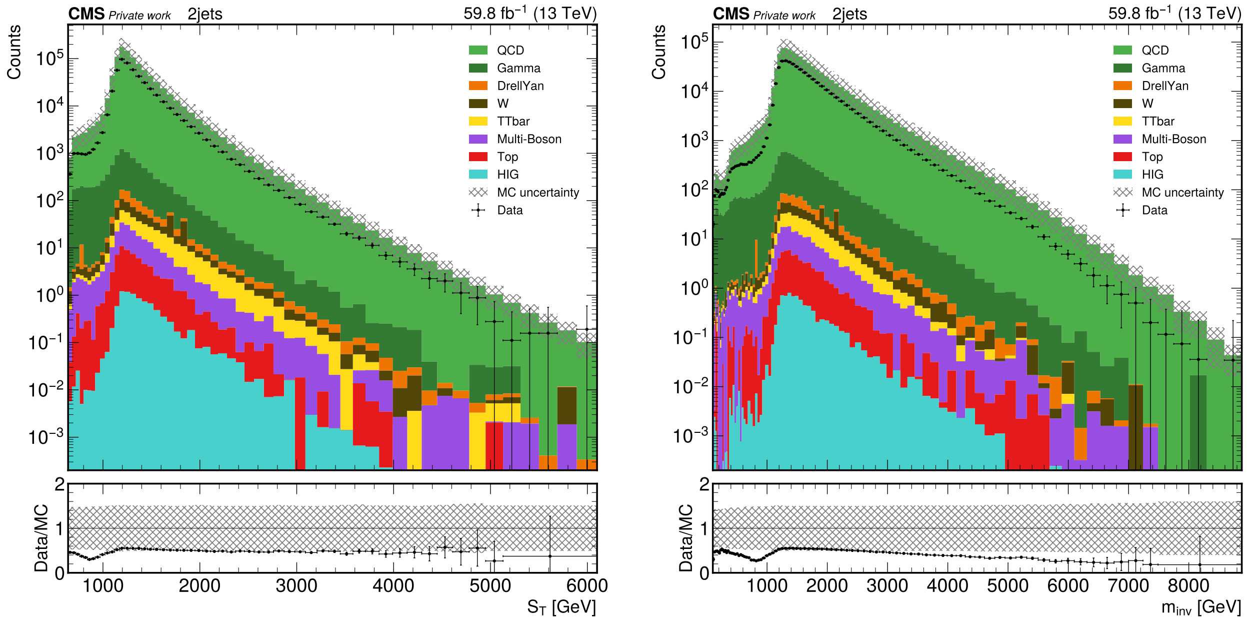
(a) Exclusive classes

(b) Jet-inclusive classes

Fig. B.8: Comparison of the integrated event counts for the classes with H_T-binned QCD samples and H_T trigger.

B.4 p_T-binned QCD samples and p_T trigger

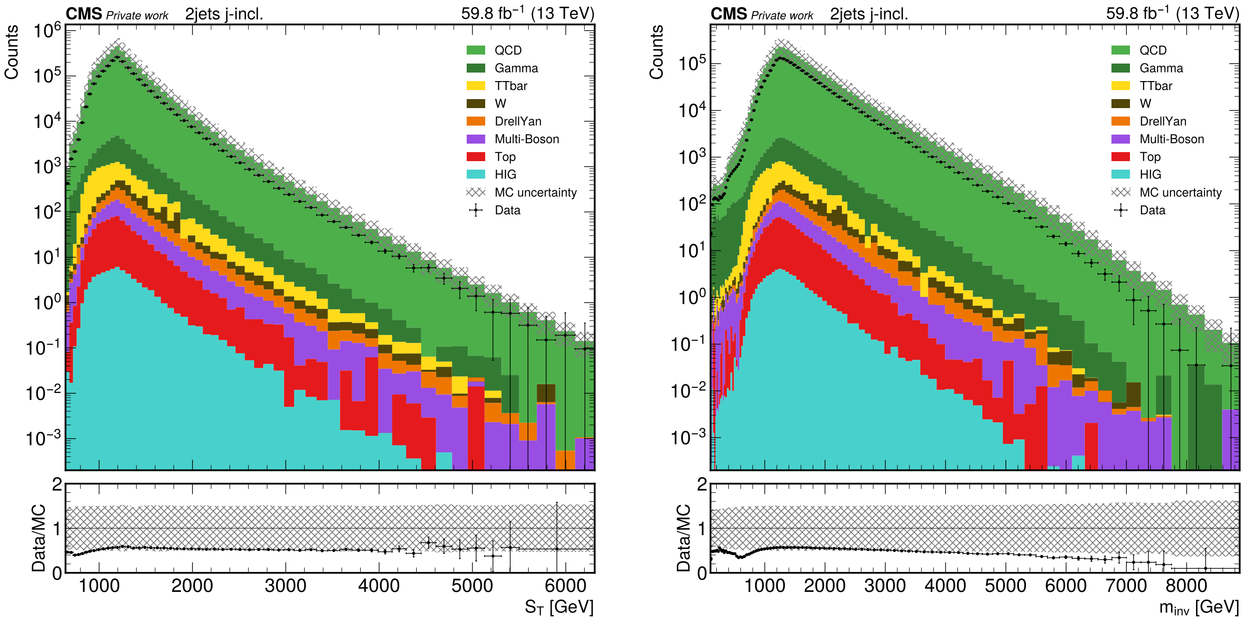
The last configuration that will be explored involves the p_T trigger and the p_T-binned QCD samples, with the $p_T > 600\,\mathrm{GeV}$ trigger efficiency cut applied for the leading jet. There were 14 exclusive and 14 jet-inclusive classes found in data and 16 exclusive and 16 jet-inclusive classes in MC.
Again, a similar deviation between data and MC is observable, with an average data/MC ratio of ≈ 0.5. However, the shape of the S_T and m_{inv} is different, especially for the exclusive event classes. This is a consequence of applying a p_T cut to ensure a high trigger efficiency instead of an H_T cut. Since only one jet with the given p_T is required for the event, it is possible that dijet events down to a transverse momenta sum of $S_T = 650\,\mathrm{GeV}$ are selected because the second jet only has to fulfill the MUSiC selection requirement of $p_T > 50\,\mathrm{GeV}$. This explains the observed tails in the exclusive class distributions for the 2jets exclusive class in fig. B.9. An energy dependence of the ratio is observed, similar to the configurations above. Distributions for the 2jets inclusive class (fig. B.10) and the 4jets exclusive class (fig. B.11) are also presented.



(a) Transverse momenta sum S_T

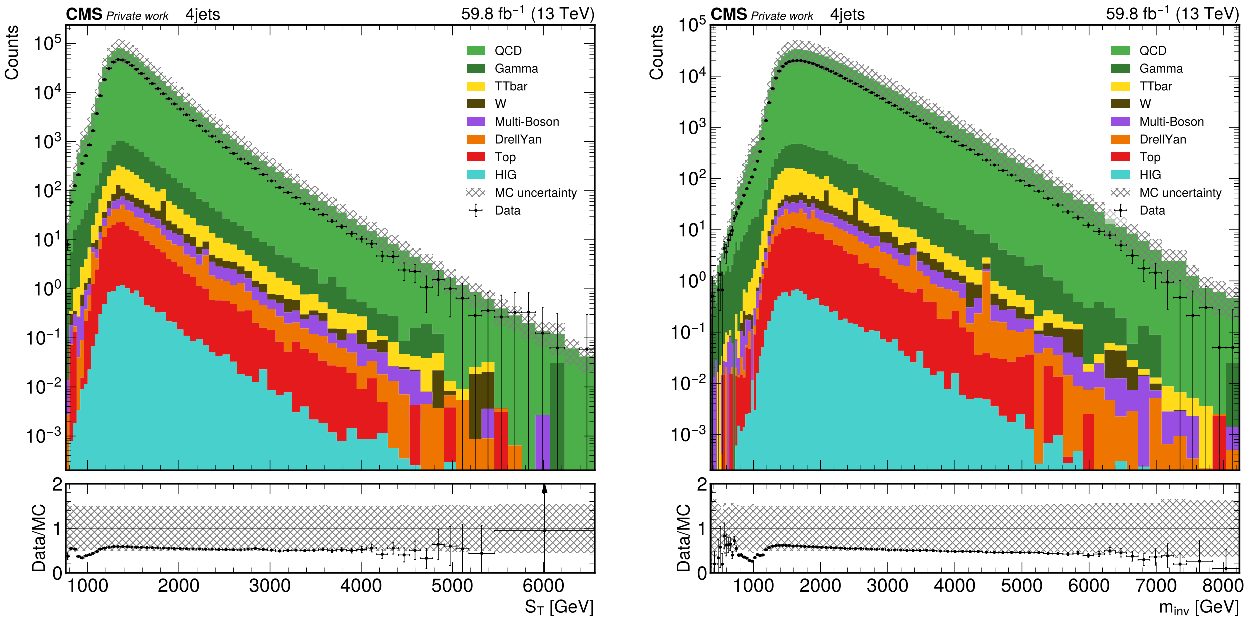
(b) Invariant mass m_{inv}

Fig. B.9: Distributions for the 2jets exclusive class with p_T-binned QCD samples and p_T trigger.



(a) Transverse momenta sum S_{T} **(b)** Invariant mass m_{inv}

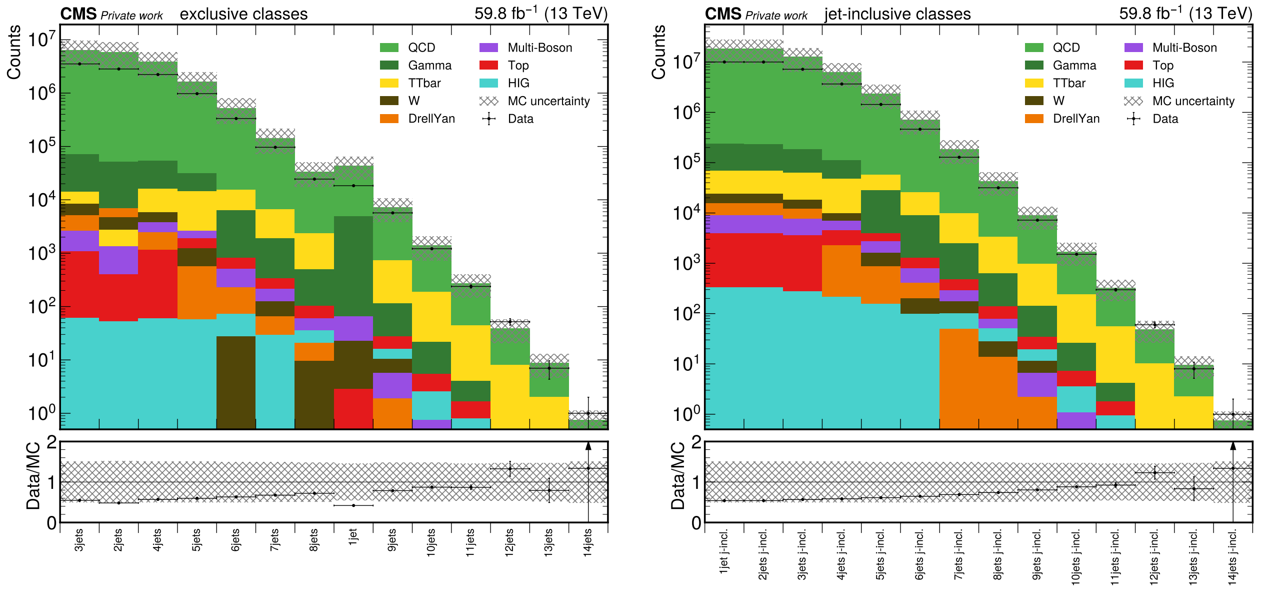
Fig. B.10: Distributions for the 2jets jet-inclusive class with p_{T}-binned QCD samples and p_{T} trigger.



(a) Transverse momenta sum S_{T} **(b)** Invariant mass m_{inv}

Fig. B.11: Distributions for the 4jets exclusive class with p_{T}-binned QCD samples and p_{T} trigger.

The integrated class counts for the exclusive classes are shown in fig. B.12. Again, a dependency of the ratio from the jet multiplicity is visible, but it is weaker than for the H_T triggered configurations presented above. The large uncertainties, allegedly caused by a few events, are apparently also visible in the 2 and 3 jet integrated class counts. Note that this configuration with the p_T trigger was not able to find the 17 jet event, this is discussed further in sec. 4.10.3.

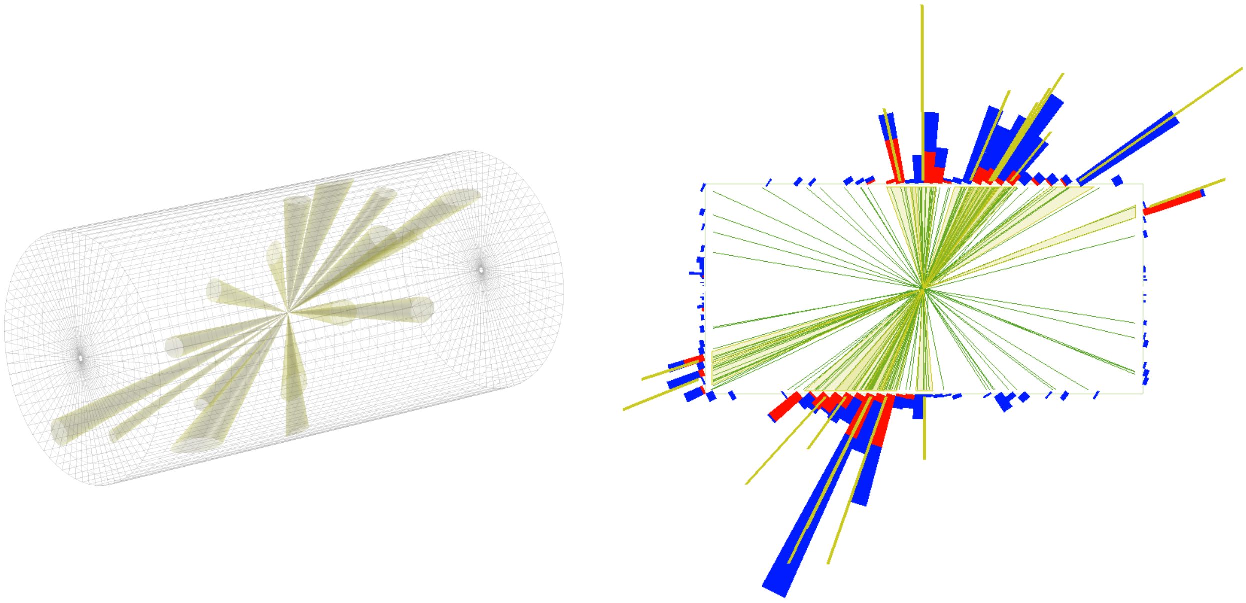


(a) Exclusive classes

(b) Jet-inclusive classes

Fig. B.12: Comparison of the integrated event counts for the classes with p_T-binned QCD samples and p_T trigger.

B.5 Details on the 17 jet event



(a) 3D view with only the reconstructed jets

(b) View from the side

Fig. B.13: Additional views of the 17 jet event.

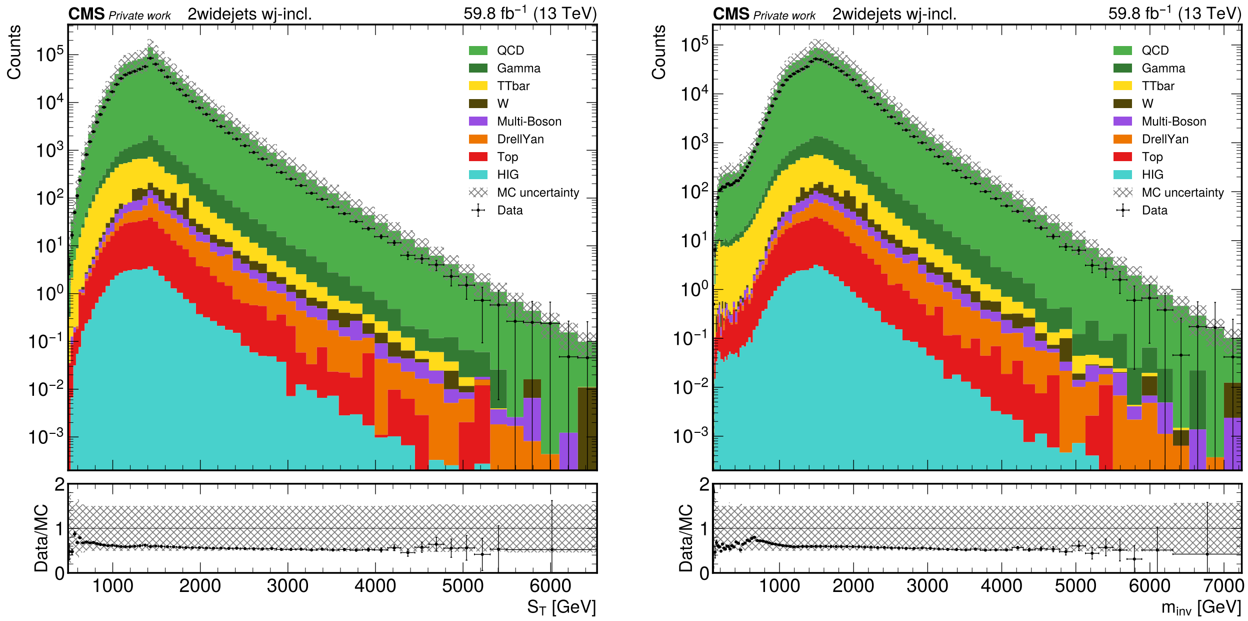
Info	Value
Date recorded	August 30th, 2018
Time recorded	22:10:51
Run number	321975
Luminosity section	396
Event number	697993631
Orbit	103619678
Crossing	1795

Tab. B.1: Event information of the 17 jet event.

C Additional content for chapter 5

C.1 Results

Fig. C.1 presents the distributions for the 2widejets widejet-inclusive class. The shape as well as the statistics seem more or less unchanged from the exclusive class, which was already presented in fig. 5.2. This indicates that the higher jet multiplicities have much fewer events than before the algorithm was applied.

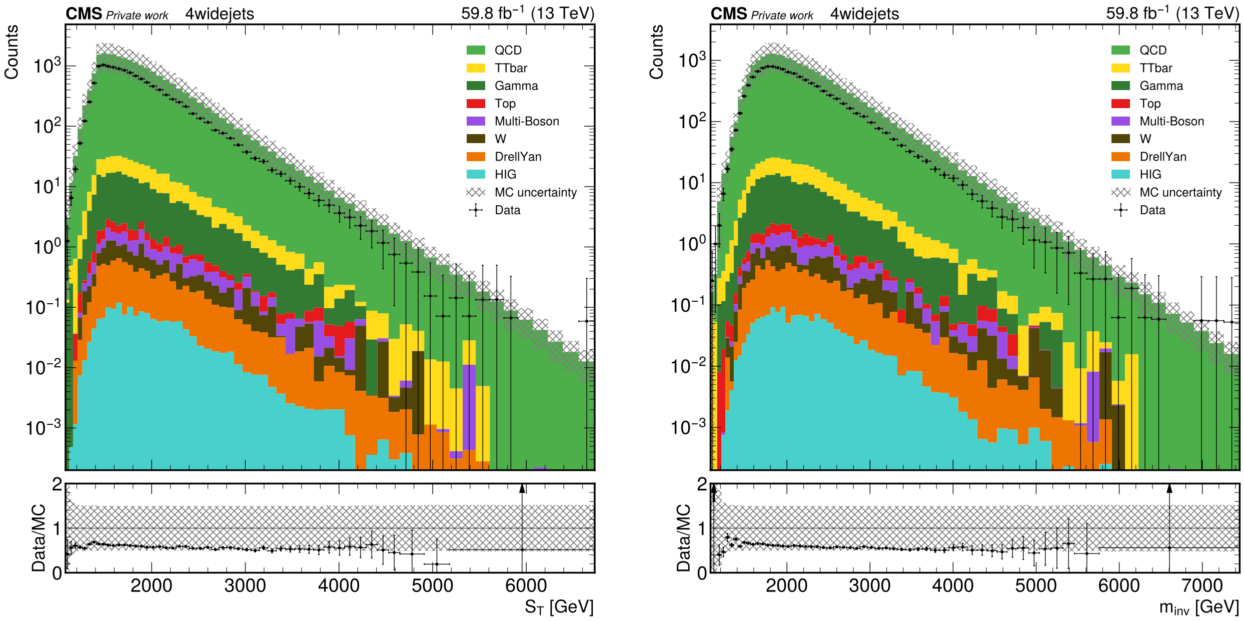


(a) Transverse momenta sum S_T

(b) Invariant mass m_inv

Fig. C.1: Distributions for the 2widejets widejet-inclusive class after applying the generalized wide jet algorithm.

This hypothesis is confirmed when looking at the 4widejets exclusive class, presented in fig. C.2. The statistics are reduced by about one order of magnitude in comparison to before applying the wide jet algorithm (see fig. B.3). Apart from the lower statistics, the ratio between data and MC is still observed to be ≈ 0.5. It appears almost constant over the entire energy range. Different from the lower multiplicity classes, the energy dependence of the ratio was not as strong for the 4widejets class (see fig. B.3), however, it can be argued that applying the algorithm still has a slightly flattening effect on the invariant mass distribution.

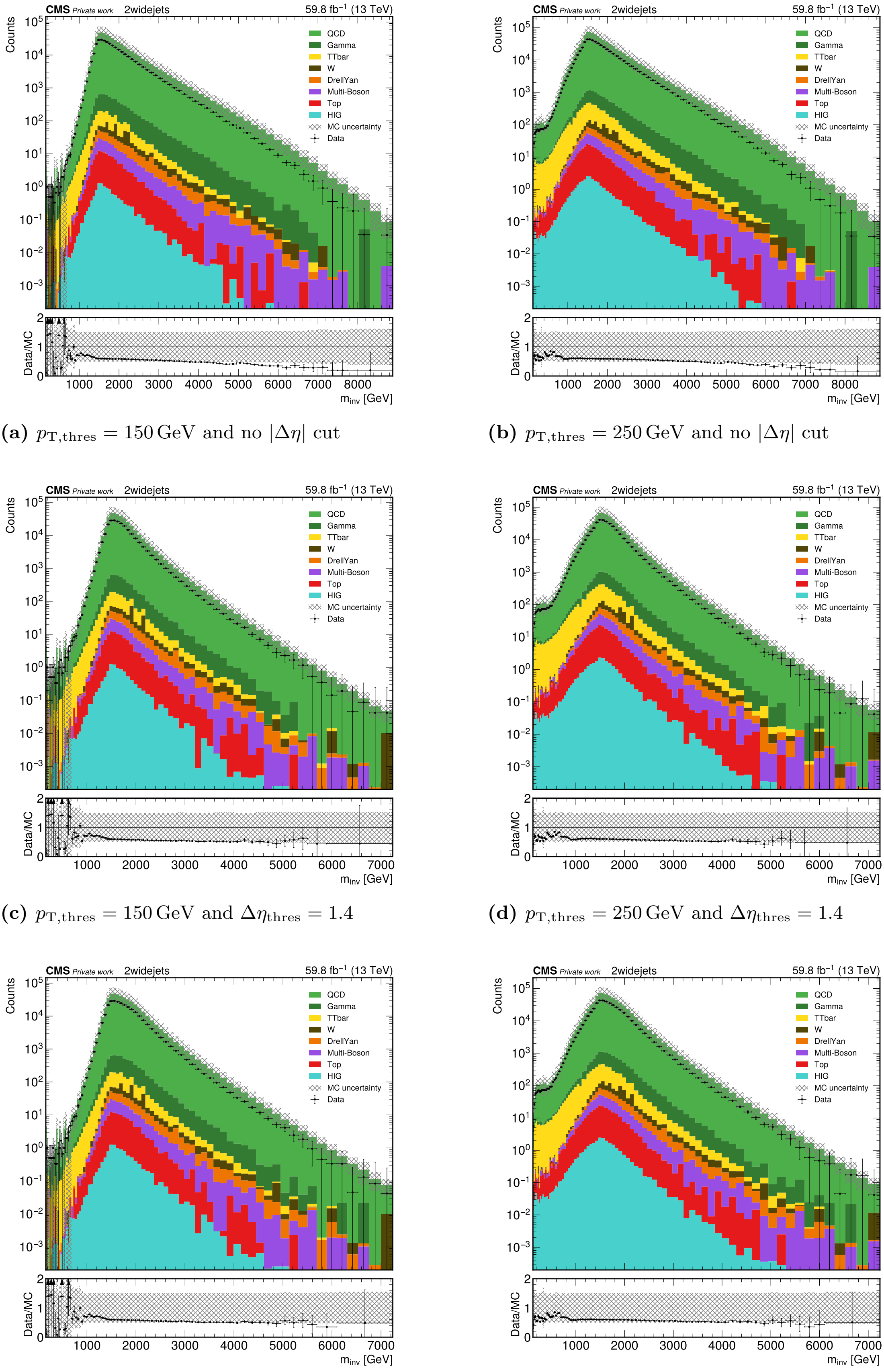


(a) Transverse momenta sum S_T

(b) Invariant mass m_inv

Fig. C.2: Distributions for the 4widejets exclusive class after applying the generalized wide jet algorithm.

C.2 Optimization



(a) $p_{\mathrm{T,thres}} = 150\,\mathrm{GeV}$ and no $|\Delta\eta|$ cut

(b) $p_{\mathrm{T,thres}} = 250\,\mathrm{GeV}$ and no $|\Delta\eta|$ cut

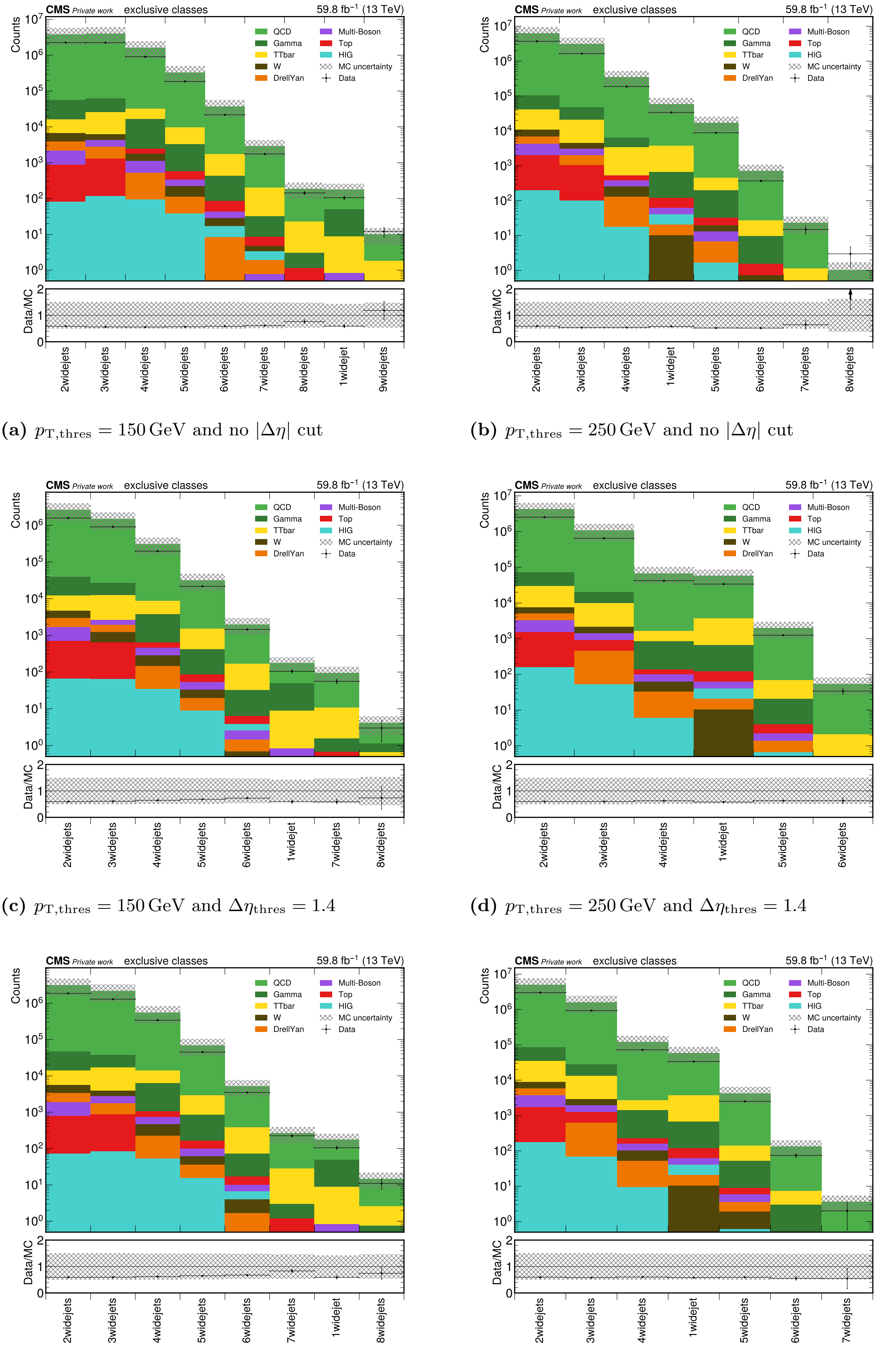
(c) $p_{\mathrm{T,thres}} = 150\,\mathrm{GeV}$ and $\Delta\eta_{\mathrm{thres}} = 1.4$

(d) $p_{\mathrm{T,thres}} = 250\,\mathrm{GeV}$ and $\Delta\eta_{\mathrm{thres}} = 1.4$

(e) $p_{\mathrm{T,thres}} = 150\,\mathrm{GeV}$ and $\Delta\eta_{\mathrm{thres}} = 1.8$

(f) $p_{\mathrm{T,thres}} = 250\,\mathrm{GeV}$ and $\Delta\eta_{\mathrm{thres}} = 1.8$

Fig. C.3: Invariant mass distributions m_{inv} for the 2widejets exclusive class for different merging and pseudorapidity cut parameters.



(a) $p_{\mathrm{T,thres}} = 150\,\mathrm{GeV}$ and no $|\Delta\eta|$ cut

(b) $p_{\mathrm{T,thres}} = 250\,\mathrm{GeV}$ and no $|\Delta\eta|$ cut

(c) $p_{\mathrm{T,thres}} = 150\,\mathrm{GeV}$ and $\Delta\eta_{\mathrm{thres}} = 1.4$

(d) $p_{\mathrm{T,thres}} = 250\,\mathrm{GeV}$ and $\Delta\eta_{\mathrm{thres}} = 1.4$

(e) $p_{\mathrm{T,thres}} = 150\,\mathrm{GeV}$ and $\Delta\eta_{\mathrm{thres}} = 1.8$

(f) $p_{\mathrm{T,thres}} = 250\,\mathrm{GeV}$ and $\Delta\eta_{\mathrm{thres}} = 1.8$

Fig. C.4: Integrated event counts for the exclusive widejet classes for different merging and pseudorapidity cut parameters.

C.3 Illustration with event displays

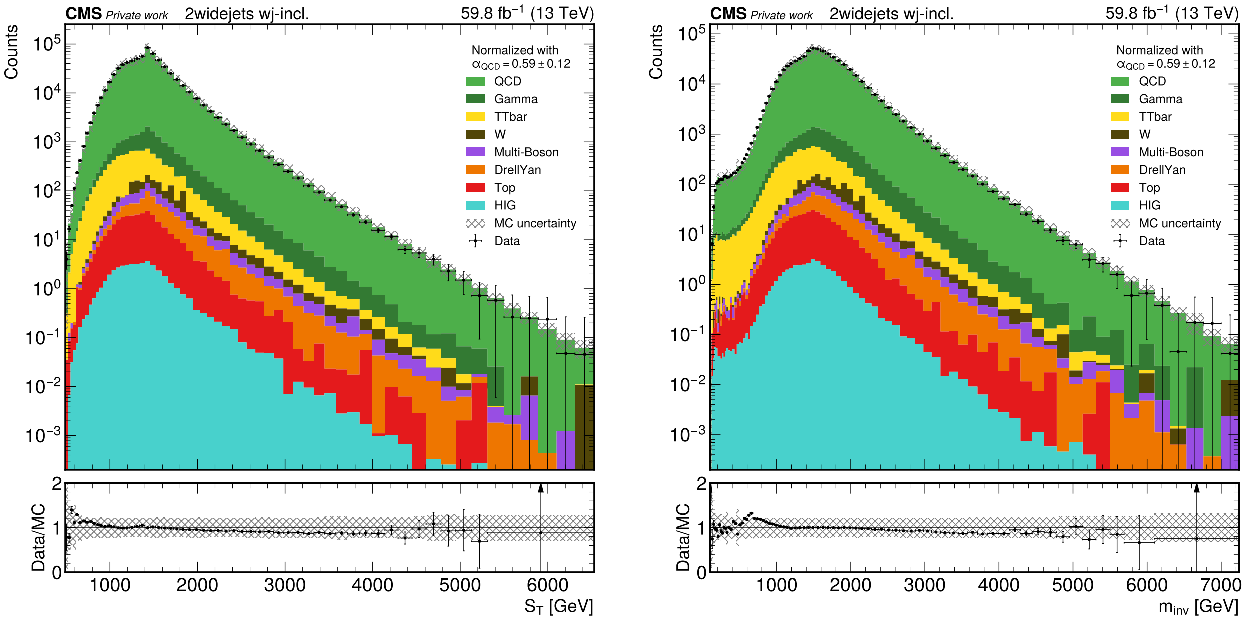
Info	Value
Date recorded	August 1$^{\mathrm{st}}$, 2018
Time recorded	19:51:20
Run number	320688
Luminosity section	427
Event number	595066336
Orbit	111934940
Crossing	2451

Tab. C.1: Event information of the presented example event that undergoes 3jet → 2jet merging.

D Additional content for chapter 6

D.1 Results

Fig. D.1 presents the 2widejets widejet-inclusive class with the QCD normalization applied. Since high jet multiplicity classes have low event counts because of the applied wide jet algirithm, the shape of the distribution is mostly similar to the exclusive class, which was presented in fig. 6.6. The normalization leads to improved agreement of data and MC.



(a) Transverse momenta sum S_T

(b) Invariant mass m_inv

Fig. D.1: Distributions for the 2widejets widejet-inclusive class with applied normalization.

List of Figures

List of Tables

List of Acronyms

ALICE A Large Ion Collider Experiment

AOD Analysis Object Data

ATLAS A Toroidal LHC Apparatus

BOOSTER Proton Synchrotron Booster

BSM Beyond the Standard Model

CERN European Organization for Nuclear Research

CMS Compact Muon Solenoid

CRAB CMS Remote Analysis Builder

CSC Cathode Strip Chambers

DAS Data Aggregation System

DESY Deutsches Elektronen-Synchrotron

DT Drift Tubes

ECAL Electromagnetic Calorimeter

FASER Forward Search Experiment

HB Hadron Barrel

HCAL Hadronic Calorimeter

HE Hadron Endcap

HLT High-Level Trigger

HO Hadron Outer

JEC Jet Energy Correction

JER Jet Energy Resolution

L1 Trigger Level 1 Trigger

LEE Look-Elsewhere Effect

LEP Large Electron-Positron Collider

LHC Large Hadron Collider

LHCb Large Hadron Collider beauty

LHCf Large Hadron Collider forward

LINAC 2 Linear accelerator 2

LINAC 4 Linear accelerator 4

LO Leading Order

MC Monte Carlo

MET Missing Transverse Energy

MoEDAL Monopole and Exotics Detector at the LHC

MUSiC Model Unspecific Search in CMS

PDFs Parton Distribution Functions

PF Particle Flow

PS Proton Synchrotron

PU Pileup

QCD Quantum Chromodynamics

QED Quantum Electrodynamics

RoI Region of Interest

RPC Resitive Plate Chambers

SM Standard Model

SND@LHC Scattering and Neutrino Detector

SPS Super Proton Synchrotron

TOTEM Total, elastic and diffractive cross-section measurement

WLCG Worldwide LHC Computing Grid

Acknowledgements

Finally, I would like to express my acknowledgements to a few people that had an influence on this thesis.

Most importantly, I would like to thank Prof. Dr. Thomas Hebbeker for the great opportunity to write my bachelor thesis at his institute on MUSiC. Thank you also to Prof. Dr. Martin Erdmann, who kindly agreed to be the second corrector for this thesis. I also would like to thank Dr. Arnd Meyer for accepting my request for a thesis in the MUSiC group, for laying out the idea for this thesis, and for his continuous interest in the progress of my work.

I would like to thank Dr. Felipe Torres da Silva de Araújo for supervising this thesis, without whose help it would not have been possible to conduct the work presented in this document. Thank you for always dedicating your time to answer all of my questions, providing me with an operational setup to work on my thesis, investing your time and knowledge to help me solve any problem that occurred during my work, and finally proofreading this thesis. And apart from your professional help, thank you for keeping to speak up against the doubts I had concerning this thesis and all the nice conversations we had during the last three months.

I would also like to thank the other ”MUSiCians”, namely M.Sc. Chinmay Seth, B.Sc. Yannik Kaiser, Lukas Karwatzki, and Patrick Kersten for helping me with smaller problems and for all the nice conversations. I really enjoyed working with you in such a friendly working environment. A special thank you to B.Sc. Yannik Kaiser for providing me with his thesis template, from which I took inspiration when creating this document.

Also, I would like to express a thank you for the pleasant food ordering Fridays to the GEM office and W'/l^* office and all the nice conversations we had during the three months.

I would like to thank my friends for the pleasant last three years of the bachelor studies. A special thank you to my roommate, who had to endure all my thoughts and worries on any university-related topic every day.

Lastly, I would like to thank my parents and my brother who have been supporting me my whole life and especially during my studies, enabling me to study lighthearted over the last three years.