DEUTSCHES ELEKTRONEN-SYNCHROTRON
Ein Forschungszentrum der Helmholtz-Gemeinschaft

DESY-THESIS-2022-013
July 2022

Measurement of the Differential Cross Section of W and Z Bosons in the Boosted Region in Proton-Proton Collisions with the CMS Experiment

by

J. Lidrych

ISSN 1435-8085

NOTKESTRASSE 85 – 22607 HAMBURG

To be sure that your reports and preprints are promptly included in the HEP literature database
send them to (if possible by air mail):

DESY Zentralbibliothek Notkestraße 85 22607 Hamburg Germany	DESY Bibliothek Platanenallee 6 15738 Zeuthen Germany

Measurement of the differential cross section of W and Z bosons in the boosted region in proton-proton collisions with the CMS experiment

Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und
Naturwissenschaften Fachbereich Physik
der Universität Hamburg

vorgelegt von

Jindřich Lidrych

Hamburg
2021

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, den 4. Juli 2021

Jindřich Lidrych

Gutachter der Dissertation :	Prof. Dr. Johannes Haller Priv.-Doz. Hannes Jung
Zusammensetzung der Prüfungskommission :	Prof. Dr. Elisabetta Gallo Dr. Alexandre Glazov Prof. Dr. Johannes Haller Priv.-Doz. Hannes Jung Prof. Dr. Sven-Olaf Moch
Vorsitzender der Prüfungskommission :	Prof. Dr. Sven-Olaf Moch
Datum der Disputation :	3. September 2021
Vorsitzender des Fach-Promotionsausschusses Physik :	Prof. Dr. Wolfgang Hansen
Leiter des Fachbereichs Physik :	Prof. Dr. Günter H. W. Sigl
Dekan der Fakultät MIN :	Prof. Dr. Heinrich Graener

Abstract

In this thesis, a measurement of the differential cross section for hadronically decaying W and Z bosons produced in proton-proton collisions is presented. Decay products of the vector bosons with very high transverse momentum p_{T} are collimated in the momentum direction of the mother particle and are merged into a large cone-size jet with the characteristic inner structure. The measurement is based on data sample from proton-proton collisions at the centre-of-mass energy of 13 TeV collected with the CMS detector at the CERN LHC in the years 2016, 2017, and 2018, corresponding to an integrated luminosity of $134\,\mathrm{fb}^{-1}$. The jet substructure techniques together with advanced data-driven method are used to identify jets originating from the hadronic decay of W and Z bosons and reduce the overwhelming contribution mainly coming from QCD multijet production. This study probes a wide range of the kinematic region of the transverse momentum from 500 to 1200 GeV. The measured cross section is presented with a full set of systematic uncertainties and the results are compared with a theoretical prediction at next-to-leading order QCD accuracy.

Furthermore, the residual data-to-simulation jet energy correction is derived from dijet events using proton-proton collisions at the centre-of-mass energy of 13 TeV recorded by the CMS detector in 2018, corresponding to an integrated luminosity of $59.7\,\mathrm{fb}^{-1}$. The data-driven estimation of the residual jet energy correction explores back-to-back dijet events by the p_{T} balance and MPF methods. The corrections for jets with the transverse momentum $p_{\mathrm{T}} = 120\,\mathrm{GeV}$ have been found to differ from unity by less than 4% in the tracker acceptance region and up to 20% in the endcap-forward transition region. Precisely calibrated jets are one of the key ingredients for most analyses including hadronic final state.

Zusammenfassung

Diese Arbeit beschreibt die Messung des differenziellen Wirkungsquerschnitts der Produktion von W und Z Bosonen in Proton-Proton Kollisionen in hadronischen Endzuständen. Die Zerfallsprodukte von Vektorbosonen mit sehr hohem Transversalimpuls p_{T} sind stark gebündelt in Richtung des Mutterteilchens und werden in einen Jet mit großem Radius und charakteristischer Struktur rekonstruiert. Die Messung basiert auf Daten aus Proton-Proton Kollisionen bei einer Schwerpunktsenergie von 13 TeV, die mit dem CMS Detektor am CERN LHC in den Jahren 2016, 2017 und 2018 aufgezeichnet wurden und einer integrierten Luminosität von 134 fb^{-1} entsprechen. Techniken zur Charakterisierung der inneren Struktur des Jets und hochentwickelte auf Daten basierte Methoden werden verwendet um die von W und Z Bosonen entstandenen Jets zu identifizieren und die großen Beiträge von QCD induzierten Jets zu reduzieren. Die Messung des differenziellen Wirkungsquerschnitts testet einen großen Bereich des Transversalimpulses von 500 bis 1200 GeV und beinhaltet vollständige Bestimmung aller relevanten systematischen Unsicherheiten. Die Ergebnisse werden verglichen mit theoretischen Vorhersagen die in nächst höhere Ordnungen in QCD berechnet wurden.

Desweiteren wurden Korrekturen zur Energie von Jets in für Unterschiede zwischen Daten und Simulation bestimmt. Dazu wurden Ereignisse mit Dijets verwendet, die in Proton-Proton Kollisionen bei einer Schwerpunktsenergie von 13 TeV mit dem CMS Detektor im Jahr 2018 aufgezeichnet wurden und einer integrierten Luminosität von 59.7 fb^{-1} entsprechen. Die Daten basierte Bestimmung dieser Korrekturen mittels der p_{T} balance und MPF Methoden nutzt die spezielle Kinematik der Dijets. Die gemessenen Korrekturen für Jets mit Transversalimpuls von $p_{\mathrm{T}} = 120\mathrm{GeV}$ sind kleiner als 4% im Akzeptanzbereich des Spurdetectors und bis zu 20% im Übergangsbereich zwischen Endcap und Forward Detektoren. Eine präzise Kalibrierung der Jets ist ein Schlüsselelement für die meisten Analysen mit hadronischen Endzuständen.

"...když nemůžeš, tak přidej!?"
Emil Zátopek

Contents

Introduction

The Standard Model of particle physics is an extremely successful quantum field theory describing our current understanding of elementary particles and interactions which act between them. In the Standard Model, the interactions are realized by exchange of gauge bosons. The gluons are mediators of the strong interaction, while the photon γ acts as mediator of the electromagnetic interaction. The last interaction included in the Standard Model, weak interaction, is realized by the exchange of massive vector bosons, namely W and Z bosons. The unification of the two latter mentioned interactions by the electroweak theory in 1960s introduces the mechanism of electroweak spontaneous symmetry breaking and the existence of a new boson, the Higgs boson, was postulated. Even though W and Z bosons were discovered already in 1983 and the Standard model of particle physics has become a widely accepted theory of the elementary particles and interactions which act between them, it took almost fifty years to find evidence of the Higgs boson. For that purpose the Large Hadron Collider, the world's largest particle accelerator, was built and the discovery of a new particle that is consistent with the Higgs boson was announced in 2012.

At the collisions energies reachable at the Large Hadron Collider, the W and Z bosons are produced beyond the energy threshold. Thus, the transverse momentum can easily exceed their masses and they gain high Lorentz-boost. The Standard Model particles with high transverse momentum give the possibility to study the Standard Model in an extreme kinematic region and to further investigate Standard Model physics. When the heavy Standard Mode particles with sufficiently large transverse momenta decay, all decay products are highly boosted and collimated within a small cone. In the case of the hadronic decay mode, the decay products create a jet with a large cone radius and characteristic inner substructure. However, jets are produced by processes in strong interaction described by Quantum Chromodynamics with a very high rate which make the measurement with hadronically decaying W and Z bosons much more challenging. The recent progress in the understanding of the inner structure of jets, referred to as jet substructure, provided new options for the identification of boosted W and Z bosons. This doctoral thesis focuses on the first differential measurement of the cross section of hadronically decaying W and Z bosons with high transverse momenta. The measurement is based on data samples from proton-proton collisions at the centre-of-mass energy of 13 TeV collected with the CMS detector at the CERN LHC in the years 2016, 2017, and 2018, corresponding to an integrated luminosity of $134\,\mathrm{fb}^{-1}$.

This thesis is organized as three parts. Premisses, the first part of this thesis, is composed of five chapters. A brief overview of the Standard Model of particle physics is discussed in Chapter 1. Not only a theoretical introduction is given, but also the proton-proton collisions as an experimental tool to study the Standard Model is discussed. Chapter 2 focuses on jets in hadronic collisions. The jets are viewed as a proxy to high-energy quarks

and gluons as well as a tool to study highly energetic vector bosons via jet substructure. Several grooming techniques and two-prong substructure variables are briefly introduced. In Chapter 3, the Large Hadron Collider and the Compact Muon Solenoid detector are described. The main principles and parameters of the collider and detector are given and explained. In Chapter 4 the physics objects reconstruction and identification is discussed, with a special focus on jet reconstruction and calibration. Each step of the factorized approach of jet energy calibration is described in a greater details. Chapter 5 is dedicated to the determination of residual jet energy corrections. A data-driven estimation of the residual jet energy correction using dijet events in the back-to-back topology is explained.

The second part of this thesis is devoted to the main topic of the present thesis, the measurement of the differential cross section of hadronically decaying W and Z bosons with high transverse momentum p_{T}. This part, the physics analysis, is composed of five chapters. Chapter 6 introduces the analysis strategy. The physical condition and the measurement strategy are discussed. In Chapter 7, event samples and physics objects selection is given. Various corrections that are necessary to be applied to MC simulations are explained. Finally, the comparisons between data and MC simulations are shown for all relevant observables used for the measurement. Chapter 8 deals with the vector boson identification. The construction of a mass decorrelated tagger together with its commissioning is presented. Chapter 9 explains the modelling of physics processes that are relevant for the measurement. A special focus is put on to the data-driven estimation of the overwhelming contribution from QCD multijet production. In Chapter 10, the description of systematic uncertainties is given. Afterwards, the measurement at detector level is performed. A data unfolding method which treats detector effects is explained and applied to the measurement. Finally, the cross section corrected for all detector effects is reported together with a full set of systematics uncertainties.

The last part of thesis, Conclusions, is composed of only one chapters. Chapter 11 summaries the first differential measurement of cross section of hadronically decaying W and Z bosons in the boosted regime at the Large Hadron Collider. An outlook is given and perspectives for future measurement with boosted W and Z bosons within the Standard Model physics is discussed.

Part I

Premisses

1

Vector bosons in the Standard Model of particle physics

"When in trouble, postulate one of more new particles." With a bit of an exaggeration, this was a successful long-running theoretical strategy in particle physics for decades [1]. Already used in 1930s by W. Pauli to save in a natural way the fundamental law of energy conservation when he postulated a neutral undetected particle, neutrino. The same trick was also adopted by the inventors of the Standard Model (SM) to avoid of spoiling the perturbative renormalizability of the electroweak theory in 1960s [2, 3, 4]. Thus, the Higgs mechanism and the existence of the Higgs boson was postulated. It took almost 50 years to justify this magic trick. In July 2012 the discovery of a new particle that is consistent with the Higgs boson, the last missing piece of puzzle of the SM at that time, was announced by the ATLAS and CMS Collaborations [5, 6].

1.1 The Standard Model of particle physics

Nowadays the SM is an extremely successful quantum field theory describing our current understanding of elementary particles and interactions which act between them. The mathematical formulation of the SM is based on the gauge group

$$\mathcal{G}_{\mathrm{SM}} = \underbrace{\mathrm{SU}(3)_C}_{\mathcal{G}_{\mathrm{QCD}}} \otimes \underbrace{\mathrm{SU}(2)_L \otimes \mathrm{U}(1)_Y}_{\mathcal{G}_{\mathrm{EW}}}, \tag{1.1}$$

which represents the symmetry group of the SM Lagrangian. The $\mathrm{SU}(3)_C$ represents the strong interaction that acts on colour charge C. The electroweak interaction that act on the left-handed fermions L and the weak hypercharge Y is described by the group $\mathrm{SU}(2)_L \otimes \mathrm{U}(1)_Y$. The symmetry group is related to the conservation law. Therefore, only processes that satisfy conservation of colour charge, electric charge and weak isospin are allowed and can be described by the SM. This theory postulates 17 elementary particles which are summarized in Table 1.1. According to the spin which they carry, they are divided in two categories: fermions, which carry half-integer spin and bosons with integer spin. Fermions are the basic building blocks of the matter, while the bosons are carriers of the interactions.

		Name	Symbol	Generation	Spin	Charge Q/e	Mass
Fermions	Quarks	up	u	I	1/2	+2/3	$2.3^{+0.7}_{-0.5}$ MeV
		down	d	I	1/2	-1/3	$4.8^{+0.5}_{-0.3}$ MeV
		charm	c	II	1/2	+2/3	95 ± 5 MeV
		strange	s	II	1/2	-1/3	1.275 ± 0.025 GeV
		top	t	III	1/2	+2/3	4.18 ± 0.03 GeV
		bottom	b	III	1/2	-1/3	173.34 ± 0.76 GeV
	Leptons	electron	e	I	1/2	-1	0.511 MeV
		electron neutrino	ν_e	I	1/2	0	< 2 eV
		muon	μ	II	1/2	-1	105.66 MeV
		muon neutrino	ν_μ	II	1/2	0	< 0.19 MeV
		tau	τ	III	1/2	-1	1776.82 ± 0.16 MeV
		tau neutrino	ν_τ	III	1/2	0	< 18.2 MeV
		Name	**Symbol**	**Interaction**	**Spin**	**Charge Q/e**	**Mass / GeV**
Bosons		gluon	g	strong	1	0	0
		photon	γ	EM	1	0	0
		W boson	$W^\pm$	weak/EM	1	± 1	80.385 ± 0.015
		Z boson	Z	weak	1	0	91.1876 ± 0.0021
		Higgs boson	H	weak/self	0	0	125.7 ± 0.4

Table 1.1 – Elementary particles of the Standard Model with three generation of fermions, gauge bosons and the Higgs bosons. Values of spin, electric charge, and mass is given for each particle.

1.1.1 Matter

In total, the SM consists of 12 fermions, which can be divided into 6 quarks and 6 leptons. All fermions have the corresponding antiparticle with the same mass but the quantum numbers with opposite sign. They can be further categorised into three generations according to the mass hierarchy and the relation between their electric charge and weak isospin. For quarks, each generation is made of up-type quarks (up, charm, top) which carry positive electric charge $Q = {}^2/_3\, e$ and down-type quarks (down, strange, bottom) with electric charge of $Q = -{}^1/_3\, e$. The weak isospin T_3 of the up-type quarks is equal to $+{}^1/_2$, while the down-type one has $T_3 = +{}^1/_2$. In addition, each quark carries one of the colour; red, blue and green. The first generation is composed from the lightest quarks (up, down) and the heaviest quarks (bottom, top) are comprised in the third generation. The remaining two quarks, relatively light strange quark with mass of 95 ± 5 MeV and heavier charm quark with mass of 1.275 ± 0.025 GeV form the second generation. In nature, the quarks form colourless objects, hadrons. The hadrons with two constituents are called mesons, while baryons are built of three quarks. The quark structure of hadrons is then used to define various quantum numbers like strangeness, charmness, bottomness to classify them. For example, the ϕ meson formed by strange quark-antiquarks pair belongs to the group of hadrons with hidden strangeness, while the Ω (sss) is a baryon with open strangeness.

For leptons, each generation is formed by electrically charged leptons, namely electron (e), muon (μ), tau (τ) and the corresponding neutrally charged partner, neutrino (ν_e, ν_μ, ν_τ). The masses of leptons varies from 0.511MeV for the electron to 1.777GeV for the tau. Since the muon is lighter than the lightest hadron, only tau can decay hadronically. Within the SM the neutrinos are assumed to be massless, however the observed neutrino oscillation

indicates non-zero masses. The neutrino oscillation is a phenomenon where neutrinos with one flavour can convert into neutrinos of a different flavour. It was proposed in 1957 [7] but for the first time observed in 1998 by the Super-Kamiokande experiment [8].

The observed mass hierarchy of fermions results in instability of heavier fermions since they can decay into the lighter one. It explains why ordinary matter is formed by fermions from the first generation.

1.1.2 Interactions

The SM includes three of four fundamental interactions, namely the strong, electromagnetic and weak. The interactions are realized by exchange of gauge bosons. The strong interaction is described by Quantum Chromodynamics, while the electromagnetic and weak interactions are formulated within the unified electroweak theory.

1.1.2.1 Quantum Chromodynamics

The theory describing the strong interaction that act between quarks is Quantum Chromodynamics (QCD). It is constructed as a non-Abelian $\mathrm{SU}(3)_C$ gauge theory. In QCD, the colour charge is the charge of the strong interaction. Since each quark carries one of the colours, it can take part to the strong interaction. The generator of $\mathrm{SU}(3)$ group is the coloured gluon octet with the following possible colour combination

$$r\bar{b}, r\bar{g}, b\bar{r}, b\bar{g}, g\bar{r}, g\bar{b}, \frac{1}{2}\left(r\bar{r} - b\bar{b}\right), \frac{1}{6}\left(r\bar{r} + b\bar{b} - 2g\bar{g}\right). \tag{1.2}$$

These gluons are massless spin-1 particles that act as mediators of the strong interaction. Since gluons carry the colour charge, they can interact with each other or even themself. The self-coupling of gluons causes several spectacular phenomena in the QCD. One of them is the colour anti-screening, which is in strong contrast to the Debye screening of electric charge known from the electromagnetism. It leads to the asymptotic freedom of quarks and the running of the coupling of the strong interaction α_s discussed later.

The Lagrangian of QCD can be written as

$$\mathcal{L}_{\mathrm{QCD}} = \sum_f \bar{\psi}_i^{(f)} \left(i\gamma_\mu D_{ij}^\mu - m_f \delta_{ij}\right) \psi_j^{(f)} - \frac{1}{4} F_{\mu\nu}^a F_a^{\mu\nu}. \tag{1.3}$$

Here γ_μ are the Dirac γ-matrices. The $\psi_j^{(f)}$ is quark spinor for a quark of flavour f with mass m_f and colour index j. The covariant derivative D_{ij}^μ is defined as

$$D_{ij}^\mu = \partial_\mu \delta_{ij} + i g_s t_{ij}^a A_a^\mu \tag{1.4}$$

and acts only in colour space. The g_s is strong coupling and is connected with α_s via the relation $\alpha_s = g_s^2/4\pi$. Finally, the A_a^μ is the gluon field running over all generators t_{ij}^a of $\mathrm{SU}(3)$. The field-strength tensor $F_{\mu\nu}^a$ is derived from the gluon field and can be formulated as

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - g_s f_{abc} A_\mu^b A_\nu^c, \tag{1.5}$$

where f_{abc} are the structure constants of the SU (3) group. As one can see from the lagragian, the masses of quarks and the strong coupling α_s are the parameters of QCD.

As a direct consequence of the self-coupling of the massless gluons, the concept of the running of the coupling α_s arises. The coupling of the strong interaction is a function of the momentum transfer Q^2 and its dependence is given by

$$\alpha_s\left(Q^2\right) = \frac{12\pi}{(33 - 2n_f)\ln\frac{Q^2}{\Lambda_{\mathrm{QCD}}}}, \tag{1.6}$$

where n_f is the number of quark flavour and $\Lambda_{\mathrm{QCD}} \sim 200\,\mathrm{GeV}$ is the QCD scale. The value of Λ_{QCD} needed to be experimentally determined and represents the scale, from where on perturbative QCD calculation can be employed. At large Q, which corresponds to the small distance, α_s decreases and the quarks can be considered as a quasi-free. This behaviour is known as asymptotic freedom. Figure 1.1 shows the running of α_s as function of Q. The world average of α_s at the momentum transfer equivalent to the mass of a Z boson is measured to be $\alpha_s\left(m_{\mathrm{Z}}\right) = 0.1185 \pm 0.0006$ [9].

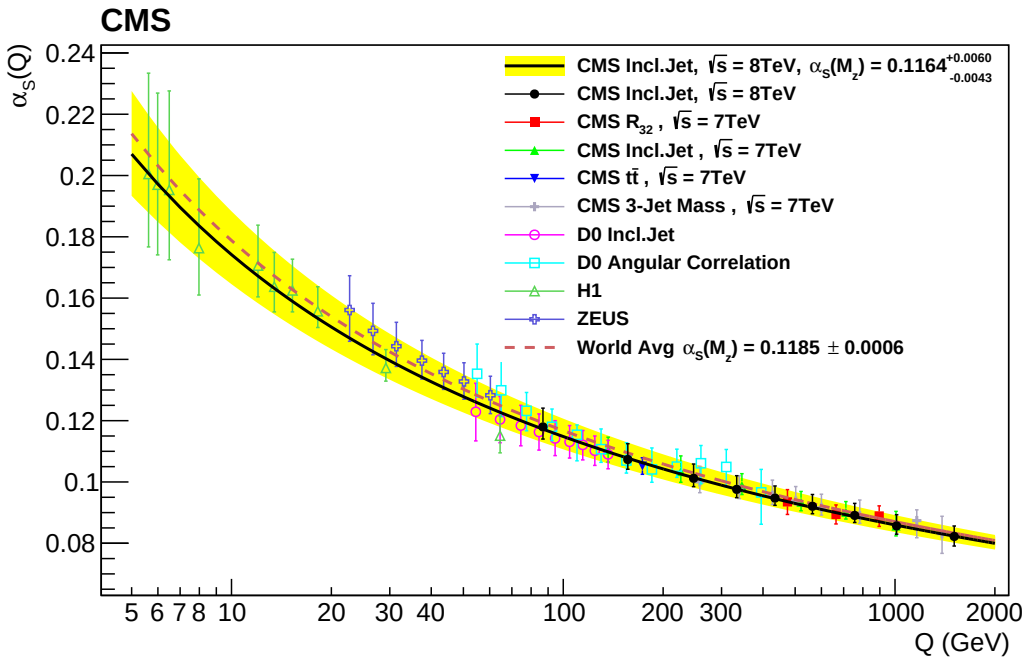

Figure 1.1 – The running of strong coupling $\alpha_s\left(Q\right)$ as a function of the scale Q determined using data from CMS and other collider experiments [9].

1.1.2.2 Electroweak theory

Already in 1860s, J.C.Maxwell formulated a set of partial differential equations that coupled electric and magnetic forces. By that he founded classical electromagnetism. The electromagnetic interaction has many important consequences on a macroscopic level of our nature. It is responsible for forming of atoms the nucleus and electrons.

The formulation of electromagnetic interaction as a relativistic field theory is provided by Quantum Electrodynamics (QED) with a symmetry group U (1). In this theory, the mediator is the massless photon γ, that acts between particles with electric charge. Thanks

to the massless mediator, the range of the electromagnetic interaction is infinite. In contrast to QCD, with increasing distance the strength of interaction decreases.

The weak interaction is responsible for flavour-changing phenomena like radioactive β^- decay. According to the electric charge exchanged during the interaction, there are two types of the weak interaction. The charged current is mediated by the charged $W^{\pm}$ boson, while the neutral Z boson acts as mediator of the neutral current. In contrast to the photon, the $W^{\pm}$ and Z bosons are massive. Therefore the range of the weak interaction is not infinity, but only 10^{-16} m. The charged current enables the interaction of two quarks with different flavour, for example the β^- decay of the neutron. In this process, one of the down quark of the neutron transforms into an up quark and a W^- boson, which then decays into the electron and the electron antineutrino as shown in Figure 1.2.

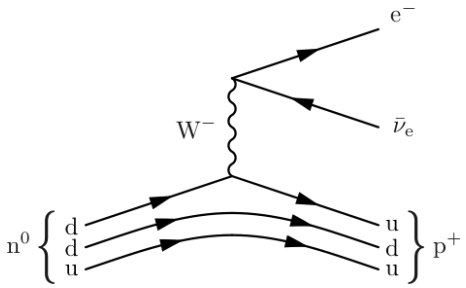

Figure 1.2 – Feynman diagram of the β^- decay of neutron.

This interaction is explained by the mixing mechanism and it means that the weak eigenstates of the quarks are different from the their mass eigenstates. Therefore quarks from the second and third generation may decay into a quark of a lower generation. The probability of such transition is quantified by the Cabbibo-Kobayashi-Maskawa (CKM) matrix

$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}, \tag{1.7}$$

where V_{ij} is the relative coupling of quark i to quark j. The probability of the transition is given by the amplitude of the elements, $|V_{ij}|^2$. The diagonal elements are close to unity, which indicate the higher probability of the transition within the same generation. The off-diagonal elements are small. However, the SM does not provide any prediction for the elements of the CKM matrix and they need to be determined experimentally. The only constraint is arising from the fact, that by construction the CKM matrix is unitary, therefore its free parameters are constrained to three mixing angles and the complex phase. The observed CP violation implies the complex phase to be different from zero.

The electroweak unification

From a historical point of view, the β^- decay of the neutron is the oldest and best known example of a process caused by the weak interaction. It is also a good starting point of the road towards the unification of the electromagnetic and weak interaction. The first theory of β^- decay was formulated by E. Fermi in 1934 [10]. It describes the β^- decay of the neutron as the point-like interaction of four fermions with spin 1/2. At that times, the Fermi's theory described the weak interaction quite well. In addition, the leptonic current in Fermi's effective Lagrangian had a pure Lorentz vector and axial vector

V-A structure that implies the parity violation in weak interaction, which was also experimentally observed. Despite of being very successful and having predictive power in 1950s, the Fermi-type theory faced several difficulties. Mainly the cross section grows quadratically with the centre-of-mass energy, which makes perturbative renormalization of this theory impossible. Sometimes referred as the weak interaction is becoming strong at high-energies. The solution of that issue was replacing the Fermi-type theory by the theory involving a dimensionless coupling constant and mediated by exchange of the intermediate vector boson $W^{\pm}$ decaying into lepton and the corresponding neutrino. As these vector bosons carry electric charge, it was immediately proposed to incorporate its electromagnetic interactions. One could consider such an attempt as the beginning of the electroweak unification. However, for various reasons, e.g. the divergence of the cross section of $\nu\bar{\nu} \to \mathrm{W}_\mathrm{L}^+\mathrm{W}_\mathrm{L}^-$ does not vanish for an arbitrary scattering angle, also this theory was not the sufficient.

In the beginning of 1960s, based on studies of Yang-Mills theories, the non-Abelian gauge symmetries were believed to be suited to control the desirable divergence cancellation. Although the proof of such statement about the properties of the non-Abelian gauge symmetries was not available yet, these proposals led Glashow [11] and Weinberg in cooperation with Salam [12, 13] to formulate the theory that unifies the weak and electromagnetic interaction on the basis of a non-Abelian gauge symmetry.

The electroweak theory unifies the electromagnetic and weak interactions. The formulation of this theory is based on non-Abelian $\mathrm{SU}(2)_L \otimes \mathrm{U}(1)_Y$ symmetry groups. Each of the symmetry group has its own generators and involves different couplings. The formulation of the electromagnetic interaction is based on $\mathrm{U}(1)_Y$ which gives rise to a single massless gauge boson B and coupling constant e corresponding to the electromagnetic coupling strength. The $\mathrm{SU}(2)_L$, describing the weak interaction, introduces three massless gauge boson W_1, W_2, W_3 and the coupling constant g which were later identified as the coupling constant of weak interaction. The physical bosons which are identified as $\mathrm{W}^{\pm}$, Z bosons and γ are obtained by mixing of these states and given as

$$\mathrm{W}^{\pm} = \frac{1}{\sqrt{2}} \left(W_1 \mp \mathrm{i} W_2\right) \tag{1.8}$$

and

$$\begin{pmatrix} \gamma \\ Z \end{pmatrix} = \begin{pmatrix} \cos\left(\theta_W\right) & \sin\left(\theta_W\right) \\ -\sin\left(\theta_W\right) & \cos\left(\theta_W\right) \end{pmatrix} \cdot \begin{pmatrix} B \\ W_3 \end{pmatrix}. \tag{1.9}$$

Here the mixing represents a rotation by the weak mixing angle θ_W, the so-called Weinberg angle [13]. The mixing angle relates the coupling constants of the electromagnetic and weak interaction by $\sin\left(\theta_W\right) = {}^{e}/_{g}$ and is measured to be $\sin^2\left(\theta_W\right) = 0.23122 \pm 0.00004$ [14]. The corresponding electric charge of the bosons is given by a relation, which connects the weak isospin T_3 and the hypercharge via

$$Q = T_3 + \frac{Y}{2}. \tag{1.10}$$

This construction gives the weak gauge bosons which carry the weak charge and therefore can interact with each other.

Spontaneous Symmetry Breaking

According to the theory proposed by Glashow, Weinberg and Salam, no mass term is present in the Lagragian. Therefore the gauge bosons as well as fermions should be massless, however from the experimental observation this is not true. In the SM, this contradiction is solved by the spontaneous symmetry breaking of the electroweak symmetry postulated by Brout, Englert [2] and Higgs [3, 4] in 1964. This mechanism introduces a new scalar field ϕ with the corresponding potential $V(\phi)$, later known as the Higgs field and Higgs potential. The Higgs field is constructed as a single weak isospin doublet of two complex scalar field, formally written as

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \tag{1.11}$$

The peculiarity of the Higgs potential $V(\phi)$ is its symmetric shape with respect to the origin as illustrated in Figure 1.3 and given by following formula

$$V(\phi) = \mu^2 \left(\phi^\dagger \phi\right) + \lambda \left(\phi^\dagger \phi\right)^2, \tag{1.12}$$

with $\mu^2 < 0$ and $\lambda > 0$. The choice of $\lambda > 0$ ensures one minimum and $\mu^2 < 0$ shifts the minimum from the origin.

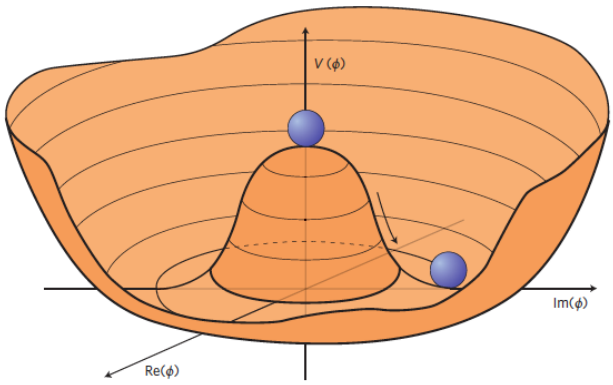

Figure 1.3 – Two-dimensional sketch of the Higgs potential $V(\phi)$ [1].

Then the symmetry of $\mathrm{SU}(2)_L \otimes \mathrm{U}(1)_Y$ is spontaneously broken by minimizing the Higgs potential for non-vanishing state of the Higgs field. Such a state of the Higgs field is given as

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \text{ with } v = \frac{2\mu}{\sqrt{\lambda}}, \tag{1.13}$$

where v is the vacuum expectation value (VEV), which depends on λ and μ. For perturbative calculation, the expansion of the Higgs field around the VEV in radial direction with the parametrization

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.14}$$

is required. Finally the real Higgs field arises from inserting Eq. 1.14 in Eq. 1.12, where the Higgs boson mass is given as $m_H = -2\mu^2$. Within the SM, there is no prediction for the Higgs mass as μ is not determined by any of the known parameter of SM. The weak gauge bosons interact with the Higgs field by that they gain mass. As shown in Ref. [1], the masses of the gauge bosons are given as

$$m_W = \frac{1}{2}gv \text{ and } m_Z = \frac{v}{2}\sqrt{g^2 + e^2}. \tag{1.15}$$

In both cases, the mass of the gauge boson is proportional to v. By using the relation for the Weinberg angle θ_W, the masses of $W^{\pm}$ and Z bosons are tied in via

$$m_W = m_Z \cos(\theta_W). \tag{1.16}$$

Finally, the couplings of the fermions to the Higgs field generate their masses. For each fermion, the mass is given as

$$m_f = y_f \frac{v}{\sqrt{2}}. \tag{1.17}$$

Here y_f is the so-called Yukawa coupling constant and characterizes the strength of the interaction between the fermion and the Higgs field. As for the Higgs mass, there are no predictions about the value of y_f within SM.

Despite the confirmation of the electroweak theory and the spontaneous symmetry breaking took a very long time, already a few years later after its formulation, there were a few hints indicating that it could be the proper theory. In 1971 G. 't Hooft formulated a rigorous proof of perturbative renormalizability of non-Abelian gauge models [15, 16], which were incorporated in the formulation of the Higgs mechanism. In addition, the discovery of weak neutral currents [17] in 1973 pointed towards models with a neutral gauge boson. The discovery of new particles with properties that are consistent with the $W^{\pm}$ and Z boson was announcement by the UA1 Collaboration at the SPS collider at CERN in 1983 [18, 19]. The search for the Higgs boson took much more longer, in total 48 years.

1.2 Proton-proton collisions as probe of the Standard model

The scattering experiments are the most powerful tools to test the SM and perturbative QCD (pQCD). They operate at the collider, where the colliding particles are accelerated and collided. New particles, that are produced, are then recorded by the detector. Within the context of this thesis, proton-proton (pp) collisions will be of particular interest and therefore they will be discussed in detail. Apart from the pp collisions, the scattering experiments involving electron-positron, electron-proton, neutrino-proton scattering or proton-antiproton and heavy ion collisions can serve as tests of pQCD. Different scattering experiments are suitable for different physics processes. The electron-positron collisions are well suited for the production and study of Z and H bosons, while the hadron-hadron collisions are more convenient for processes initiated by gluon-gluon, quark-quark or quark-gluon interactions. Finally, the heavy ion collisions are considered as a tool to create the so-called quark-gluon plasma and to study the hadronic matter under extreme conditions.

1.2.1 Structure of protons and the factorization theorem

The proton is a composed object of elementary particles. These particles are the three valance quarks (uud) embedded in the sea of gluons and quark-antiquark pairs. The number of partons in the proton depends on the scale probed by the momentum transfer Q^2 of the scattering process. The probability of finding the parton i in the proton that carries a fraction of the total proton momentum x_i is given by the parton distribution functions (PDFs) defined as $f\left(x,Q^2=\mu^2\right)$, often used the momentum weighted $xf\left(x,Q^2=\mu^2\right)$. These functions are believed to being universal, i.e. they can be determined from one physical processes and then they can be used to predict the cross section of any other process applying the factorization theorem. However, they depend on the momentum transfer and parton flavour of the involved partons. Figure 1.6 shows an example of PDFs of the NNPDF3.1 set with $\alpha_S = 0.118$. For the region of low momentum transfer $Q^2 = \mu^2 = 10\,\text{GeV}^2$ as shown in Figure 1.6 (right), there is a larger probability to find valance quarks that carry a large fraction of the total proton momentum. By increasing the momentum transfer, the contribution of sea partons becomes more important. At all scales, the contribution from the sea partons is dominant in low$-x$ region. The gluon PDF rapidly increases with decreasing x. Compared to light quark PDF, the heavy quark PDF is smaller. Once the PDF is known at an initial scale, it is possible to evaluate it to higher scale by using Dokshitzer-Gribov-Lipatov- Altarelli-Parisi (DGLAP) [20, 21, 22] evolution equations.

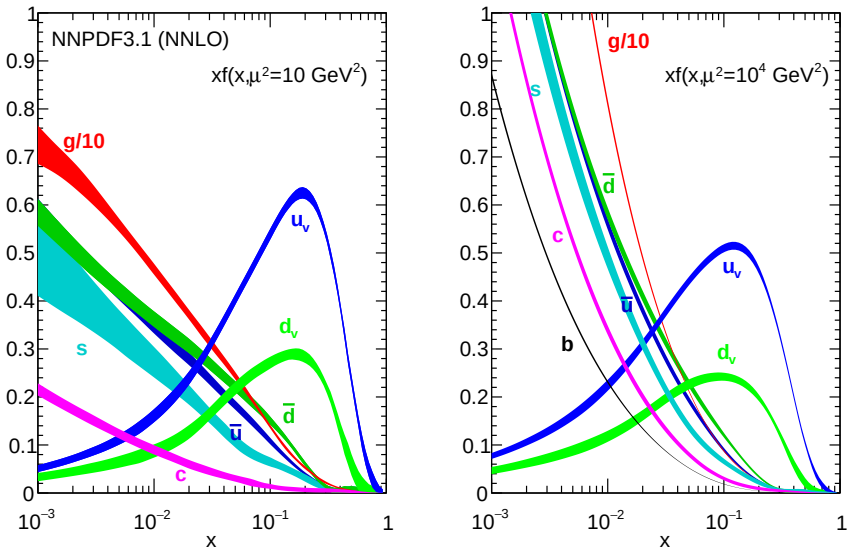

Figure 1.4 – The momentum weighted parton distribution functions at next-to-next-leading order obtained by NNPDF Collaboration [23] at the scale $Q^2 = \mu^2 = 10\,\text{GeV}^2$(left) and $Q^2 = \mu^2 = 10^4\,\text{GeV}^2$(right).

1.2.1.1 Collinear factorization

The collisions of two protons can be seen as an scattering of two partons 1 and 2 carrying the fraction of the proton momenta x_1 and x_2, respectively. In collinear factorization, where the transverse momenta of the incoming partons are considered to be negligible, the four-momenta of parton can be written as

$$
\begin{aligned}
p_1 &= \frac{\sqrt{s}}{2}\,(x_1,0,0,x_1)\,, \\
p_2 &= \frac{\sqrt{s}}{2}\,(x_2,0,0,-x_2)\,,
\end{aligned}
\tag{1.18}
$$

where s stands for the square centre-of-mass energy of the pp collisions, i.e. $s = (P_1 + P_2)^2$. Consequently, the square centre-of-mass energy of parton-parton cross section is given as

$$
\hat{s} = (p_1 + p_2)^2 = x_1 x_2 s = Q^2. \tag{1.19}
$$

Figure 1.5 shows an illustration of the production of a vector boson in pp collision. The meaning of symbols is consistent with the discussion above. The produced vector boson V then decays into pair of 2 fermions f.

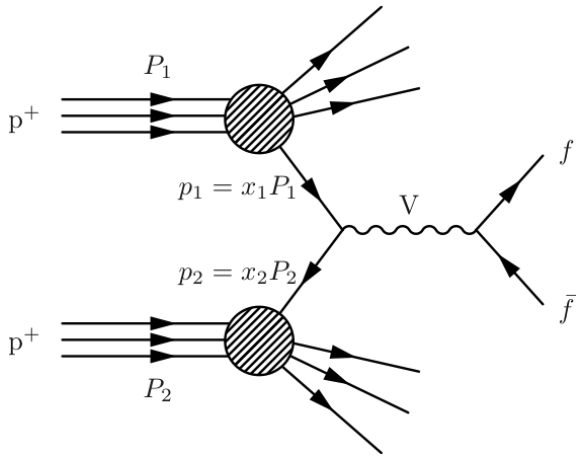

Figure 1.5 – Illustration of the proton-proton collision and the production of a vector boson decaying into two fermions.

According to the factorization theorem, any hadron-hadron cross section can be written as a product of two non-perturbative PDFs and the perturbative parton-parton cross-section $\hat{\sigma}$. For our example shown in Figure 1.5, the cross section is formally written as

$$
\sigma_{p_1 p_2 \to X} = \sum_{i,j} \int_0^1 \mathrm{d}x_1 f_{i,p_1}\,(x_1,\mu_\mathrm{F}) \int_0^1 \mathrm{d}x_2 f_{j,p_2}\,(x_2,\mu_\mathrm{F})\,\hat{\sigma}_{ij\to X}\,(x_1 P_1, x_2 P_2, \alpha_\mathrm{s}, \mu_\mathrm{F})\,. \tag{1.20}
$$

Here the sum runs over all indices i,j representing all possible parton types in the incoming proton and μ_F stands for the factorization scale. Within the framework of collinear factorization, we are able to calculate the cross section for a wide range of processes at various scales. However, there are some limitations arising from the assumption about the neglection of transverse momenta of the incoming partons. For processes with low momentum transfer and/or in the low-x region, this assumption is not valid anymore. In low-x region, where the gluons dominate, the longitudinal momenta of partons are not larger than its transverse momenta. This fact is incorporated into the generalized parton distributions (GPD) and the transverse momentum dependent distributions (TMD) PDF,

where the transverse momenta of partons are considered in the calculation of the PDF and the partonic cross section.

The GPDs encode the transverse position of a parton in a nucleon. In contrast to the PDF, they are not interpreted as parton densities, but as probability amplitudes.

1.2.1.2 Transverse momentum dependent PDFs and k_t factorization

In order to obtain TMD PDFs in a wider kinematic range, the parton branching (PB) method can be employed [24, 25]. The PB is a novel method to numerically solve the DGLAP evolution equations. In an iterative procedure, it uses the concept of resolvable and non-resolvable branching and applies Sudakov form factors (discussed later) to describe the probability for non-resolvable branchings from one evolution scale to another. The kinematics of the splitting are known in each step, and therefore it enables us to extract the TMDs. A more detailed description of the PB method is discussed in Ref [24, 25]. Sometimes the TMD PDFs are called as unintegrated PDF(uPDF). The integration of the TMD over the transverse momentum results in PDF. Formally written as

$$f_{i,p_1}\left(x_1,\mu_F\right) = \int \mathrm{d}^2\mathbf{k_t} A_{i,p_1}\left(x_1,\mathbf{k_t},\mu_F\right), \tag{1.21}$$

where the $A_{i,p_1}\left(x_1,\mathbf{k_t},\mu_F\right)$ is TMD PDF. Here $\mathbf{k_t}$ is the euclidean transverse momentum vector. Figure 1.6 shows TMDs obtained by PB method from fit the deep inelastic scattering (DIS) measurements from HERA [26]. Figure 1.6 (left) shows the TMD for up and down quark as function of their transverse momentum for $x = 0.01$ at the scale $\mu = 10$GeV. The TMD is shown as function of x for fixed $\mathbf{k_t} = 5$ GeV in Figure 1.6.

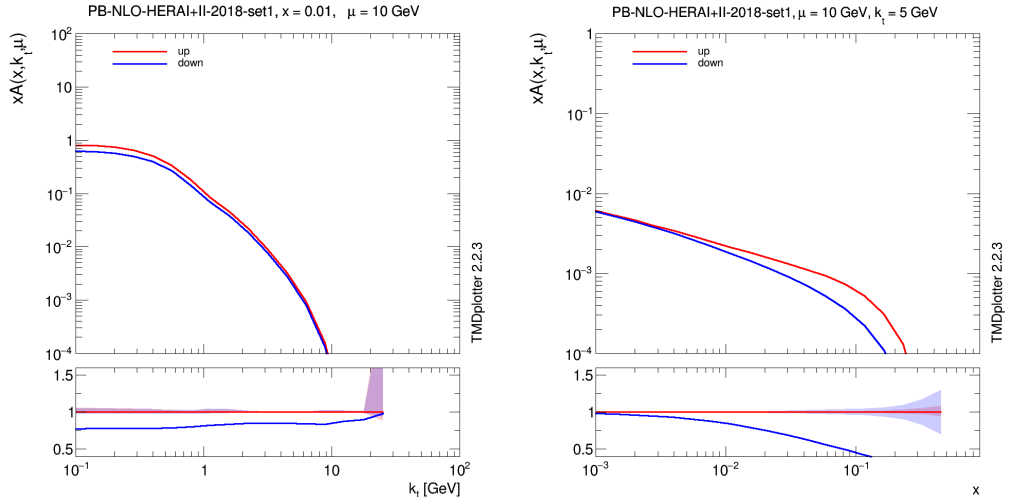

Figure 1.6 – Transverse momentum dependent parton distribution function [27] for up and down quark as function of their transverse momentum $\mathbf{k_t}$ for $x = 0.01$ at the scale $\mu = 10$ GeV (left) and as function of the momentum fraction x for fixed $\mathbf{k_t} = 5$ GeV(right).

Within the framework of TMD PDFs and k_t factorization, any hadron-hadron cross section

is given as

$$\sigma_{p_1p_2\to X} = \sum_{i,j} \int_0^1 \mathrm{d}x_1 \int \mathrm{d}^2\mathbf{k_{t1}} A_{i,p_1}\left(x_1,\mathbf{k_{t1}},\mu_\mathrm{F}\right) \cdot \int_0^1 \mathrm{d}x_2 \int \mathrm{d}^2\mathbf{k_{t2}} A_{i,p_2}\left(x_2,\mathbf{k_{t2}},\mu_\mathrm{F}\right) \hat{\sigma}_{ij\to X}\left(x_1P_1,x_2P_2,\alpha_\mathrm{s},\mu_\mathrm{F}\right). \tag{1.22}$$

In addition, the TMD PDF can encode information about the polarization degree of freedom and can be related to the longitudinally polarized and transversely polarized quark parton distributions and Sivers function [14]. This is especially important for spin physics, where e.g. the origin of spin of proton is studied. In context of the measurement with vector bosons, they are important to describe the production of the vector bosons with the low transverse momenta, where soft-gluon perturbative resummations and nonperturbative contributions play a role. Recently, it was shown that the TMD PDF enables a good global description of Drell-Yan production with low transverse momenta over wide range of the collision energies [28].

1.2.2 Matrix elements

The cross section calculation involves the parton-parton cross section $\hat{\sigma}$ that mathematically formulated is the probability amplitude of the transition from the initial state into the final state integrated over the phase space for the final state partons. The cross section with the final-state particle X with k additional partons can be formally written as

$$\sigma_\mathrm{F} = \sum_{k=0}^{N_\mathrm{legs}} \int \mathrm{d}\Phi_{X+k} \left| \sum_{l=0}^{N_\mathrm{loops}} \mathcal{M}^{(l)}_{X+k} \right|^2, \tag{1.23}$$

where Φ_{X+k} is the phase space of the final state particle X with k additional partons (legs) and $\mathcal{M}^{(l)}_{X+k}$ is the matrix elements with N_loops loops for the given leg l. A leg is a real correction to the cross section as it produces a new particle in the final state, while the loops are virtual corrections, i.e. they enter into the calculation but they do not produce any new particles in the final state. Following the choice of number legs and loops, namely k and l, a certain processes enter into the calculation of the total parton-parton cross section. Thus, according to the value of k and l, the final state processes are recognized as:

- $k = 0, l = 0$: production of the final state particle X at the leading order(LO),
- $k = n, l = 0$: production of the final state particle X with n additional partons at LO,
- $k + l \leq n$: production of the final state particle X at the next^n-to-leading order (N^{n}LO) which also includes processes with additional partons, namely from like X with one additional parton at N^{n-1}LO until X with n additional partons at N^{n-n}LO.

Figure 1.8 shows a Feynman diagram, that graphically represents the ME. The production of a Z boson with various number of additional partons are shown. The produced Z boson then decays into the quark-antiquark pair. The first two plots show the production at LO. The difference between these two diagrams is the exchanged virtual gluon by the incoming quarks shown in the second plot. The Z boson production at NLO and NNLO is shown in third and fourth plot, respectively.

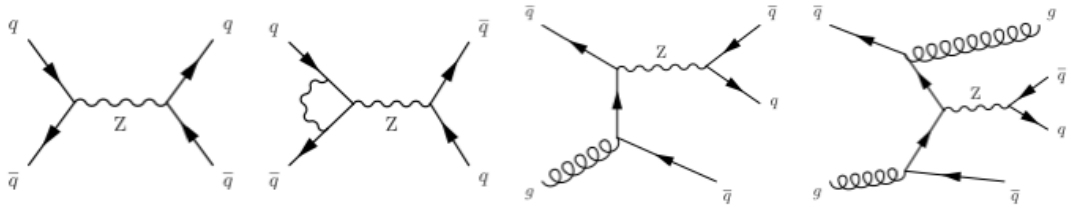

Figure 1.7 – Feynman diagrams of the Z boson production in proton-proton collisions at various order. From left to right: leading order, leading order with exchange of virtual gluon, next-to-leading order, and next-to-next-leading order.

1.2.3 Parton shower

The high energetic partons can loose energy by the radiation of extra partons. This phenomenon viewed as parton splitting is described by the concept of parton shower, where it is treated as an approximation of higher-order real emission corrections to the hard scattering. At LO only four different splittings of partons are possible, namely $q \to qg$, $q \to gq$, $g \to gg$, and $g \to q\bar{q}$. Each of the splitting is expressed by the DGLAP splitting function, that can be interpreted as the probability of parton j with momentum fraction ξ to emit a parton and become parton i with a momentum fraction x. The evolution of a parton from the initial scale t_0 to the final scale t is then described by DGLAP evolution equation and the probability of no-branching between these two scales and splitting at the scale t is given by

$$P_a(t) = \Gamma_a(t) \exp\left(-\int_{t_0}^{t} \mathrm{d}t' \Gamma_a(t')\right). \tag{1.24}$$

Here $\exp\left(-\int_{t_0}^{t} \mathrm{d}t' \Gamma_a(t')\right)$ is the Sudakov form factor. With every new splitting, the partons loose energy, thus the scale Q of the new splitting is smaller and therefore α_S increases. The splitting of partons lasts until a certain scale Q_{min}, where the pQCD breaks down. At this scale, the hadronization of the parton begins. The parton shower can be initialized by the incoming as well as outcoming partons. In the first case, we talk about the initial-state radiation (ISR), the latter case is referred as the final-state radiation (FSR). Since the incoming partons momenta are already fixed and used in the ME calculation, ISR is handled by the backward evolution.

1.2.4 Hadronization and decay

The hadronization typically starts at the scale of $Q_0^2 \sim 1\,\mathrm{GeV}^2$ where pQCD does not work anymore. The rising of α_{S} makes impossible further perturbative parton splitting, but the gain in energy leads to creation of a new pair of partons and the original parton looses some of its energy. The hadronization lasts until no energy is left to create a new pairs of partons. At the end of this process, the partons form colourless objects, hadrons. As the hadronization is nonperturbative process, the analytical calculation cannot be easily performed. There are two common phenomenological models describing the hadronization: the Lund fragmentation model [29] and the cluster model [30]. The formed hadrons are usually unstable. Thus the sequential decay of these hadrons then occurs.

1.2.5 Underlying event

The pp collisions does not consist only from the hard interaction, but it also includes additional soft particle production from the proton remnants. In addition, multiparton interaction can occur, if the probability for interaction is large enough. These contributions are commonly referred to as underlying event (UE). Similarly to the hadronization, UE cannot be analytically calculated.

1.2.6 Monte Carlo event generators

For a realistic simulation of pp collisions, the Monte Carlo (MC) event generators have to properly treat all the individual steps that were discussed earlier. The complexity of such collisions is shown in Figure 1.8.

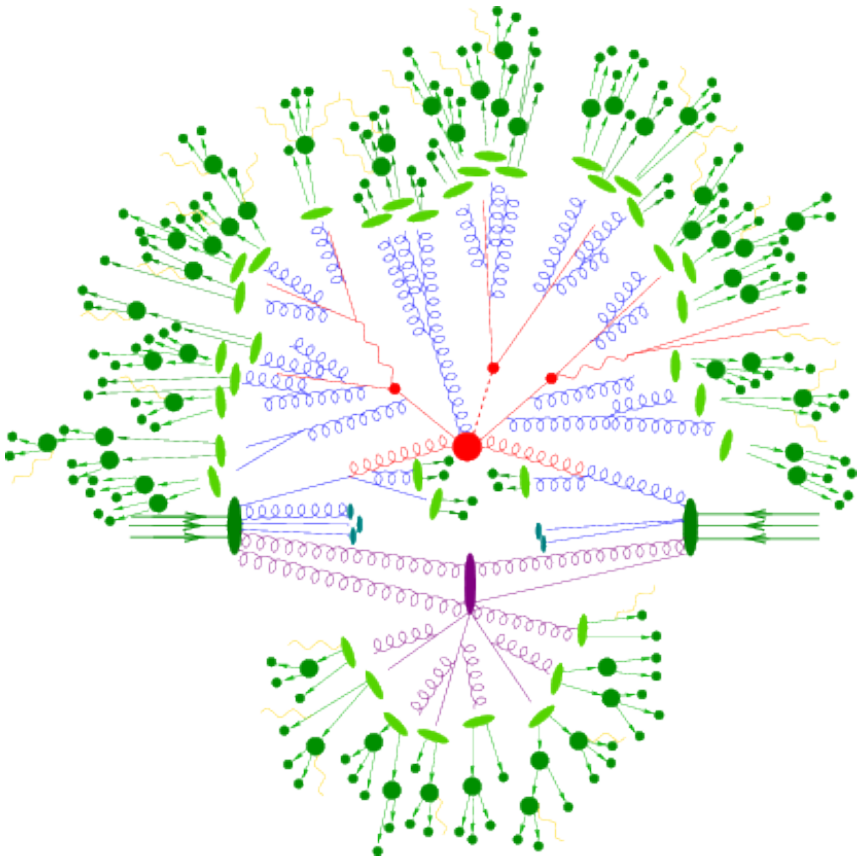

Figure 1.8 – Schematic illustration of the MC simulation [31] of proton-proton collision. The hard collisions is represented by the red blob in the centre and is surrounded by the parton shower. The purple marker indicate a secondary scattering event representing by the multiple parton interactions. Hadronization, the transition from parton to hadron, is represented by light green blobs. The dark green markers stands for hadron decays.

Each MC generator needs a certain physics parameters as input. The exact value of these parameters is of concern of tuning, where the parameters are adjusted to reach a reasonable agreement with data. The extracted set of parameters are then called as tune. Typically, the PDF set and corresponding parameters for the non-perturbative effects, namely those related to UE, are of particular interest for tuning, while the pQCD are well analytically calculated without any tuning. In addition, the matching scale between

the matrix element and the PS is not arbitrary. The PS evolution can already include the hard radiation of parton, which can be equivalent to the phase space configuration from the ME calculation. Therefore it is needed to avoid such case as it would lead to double counting of the same phase space configuration. There are many methods how to do it, but most popular and widely used are the MLM [32] and FxFx [33] matching. The former one is designed for the LO ME, while the latter one is used for the NLO ME. The main principle of the both matching algorithms is quite similar and is based on the event veto, i.e. event is rejected if the parton from the PS is emitted above the scale of partons involved in the ME calculation. Since the FxFx matching is designed for the NLO ME calculation, it treats not only the real corrections, but also the virtual corrections.

From a large variety of the MC event generators, we listed a few of them, mainly those which will be later used.

- Pythia8 [34]: a LO $2 \to 2$ ME generator with higher-order corrections that are implemented by the PS in leading-log approximation (LLA). Pythia utilizes a PS that is p_{T}-ordered. This means that the emitted partons by ISR and FSR are ordered by decreasing transverse momentum. The hadronization is supplemented by the Lund String model.
- Herwig++/Herwig7 [35]: a LO as well as NLO event generator. Unlike to Pythia, Herwig uses the angular-ordering for the PS. It means that the emitted parton are ordered in θ, where θ is the angle between the emitted parton and the parent parton. In this case, the scale is given as $Q^2 = 2E_a^2\,(1 - \cos\theta)$, where E_a is the energy of the parent parton. The cluster model, that is based on nonperturbative gluon splitting, is employed for the modelling of the hadronization.
- MadGraph5_aMCatNLO [36]: a LO and NLO ME generator. The decay of the resonances that are involved in the ME is performed by MadSpin [37]. It does not simulate the PS nor UE. Thus the generated events need to be interfaced to Pythia or Herwig. The ME-PS matching uses the MLM or FxFx method according to the accuracy of the ME.
- Powheg [38]: a NLO $2 \to 2$ ME generator with one additional hard emitted parton. Similarly to MadGraph5_aMCatNLO, the generated events are then interfaced by Pythia8 or Herwig to model the PS and UE. In order to avoid double counting during ME-PS matching, the Powheg method [38] is used.
- Sherpa [31]: a LO and NLO MC generator. The tree and one-loop ME for processes at NLO QCD and NLO EW are calculated by OpenLoops [39]. In contrast to Powheg or MadGraph5_aMCatNLO, it also simulate the PS and UE. The hadronization is parametrized by the cluster model.
- Cascade [40]: a PS MC generator. It adds transverse momenta to initial partons according to the TMD-PDF and the ISR is done following the TMD, while the FSR is treated by Pythia6 [41]. A ME at LO or NLO are calculated by MadGraph5_aMCatNLO.

1.2.6.1 Detector simulation

Finally, the generated events are propagated through the detector to simulate the detector response and model the readout signals from subdetectors. The output of the detector simulation is stored in the same data format as real data. This procedure is referred

as a full detector simulations and the GEANT4 [42] software is employed. This procedure is very time and resource consuming, hence sometimes the simplified method, FAST SIMULATION [43], is preferred rather.

1.3 Shortcoming of the Standard model

Nowadays the SM is an extremely successful theory and almost all experimental measurements are consistent with SM predictions. Figure 1.9 shows the comparison of the total production cross section measurements with SM predictions. Even the cross section ranges over 13 order of magnitude, the measured values are consistent with SM predictions.

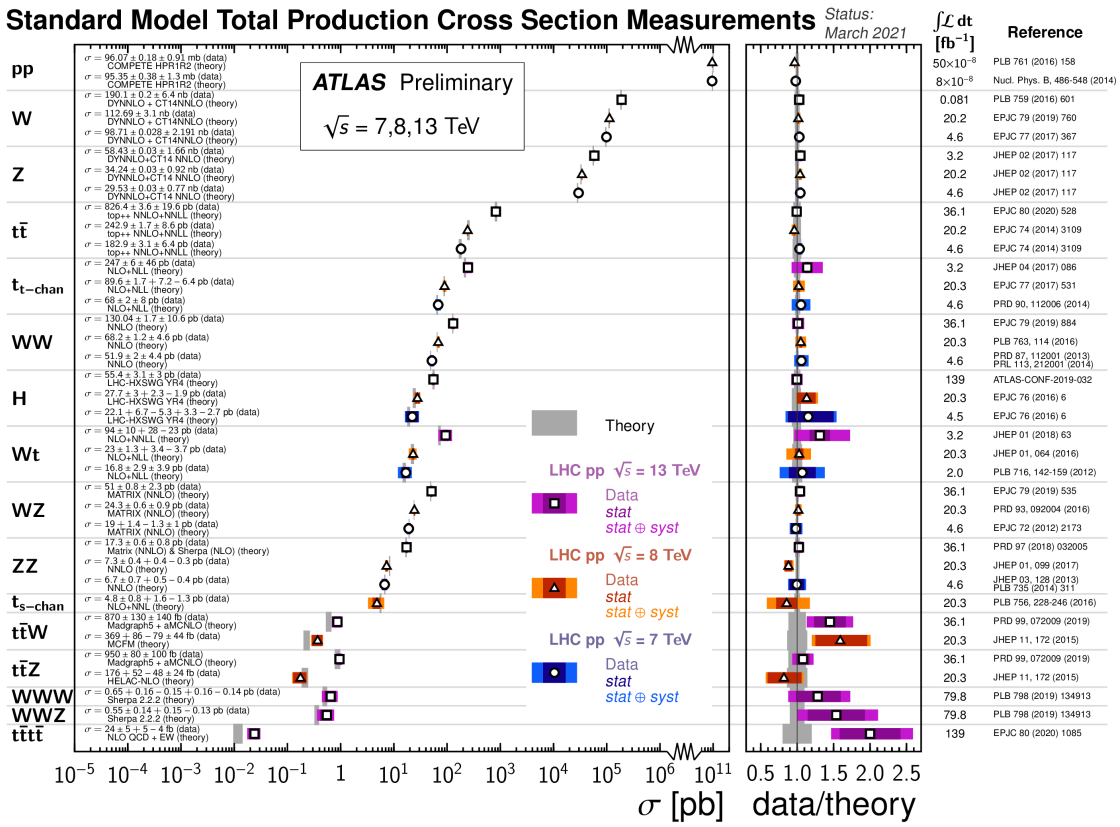

Figure 1.9 – Overview of Standard Model production cross section measurements by the ATLAS Collaboration [44].

Despite the very strong power of the SM as the theory not only in describing the experimental measurements, but also in predicting new phenomena and particles which waited to be discovered, there are certain shortcoming of SM. Some of the limitations arise from the experimental observations, mainly from the astroparticles physics. We can mention a few examples:

- Neutrino oscillations and mass: Within the SM, the neutrinos are believed to be massless particles. However in order to describe the neutrino oscillations, they need to have a non-zero mass.
- Gravity: Only three of four fundamental forces are incorporated in the SM. The SM does not describe the gravitation interaction. Its effects will become visible at the energies close to the Planck scale $M_P \approx 10^{19}$ GeV.

- Matter-antimatter asymmetry: After the Big Bang, when the universe was created, matter and antimatter is believed to be produced almost in equal amounts. Nevertheless nowadays matter dominates over antimatter and the effect arising of CP violation is too small for its explanation within the SM.

The other kind of limitations are more connected to conceptual problems and the relevance of these issues can be debated. Among these are:

- Ad-hoc-ness and free parameters: The SM has many free parameters, which need to be provided. The mass of fermions, coupling strengths, the Higgs mass etc. are not predicted by the SM. In addition, the origin of some regularities is unclear. Why are there only three generations of quarks and leptons. Why is the electric charge of up-type and down-type quarks equal exactly to $Q = {}^{2}/_{3}\mathrm{e}$ and $Q = -{}^{1}/_{3}\mathrm{e}$, respectively. Why is the up quark four orders of magnitude lighter than the top quark?
- Hierarchy and "fine-tuning" problem: Within the SM there is not any explanation for the Hierarchy problem. The quantum loop correction to the Higgs boson mass are proportional to the Planck scale and therefore the mass of the Higgs boson should be of similar order. Only "fine-tuning" would lead to a cancellation of such huge corrections and bring the Higgs mass to $m_H \approx 125\,\mathrm{GeV}$.

Thus the SM is believed to be incomplete.

1.4 Vector bosons at Hadron Colliders

1.4.1 Production, decay and properties

In pp collisions, the dominant process of the vector boson production is quark-antiquark annihilation, known as Drell-Yan process [45]. For $\mathrm{W}^{\pm}$ boson production, the quark and anti-quark need to be of different flavour and the main contribution comes from $\mathrm{u}\bar{\mathrm{d}}$ and $\bar{\mathrm{u}}\mathrm{d}$ annihilation, respectively. As the proton consists of two up quarks and one down quark, the antiquarks have to originate from the sea quarks, while the quarks can be also a valance quark. Therefore the higher probability of finding up quarks in the proton leads to asymmetric production of W^{+} over W^{-} bosons. The Z bosons are produced by the annihilation of the quark and anti-quark pair with the same flavour.

The decay mode and its width together with the branching ratio of the vector bosons can be calculated within the EW theory. The $\mathrm{W}^{\pm}$ and Z bosons decay modes are dominated by the fully hadronic channels with two quarks in the final state. More than two third of the produced vector boson decay in this channel. The rest of the $\mathrm{W}^{\pm}$ bosons decay into lepton and the corresponding neutrino. In case of the Z boson, the subdominant decay mode after the fully hadronic mode is the decay into two neutrinos with $\sim 20\%$ of all decays. The remaining 10% of decays are realized by a decay into two same flavour opposite charged leptons. The most common decay mode of the vector bosons together with the decay width and the branching ratio is summarized in Table 1.2.

1.4.2 Discovery of W and Z bosons

The important step for the discovery of $\mathrm{W}^{\pm}$ and Z bosons was the observation of weak neutral current [17] in 1973 that pointed towards models with a neutral gauge boson. In addition, it provided an estimation about the value of the Weinberg angle, which can be further utilized to predict the mass of $\mathrm{W}^{\pm}$ and Z bosons. Already in 1981, the accuracy of the prediction assuming the validity of the electroweak theory has ca 2%

Decay mode	Decay width/ GeV	Branching ratio
$W^+ \to ud$	0.695	33.349 %
$W^+ \to cs$	0.694	33.349 %
$W^+ \to l^+\nu_{l^+}$	0.695	33.322 %
$Z \to d\bar{d}$	0.379	15.233 %
$Z \to u\bar{u}$	0.296	11.897 %
$Z \to s\bar{s}$	0.379	15.233 %
$Z \to c\bar{c}$	0.296	11.879 %
$Z \to b\bar{b}$	0.373	14.981 %
$Z \to l^-l^+$	0.258	10.379 %
$Z \to \nu_{l^-}\nu_{l^+}$	0.507	20.398 %

Table 1.2 – Decay mode, width and branching ratio of the W and Z bosons.

relative uncertainty and the Born approximation of the $W^{\pm}$ and Z boson masses were estimated to be 79.5 ± 2.6 GeV and 90.0 ± 2.1 GeV, respectively [46]. According to the Improved Quark Parton model (IQPM), the production of such heavy vector bosons should require a pp collisions at a centre-of-mass energy larger than three times the boson masses though. The IQPM is extension of the QPM, which was formulated by Feynman [47], Bjorken and Paschos [48] in 1969. In this model, the proton is considered as composed objects of three valence quarks surrounded by the sea of gluons and virtual quark antiquark pairs. Within this model, each valance quark was assumed to carry ca one third of the proton momentum, therefore the pp collisions at $\sqrt{s} = 300$ GeV were needed. Such collisions energies were achieved in proton-antiproton collisions by SPS accelerator located at CERN in 1981, which enabled to discover vector bosons.

The events, where $W^{\pm}$ bosons were produced and decaying into lepton and the corresponding neutrino were expected to have following signature; the presence of the lepton with high transverse momentum and high missing transverse momentum from the undetected neutrino in the opposite azimuthal direction to the lepton direction. In addition, the spectrum of transverse momentum distribution of leptons should exhibit the Jacobian peak around the value corresponding to half of the $W^{\pm}$ boson mass. The search for the Z boson relied on the leptonic decay channel, where two leptons with high transverse momentum are expected. The discovery of $W^{\pm}$ and Z bosons was announced by the UA1 Collaboration at the SPS collider in 1983 [18, 19]. The left plot in Figure 1.10 shows the scatter plot of the transverse momentum of electrons and the missing transverse momentum. In total, 6 events were interpreted as $W^{\pm} \to e^{\pm}\nu_e$. The right plot in Figure 1.10 shows the invariant mass distribution of lepton pairs. After applying two selection criteria, a peak around 90 GeV was identify as the Z boson.

1.4.3 W and Z bosons and the LHC

Nowadays the measurements of W and Z bosons serve as a precision test of the SM. Theoretical prediction of the W and Z bosons production cross section are available at NNLO accuracy in perturbative QCD [50]. Thanks to the higher accuracy in pQCD calculations, the uncertainties arising from the choice of the renormalization (μ_R) and the factorization (μ_F) scale have been significantly reduced to the order of one percent [51, 52]. Moreover, electroweak corrections recently became available at NLO accuracy. They are important mainly for the production of vector bosons with high transverse momentum [53]. Therefore the measurement of $W^{\pm}$ and Z bosons production can constrain existing theoretical

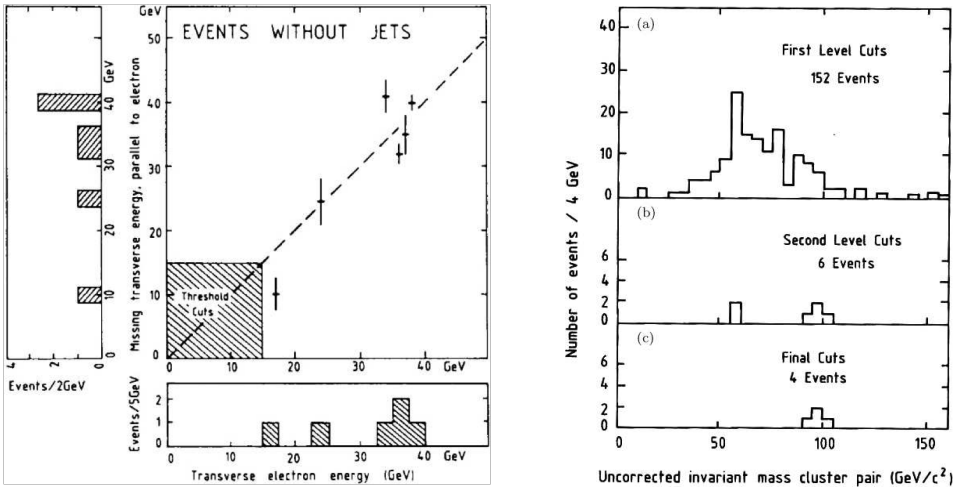

Figure 1.10 – The scatter plot of the transverse momentum of electrons and the missing transverse momentum with six events where W bosons were produced (left). The invariant mass distribution of lepton pairs (right). After applying selection criteria, a peak around 90 GeV was identify as the Z boson [49].

predictions and in addition provide information about the proton structure.

Z boson

The decay mode with the lepton pair in the final state is considered as a golden channel for measurements of Z bosons. The leptons serve as very clean experimental signatures for the reconstruction and event triggering. In addition, such final state is almost background free. There is only a very small contribution of different processes with the same signatures. Therefore such measurements can be very precise, nowadays dominated by the uncertainty related to the integrated luminosity [54]. Figure 1.11 shows one of the latest measurement [54] of the Z boson cross section from the Large Hadron Collider (LHC). The measurement covers a wide range of the transverse momentum of the Z boson from a very low value up to the TeV scale.

W boson

Compared to the Z boson, the measurement of W boson in leptonic decay channel is more difficult to observe due to the undetected neutrino resulting in the presence of missing transverse momentum. One of the most challenging measurement nowadays is the W boson mass measurement. From the experimental side, the measurement requires precise detector calibration for the determination of the missing transverse momentum and the lepton reconstruction. On the other hand, the theoretical challenges arise from the modelling of the transverse momentum spectrum of Z boson in the very low region, where the non-perturbative and resummation contributions play an important role. Figure 1.12 shows the results from the W boson measurement at the LHC [55]. The mass of W boson is determined to be $m_{\mathrm{W}} = (80370 \pm 19)$ MeV [55].

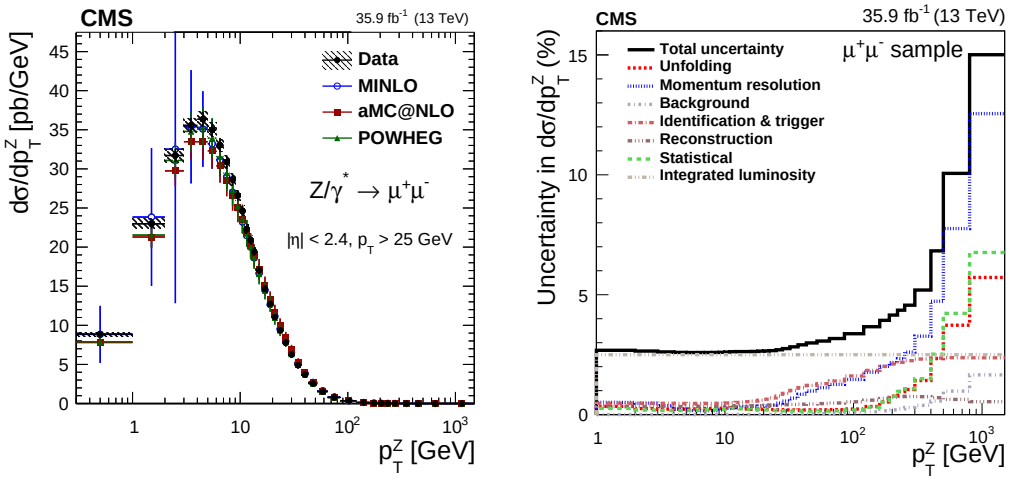

Figure 1.11 – The measured cross section [54] of the Z boson in bins of transverse momentum p_{T} and compared to the various predictions(left) and the systematic uncertainties from various sources(right).

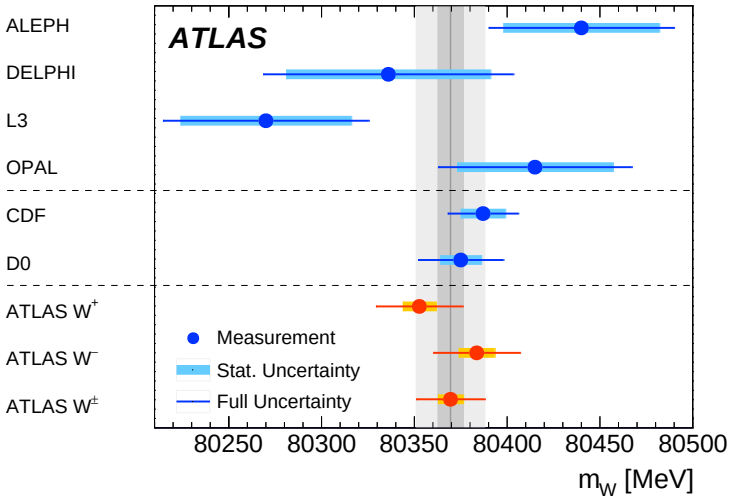

Figure 1.12 – The W boson mass as measured by the ATLAS Collaboration [55] and other collider experiments.

2

Jets in hadronic collisions

When studying high-energy pp collisions, there are many processes involving quarks and gluons in the final state. In addition, the quarks can originate from the hadronic decay of heavy resonances. However as discussed in the previous section, these partons are not observed as free particles. The evolution of these elementary particles carrying colour charge is governed by QCD and leads to collimated sprays of stable particle. Then these particles are recombined into hadronic jets with the aim to give access to kinematic quantities of the original partons. The jets are viewed as proxy to the high-energy quarks and gluons.

2.1 Jet clustering algorithms

In the early days, the information about all produced particles from the whole event was utilized to measure event shape variables. For an example, a thrust gives information about the events and characterizes the compatibility with the event containing two oppositely directed jets. The main advantage of the event shape variables is the fact that they are very easily calculable in pQCD. Therefore they are still popular nowadays [56]. But a proper definition of jet algorithms is needed for study the dynamics of quarks and gluons in more detail. When defining a jet clustering algorithm, several important properties are required. Among those are:

- An algorithm has to be infrared and collinear (IRC) safe. These properties correspond to the requirements that the presence of an additional soft emission and collinear splitting does not change the set of hard jets. They are extremely important requirements as there are always soft emission in the parton shower and collinear splitting related to the fragmentation processes.
- An algorithm needs to be applicable for the experimental analysis and theoretical calculations. In case of the theoretical calculations, the algorithm needs to be valid at any order of the perturbative calculations.

A wide variety of the jet algorithms has been developed over time, but nowadays the most popular jet algorithms, namely the anti-k_{T} and the Cambridge/Aachen algorithm, belong to the group of sequential recombination algorithms. They systematically go through all particles in the events and create pairs of particles. If certain rules are satisfied, the pair of two particles is recombined into a new object. For the sequential recombination algorithms, the rules are following:

1. For all possible pairs of particles i and j, the longitudinally invariant distances defined

$$d_{ij} = \min\left(p_{\mathrm{T},i}^{2p}, p_{\mathrm{T},j}^{2p}\right)\frac{\Delta R_{ij}^2}{R^2} \quad \text{and} \quad d_{iB} = p_{\mathrm{T},i}^{2p} \tag{2.1}$$

   is calculated. Here $p_{\mathrm{T},i}$ and $p_{\mathrm{T},j}$ are the transverse momentum of particles i and j, respectively. The distance between these particles in the $\eta - \phi$ plane is given by $\Delta R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$, where η and ϕ are the pseudorapidity and azimuthal angle of given particle. The d_{iB} is a beam distance, while d_{ij} is the so-called k_{T} distance. The parameter R is the cone size radius, sometimes called as a resolution and is constant. It characterizes the cone width of the clustered jets. The parameter p is characteristic parameter for each sequential recombination algorithms and controls the power of energy and the geometrical scales.

2. Find the smallest value of d_{ij} and d_{iB}.

3. If this is d_{ij}, then the particles i and j are combined into a new particles k. The kinematic properties of the new particles are the p_{T}-weighted average of the quantities of the original particles i and j, namely

$$p_{\mathrm{T},k} = p_{\mathrm{T},i} + p_{\mathrm{T},j}, \tag{2.2}$$

$$\eta_k = \frac{\eta_i p_{\mathrm{T},i} + \eta_j p_{\mathrm{T},j}}{p_{\mathrm{T},k}}, \tag{2.3}$$

$$\phi_k = \frac{\phi_i p_{\mathrm{T},i} + \phi_j p_{\mathrm{T},j}}{p_{\mathrm{T},k}}. \tag{2.4}$$

   The particles i and j are removed from the list, while the particle k is inserted into the list. Then the algorithm returns to the step 1.

4. If the smallest value is d_{iB}, then the particle i is defined as a jet and is removed from the list. Then the algorithm returns to the step 1.

5. The algorithm proceeds until no particle is left in the list.

This set of rules ensures that the sequential recombination algorithms are IRC safe. The most frequent choice of the value of parameter p is one of the following integer: -1, 0, 1. The choice of the parameter p has a certain consequences for the clustering history as it will be discussed later.

The k_{T} algorithm

The k_{T} algorithm [57] is obtained for $p = 1$ and it is the oldest sequential recombination algorithm. The choice of $p = 1$ leads to the fact that the clustering follows the QCD branching in the reverse direction. It means that the softest particles are clustered firstly, while the hardest particles are recombined in the end. Hence the jets have the irregular shapes. In addition, as a consequence of the clustering history, a few soft particles can be identify as a jets. Therefore the k_{T} algorithm is sensitive to the UE activity and the pileup. Nowadays this algorithm is not preferred for the usage at the current hadron colliders.

The anti-k_T algorithm

This algorithm [58] is obtained for $p = -1$. In contrast to the k_T algorithm, the hardest partons are clustered first. Thus, the jets are growing around the hardest partons and the yielding jets have a regular and well defined shape. Moreover, the choice of $p = -1$ gives the higher robustness against the UE activity and pileup. Because of these properties, the anti-k_T algorithm is nowadays one of the most popular and widely used jet clustering algorithms.

The Cambridge/Aachen algorithm

The Cambridge/Aachen (CA) algorithm [59] is obtained for $p = 0$. In the previous two cases, the energy scale of the partons played an important role for the clustering, but by choosing of $p = 0$, the jet clustering is based only on spatial separation. Hence, the CA algorithm follows the QCD branching in angles. As it will be discussed later, this is especially important when studying the inner structure of the jets.

Figure 2.1 shows the comparison between the different clustering algorithms for the same event. The jet area is shown in $\eta - \phi$ plane. Compared to the CA and k_T algorithm, the jets clustered by the anti-k_T algorithm have a regular shape as one can see in Figure 2.1.

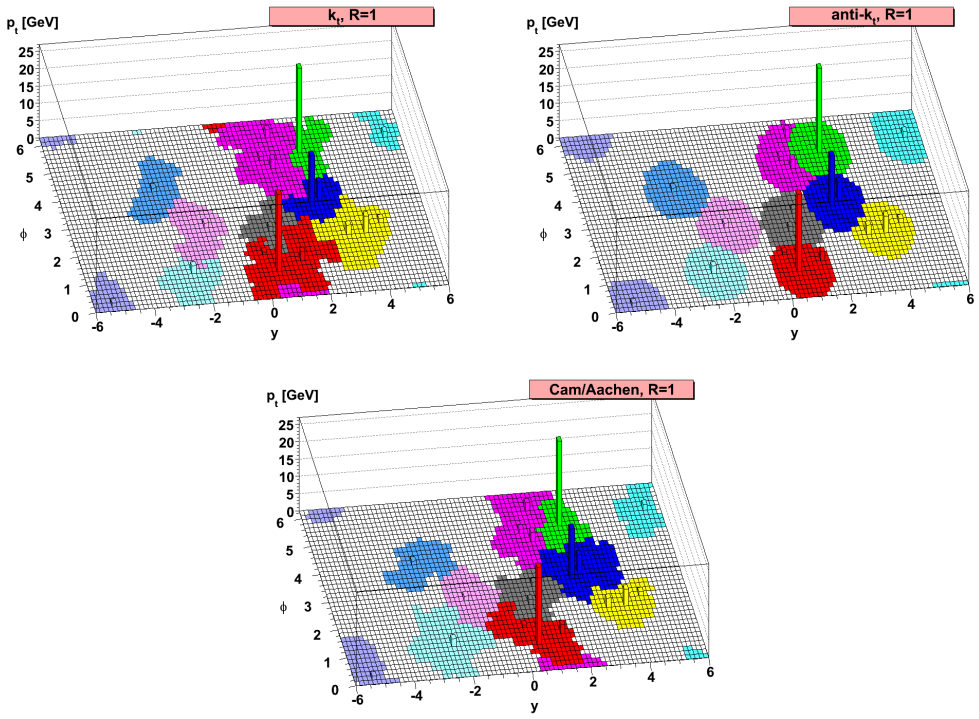

Figure 2.1 – Illustration of the three jet clustering algorithms (k_T, anti-k_T, and Cambridge/Aachen) applied to the same event [58].

2.2 Jet substructure

For a several decades, jets are playing crucial role in particle physics. Thanks to relatively high cross section, the inclusive jet cross section measurement is considered as one of the textbook measurement that probes the QCD at multiple scales. In pp collisions at the

centre-of-mass energy of 13 TeV, the inclusive measurement [60] covers the range from $p_{\mathrm{T}} \approx 100$ GeV to $p_{\mathrm{T}} \approx 2$ TeV and the measured cross-section varies over of 10 orders of magnitude. Hence, such measurement is suitable for testing and verifying various aspect of QCD.

In contrast to that, jet substructure is a rather young field in particle physics. At the collisions energies reachable at the LHC, the electroweak bosons, namely W, Z and H bosons, together with top quarks are produced beyond the energy threshold. Thus, the transverse momentum can easily exceed their masses and they gain high Lorentz-boost. The dominant decay channel of these resonances is to quarks that are therefore highly boosted and collimated in the direction of the mother particle. The small opening angle between them, which can be estimated [1] as

$$\Delta R_{12} \approx \frac{2m}{p_{\mathrm{T}}} \tag{2.5}$$

results in the topology, where the heavy resonance with mass m and the transverse momentum p_{T} is reconstructed as one large and massive jet, a so-called fat jet. In the past, such topologies were unpopular because of the low separation power between the event of interest and the overwhelming contribution mainly coming from QCD multijet production. However recent progress in the understanding of the inner structure of jets, referred to as jet substructure, provided new options for the identification and study of boosted topologies. The scientific studies of the boosted objects and jet substructure cover the wide range of the spectrum of the topics from the precision SM measurement including the study of jet substructure for various jet flavour [61], the determination of α_S [61] and top quark mass [62], to searches [63] of physics beyond SM (BSM). Since in many extensions of SM, there are heavy resonances at the TeV scale that predominantly decay into SM bosons.

From a wide variety of the jet substructure variables, we will briefly discuss the jet mass and two substructure variables that exploit the inner structure of jets for the W and Z bosons tagging. More comprehensive discussion can be found in Ref. [64, 56].

2.2.1 Jet mass and grooming algorithm

A good starting point to identify the boosted W and Z bosons from the QCD multijet production is through the jet mass. The jet invariant mass is defined as

1. It is derived from the invariant mass m of two body decay. Here the decay product are two quarks with energies E_1 and E_2 and four momenta p_1 and p_2, respectively. The invariant mass is then given as

$$m^2 = (p_1 + p_2)^2 \approx 2E_1E_2\,(1 - \cos\theta_{12}) \approx E^2 z\,(1 - z)\,\theta_{12},$$

where the last approximation is valid for the very small opening between the decay products, i.e. $\theta_{12} \ll 1$. Here z and $(1 - z)$ are energy fractions that are carried by the decay products. The opening angle roughly corresponds to the ΔR_{12} in $\eta - \phi$ plane. In case of the symmetric decay($z = 1/2$) in the boosted regime ($p_{\mathrm{T}} \ll m$), we can write

$$\Delta R_{12} = \frac{m}{E\sqrt{z\,(1 - z)}} \approx \frac{2m}{p_{\mathrm{T}}}.$$

$$m_{\text{jet}}^2 = \left(\sum_{i \in \text{jet}} p_i \right)^2 , \tag{2.6}$$

where the sum runs over all constituents of the jet. The jet invariant mass is quite often called the ungroomed jet mass or plain jet mass. In practice, the ungroomed jet mass is not used for the vector boson tagging as is very sensitive to the soft and wide-angle radiations. The mass of jets initiated by quarks and gluons is generated by the parton shower, that is driven by the Sudakov form factor. Hence, it is very unlikely to have jets with $m_{\text{jet}} \sim 0$ GeV. The ungroomed jet mass distribution exhibits a large peak located around $m_{\text{jet}}/p_{\text{T}} \approx 0.1$. This peak is referred as a Sudakov peak and above the peak, the jet mass distribution decreases.

Several grooming techniques were developed and introduced as a tool to reduce the Sudakov peak by removing the soft and/or wide-angle radiations from the jets. The main objective of the grooming technique is to help to resolve the "partonic" mass of the jets. The grooming algorithms re-cluster the jets while the soft and/or wide-angle contributions are rejected.

Trimming

The trimming algorithm [65] re-clusters the original jet with the cone size R into subjets with the characteristic radius $R_{\text{sub}} < R$. It uses the k_{T} algorithm for the subjets clustering, therefore it starts with the soft particles. The subjets with a momentum fraction smaller than a certain value f_{cut},

$$p_{\text{T, subjet}}/p_{\text{T, jet}} < f_{\text{cut}}, \tag{2.7}$$

are then removed. The remaining subjets form the trimmed jet. The tunable parameters of the trimming technique are R_{sub} and f_{cut}. As the k_{T} algorithm is used, the geometrical scale of the subjets is not explored.

Pruning

In contrast to trimming, the pruning technique [66] with two parameters z_{cut} and r_{cut} aims in addition to reduce the contribution from the wide-angle radiation. The original jet is re-clustered by the CA algorithm and at each step of re-clustering it requires

$$\frac{\min\left(p_{\text{T,a}}, p_{\text{T,b}}\right)}{p_{\text{T,a}} + p_{\text{T,b}}} > z_{\text{cut}} \quad \text{and} \quad \Delta R_{a,b} < D_{\text{cut}} = \frac{2 r_{\text{out}} m_{\text{jet}}}{p_{\text{T}}}. \tag{2.8}$$

Here $p_{\text{T,a}}$ and $p_{\text{T,b}}$ are the transverse momenta of the objects (single particle or group of particles) a and b, respectively. The first requirement checks the hardness of the combination, while the relative angle between the recombined objects is examined by the second condition. If these requirements are met, the objects a and b are recombined, otherwise the object with lower transverse momentum is rejected.

Soft Drop and Modified Mass Drop Tagger

The soft drop [67] technique uses the CA algorithm to create a clustering pairwise tree with an angular-ordered structure by going backwards in the sequence in which the jets

were build. The algorithm depends on a soft threshold z_{cut} and the angular exponent β. At each step of the clustering, when two objects are merged into one object, the following soft drop condition

$$\frac{\min\left(p_{\text{T},1}, p_{\text{T},2}\right)}{p_{\text{T},1} + p_{\text{T},2}} > z_{\text{cut}} \left(\frac{\Delta R_{12}}{R_0}\right)^{\beta}, \tag{2.9}$$

needs to be fulfiled. Here the parameter z_{cut} defines the scale what is considered as a soft radiation, while the parameter β controls the grooming procedure as a function of the angular separation of the two subjets. One can immediately see, that for $\beta \to \infty$, the grooming technique returns the whole jet. By setting $\beta > 0$, the soft radiation is removed from the jet, while a certain fraction of the soft-collinear radiation is kept. Removing of collinear as well as soft radiation can be achieved with $\beta < 0$. The soft drop with $\beta = 0$ and $z_{\text{cut}} = 0.1$ is identical with the modified mass drop tagger (mMDT) [68] and returns the jet with exactly two subjets. All soft radiation wider than this two-prong structure is removed. This is especially beneficial for the tagging of W, Z and H jets as the two-body decay of these bosons results in the jet with the two subjets. In addition, the mass of each subjets is expected to be smaller than the mass of the jet. Unlike to the QCD jets, where the mass is generated by the continuous soft radiation, the mass of hardest subjet is closed to the jet mass.

Nowadays the soft drop algorithm with $\beta = 0$ and $z_{\text{cut}} = 0.1$ is one of the most popular and frequently used grooming technique. The jet mass and the grooming technique is not useful only for the heavy resonance tagging, but also for the SM measurement. Figure 2.2 shows the jet mass distribution measured in dijet events. The study was performed for the ungroomed and groomed jet mass in order to gain insight into the soft radiation, which is more challenging for the MC event generators. One can nicely see the Sudakov peak in the ungroomed jet mass, while the grooming technique pushes the jet mass distribution closer to the zero. Once the soft radiation is removed by the soft drop algorithm, a better agreement with the MC prediction of PYTHIA8 is achieved.

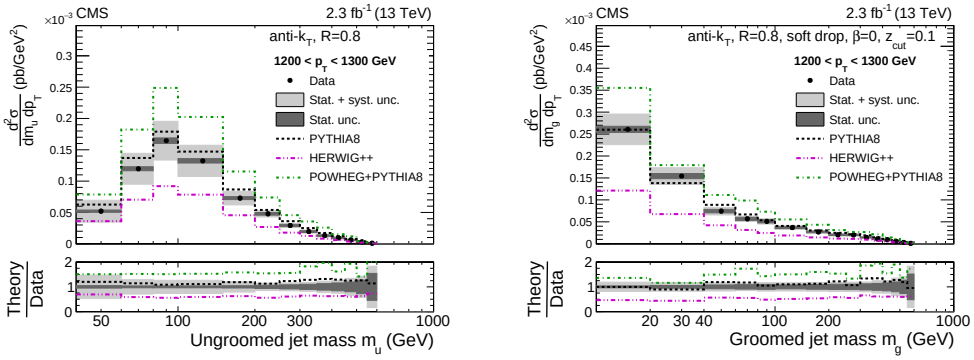

Figure 2.2 – Normalized cross section for the ungroomed (left) and groomed (right) jet mass for the same p_{T} bin [69].

2.2.2 Two-prong substructure

Apart from the jet mass, additional observables are needed to distinguish the W and Z bosons from the QCD jets. The main idea behind these observations is the fact that the W and Z bosons decay into two quarks and form the jets with two subjets. On the

other hand, the high energetic partons continuously radiate and therefore the formed QCD jet consists typically of several large angle and asymmetric splitting. Examples of such observables, that aims to count the number of hard prong structure inside the jets, are the N-subjettiness [70] and the energy correlation functions (ECFs) [71].

N-subjettiness

This algorithm measures the n-subjettiness variable [70], τ_N, defined as

$$\tau_N = \frac{1}{d_0} \sum_{1 \leq i \leq n_j} p_{\mathrm{T},\,i} \min\left(\Delta R_{1,i}, \Delta R_{2,i} \ldots \Delta R_{N,i}\right), \tag{2.10}$$

where the sum runs over all jet constituents with transverse momentum $p_{\mathrm{T},\,i}$. Here $\Delta R_{1,i}$ stands for the angular separation between the jet constituents and candidate subjet axes obtained by the optimization procedure which minimizes N-subjettiness. The normalization factor d_0 is given by

$$d_0 = \sum_{1 \leq i \leq n_j} p_{\mathrm{T},k} R_0, \tag{2.11}$$

where R_0 is the jet cone size. Then the ratio of the n-subjettiness is utilized to measure the compatibility of a jet with N-axis hypothesis. The jets with $\tau_{\mathrm{N}} \gg 0$ have most likely more than N subjets, while $\tau_{\mathrm{N}} = 0$ indicates that most of the radiation is aligned along the N subjet axes. It means that jets originating from W and Z bosons decays are expected to have large τ_1 and small τ_2. For the tagging of a jet with two-prong structure, the ratio $\tau_2/\tau_1 = \tau_{21}$ is used instead of just the variables τ_1 and τ_2. The reason for that is if the QCD jet has large τ_1, it typically tends to have similarly large τ_2.

Energy correlation functions

The generalized ECFs [71] based on ECFs measure the correlation of the particles inside the jet. In the most general case, the n-point ECF with the angular exponent β for v pairwise angles is defined as

$${}_v e_n^{\beta} = \sum_{1 \leq i_1 < i_2 \cdots < i_n \leq n_j} \left[\prod_{1 \leq k \leq n} \frac{p_{\mathrm{T},\,i_k}}{p_{\mathrm{T}}} \right] \min \left\{ \prod_{m=1}^{v} \Delta R_{i_j,i_k}^{\beta} | i_j < i_k \in i_1, i_2 \ldots i_N \right\}. \tag{2.12}$$

Here the jet with the transverse momentum p_{T} has n_j constituents with transverse momenta $p_{\mathrm{T},\,i_k}$. As one can see, the definition of ECF is subjet axis free which is considered as an advantage compared to the N-subjettiness. The ECFs measure the number of centers of hard radiations inside the jet. For the tagging of jets with two-prong substructure, the ratio of 2- and 3-point ECFs , which are defined as

$${}_1 e_2^{\beta} = \sum_{1 \leq i < j \leq n_j} \frac{p_{\mathrm{T},\,i}}{p_{\mathrm{T}}} \frac{p_{\mathrm{T},\,j}}{p_{\mathrm{T}}} \Delta R_{ij}^{\beta}, \tag{2.13}$$

$${}_2 e_3^{\beta} = \sum_{1 \leq i < j < k \leq n_j} \frac{p_{\mathrm{T},\,i}}{p_{\mathrm{T}}} \frac{p_{\mathrm{T},\,j}}{p_{\mathrm{T}}} \frac{p_{\mathrm{T},\,k}}{p_{\mathrm{T}}} \min\left(\Delta R_{ij}^{\beta} \Delta R_{ik}^{\beta}, \Delta R_{ij}^{\beta} \Delta R_{jk}^{\beta}, \Delta R_{jk}^{\beta} \Delta R_{ik}^{\beta} \right), \tag{2.14}$$

is recommended. For the QCD jets, it is typically expected ${}_2e_3 \sim {}_1e_2$ as 1-point correlations are dominating, while the two-prong structure results in large 2-point correlations, i.e. ${}_2e_3 \ll {}_1e_2$. In particular, the following dimensionless ratio defined as

$$N_2^{\beta=1} = \frac{{}_2e_3^{\beta=1}}{\left({}_1e_2^{\beta=1}\right)^2} \tag{2.15}$$

shows the best separation power to distinguish the jets coming the boosted W and Z bosons decay from the QCD jets.

2.3 Vector boson tagging

The identification of W and Z bosons from the QCD jet usually relies on applying selection criteria related to the groomed jet mass and two-prong substructure variables. The exact value of cuts are analysis dependent and need to be optimized in order to achieve the best possible signal-to-background efficiency.

2.3.1 Mass decorrelation

The jet substructure variables are correlated with the jet mass and applying cuts on the jet substructure results in an undesired mass sculpting of the jet mass distribution. It means that after applying cuts, the groomed jet mass distribution for the QCD jet is not steeply falling any more. In addition, the shape and the position of sculpted distribution depends on the jet p_{T}. For most of the analyses, it is an unwanted behaviour as the signal extraction typically relies on a data-driven technique that utilizes the side bands to predict the shape of the background in the signal region. For searches in BSM physics, this can lead to a situation, when a possible signal will sit on the top of peaked background. There are several techniques to solve this issue by performing a decorrelation of the tagger from the jet mass. One of the important aspects of the decorrelation method is the attempt to preserve the separation power of variables which are being decorrelated.

Designed decorrelated tagger

In Ref. [72], the construction of a designed decorrelated tagger (DDT) based on τ_{21} variable was proposed. The main idea of the decorrelation method is to flatten the τ_{21} dependence on ρ' for various jet p_{T} range. Here ρ' is a dimensionless variable, defined as

$$\rho' = \log\left(\frac{m^2}{p_{\mathrm{T}} \cdot 1\ \mathrm{GeV}}\right), \tag{2.16}$$

that ties the jet mass m and the transverse momentum p_{T}. Then the linear transformation of τ_{21} according the following formula

$$\tau_{21}^{\mathrm{DDT}} = \tau_{21} + M \cdot \rho' \tag{2.17}$$

can be done. The parameter M needs to be determined from the MC simulations for QCD jets. The only assumption and therefore limitation of this technique is on the linearity of τ_{21} versus ρ'. Figure 2.3 shows the soft drop jet mass distributions after applying the selection criteria on τ_{21} and τ_{21}^{DDT}, respectively. As one can clearly see, the DDT removed the observed mass sculpting.

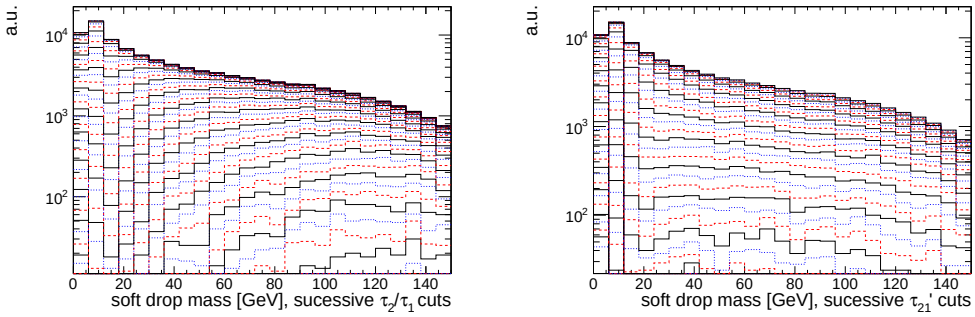

Figure 2.3 – Soft drop jet mass distribution for different jet p_{T} bins after applying various cuts on τ_{21}(left) and τ_{21}^{DDT}(right) [72].

Fixed-efficiency regression

The concept of the DDT technique can be further extended and generalized by fixed-efficiency regression. For given jet substructure variable in bins of a specific observable, let us measure its profile. Then we can find a quantile X and corresponding value of the jet substructure such that it divides events into groups with X and $(1-X)\,\%$ efficiency. For example, one wants to apply the selection on τ_{21} variable and keeps only $X\%$ background (QCD) contribution in the given bins of the jet mass m distribution. Thus, one needs to find $\tau_{21}\left(m;\,\varepsilon_{\mathrm{QCD}}=X\%\right)$ in bins of the jet mass and then perform the transformation as

$$\tau_{21}^{\mathrm{DDT}}\left(m\right)=\tau_{21}\left(m\right)-\tau_{21}\left(m;\,\varepsilon_{\mathrm{QCD}}=X\%\right). \tag{2.18}$$

The selection $\tau_{21}^{\mathrm{DDT}}\left(m\right)<0$, which is equivalent to $\tau_{21}\left(m\right)<\tau_{21}\left(m;\,\varepsilon_{\mathrm{QCD}}=X\%\right)$, yields the fixed background efficiency in the given bin. Additional improvement of this decorrelation technique can be achieved by doing the decorrelation double differentially not only as a function of the jet mass, but also as the jet p_{T} or other suitable variables.

In summary, let us stress that both of the decorrelation methods, DDT and fixed-efficiency regression, rely fully on the MC predictions. Therefore the mismodelling of the jet variables used for the decorrelation technique results in a residual disagreement and effects which are not visible in the MC simulations. In addition, the limited statistics of the simulated MC samples can lead to a non-smooth behaviour during the decorrelation.

2.3.2 Machine learning based taggers

For completeness, let us briefly discuss heavy resonance taggers based on the machine learning. In recent years, machine learning (ML) and deep learning (DL) techniques are exploited to identify and classify jets originating from the hadronic decay of boosted W, Z, H bosons and top quarks. As input, the taggers use the so-called low level variables like four-momenta of the jet constituents as well as information about secondary vertices and so on. In most cases, the taggers are trained by supervised learning. It means that the simulated MC samples contain well defined types of jets and therefore the tagger can learn the properties of the jets originating from boson decay and QCD jets. The discrimination is then constructed. Even ML taggers suffer from the mass sculpting. From a wide spectrum of ML, we can highlight DeepAK8 [73], ParticleNet [74] and LundNet [75] tagger. The DeepAK8 is multi-class classifiers tagger, which is recently commissioned by the CMS Collaboration. It aims to distinguish the jets originating from the hadronic decay

of W, Z, H bosons and top quark versus QCD jets. In addition, it is trained for heavy flavour tagging, i.e. the decay channels like $Z \to b\bar{b}$, $Z \to c\bar{c}$, $H \to b\bar{b}$, and $H \to c\bar{c}$ are also identified. The next generation of this tagger is the so-called ParticleNet tagger, where the jets are represented as a clouds of particles. Finally, the LundNet tagger utilizes the jet Lund plane [76] for the tagging as the information about the inner structure is encoded in it.

3

The Large Hadron Collider and the Compact Muon Solenoid

The Large Hadron Collider (LHC) [77], located at *Conseil Europen pour la Recherche Nuclaire* (CERN) in Geneva, is the world's largest particle accelerator ever built. It is situated in 26.7 km circular tunnel previously used until 2001 for the Large Electron Positron (LEP) collider [78] at the depth of 45 m to 170 m under surface close to the border between France and Switzerland. The LHC is a superconducting two-ring hadron accelerator and collider designed for pp as well as heavy ion collisions. Its construction started in 1998 and after the commission in 2008, the first pp collision happened in 2009.

3.1 LHC Injection Chain

A schematic illustration of the CERN accelerator complex with the complete injection chain is shown in Figure 3.1. As one can see, the LHC [77] is the last chain of the accelerator complex. The proton bunch is created in the duoplasmatron source, where the hydrogen atoms are ionized by the magnetic field and create a bunch with the energy of 90 keV. Then, the proton bunch is injected into the Linac2 and pre-accelerated to 50 MeV. Consecutively, the bunch is injected into the Proton Synchrotron Boosted (PSB) and the Proton Synchrotron (PS). The bunch leaves the PS with energy of 25 GeV. Before the injection into the LHC, the bunch is sent to the Super Proton Synchrotron (SPS) where it reaches the energy of 450 GeV. Finally, the bunches are injected into the LHC separately into two counter-circulating beams (a clockwise and anticlockwise direction) and accelerated to the final energy (7, 8 or 13 TeV). The proton beam is formed from up to 2808 bunches and each of them consists of 1.15×10^{11} protons. The bunches in the LHC are separated by 25 ns. In case of the heavy ion physics, the beam consists of 600 lead bunches and each of them contains 7×10^{7} lead ions. The heavy ion bunches are pre-accelerated in a similar way as the proton bunches, just with one small difference - the pre-acceleration of heavy-ion bunches starts in Linac3.

3.2 LHC Layout

The LHC itself consists of eight arcs per beam separated by eight interaction points (IP). The beams cross and collide in four of them. These IPs are occupied by the experiments, which will be discussed later. The remaining IPs house accelerator equipment. The beam cleaning, namely the momentum and betatron cleaning is done at IP3 and IP7,

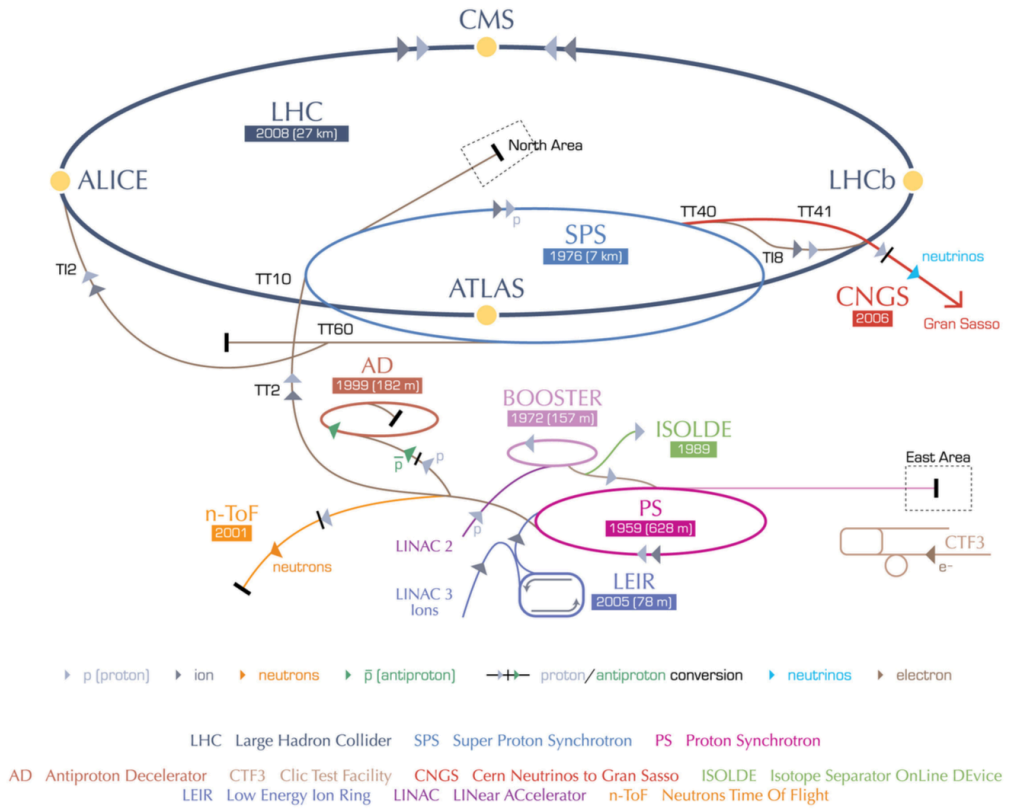

Figure 3.1 – The CERN accelerator complex with the complete injection chain including all pre-accelerators [79].

respectively. The beam acceleration system is located at IP4. Finally, the beam dump is done by the kicker magnet at IP6. The layout of the LHC is shown in Figure 3.2.

Each arc contains 23 cells, with the classical `FODO` structure [80]. In total, the LHC consists of 184 cells, each of them being 106.9 m length. The `FODO` cell is a magnetic structure with the regular arrangement in the following order: the focusing quadrupole, three dipole magnets, the defocusing quadrupole, three dipole magnets again and the focusing quadrupole. The shortcut `FODO` is related to the fact that the dipole magnets do not contribute in first order to the focusing (zero focusing = `0`). The schematic layout of the `FODO` cells is shown in Figure 3.3. Each of the dipole magnets, made of the NbTi magnets, weights 35 tonnes and is 15 m long. They are cooled down to 1.9 K by the liquid helium and produce the magnetic field with the strength of 8.3 T. Such a strong magnetic field keeps the beam motion in a circular trajectory. In addition, between the focusing and defocusing quadrupole, the other type of magnets like sextupole and decapole magnets are installed for a fine-tuning of the beam. Before entering the interaction points, the bunches are squeezed closer together to increase the probability of the interaction. At IP1 and IP5, the diameter of the beam is in the order of $\sim 10\,\mu$m.

3.2.1 LHC detectors

Currently, there are four major experiments at the LHC located at the interaction points, namely:

- A Large Ion Collider Experiment (ALICE) [81],

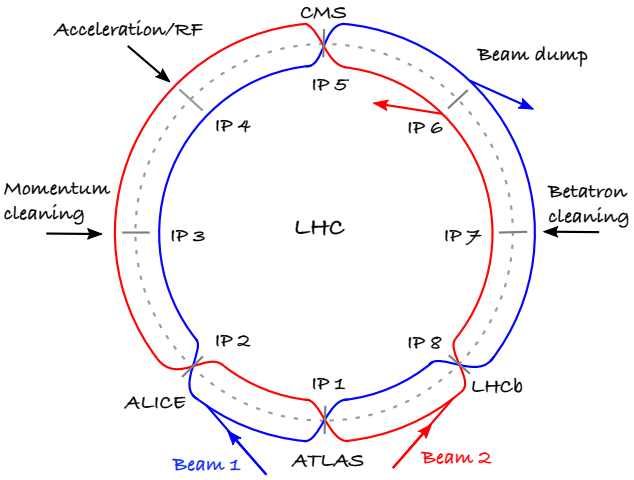

Figure 3.2 – The layout of the LHC.

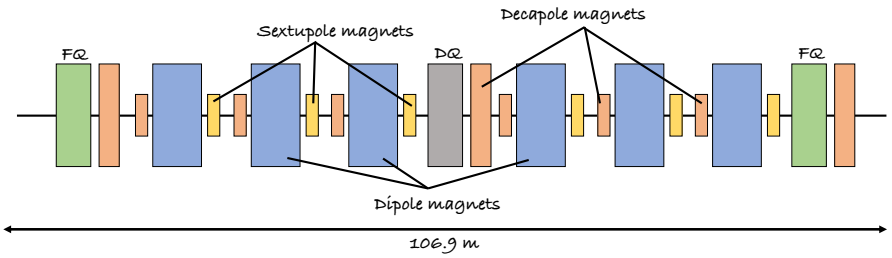

Figure 3.3 – The schematic layout of the FODO cell.

- A Toroidal LHC Apparatus (ATLAS) [82],
- Compact Muon Solenoid (CMS) [83],
- Large Hadron Collider beauty (LHCb) [84],

The ATLAS and CMS experiments are general multi-purpose detectors built to cover the widest possible range of physics at the LHC. Although the different technical concepts and design were used for their construction, they reach comparable performance and they can independently confirm and cross check the results from the same physics measurements. In addition, the combination of results from both experiments can reduce the statistical uncertainties. The main physics goal, the discovery of the Higgs boson, was achieved in 2012, when both experiments reported on the observation of the Higgs boson [5, 6]. The physics program of the ATLAS and CMS experiments covers a wide range of topics: from precision measurement of SM physics, Higgs physics to the searches for the signature of the BSM physics, dark matter and supersymmetry. In addition, both experiments are active in heavy ion physics and achieve competitive results to the ALICE experiment. In this thesis, we shall use the data recorded by the CMS experiment, therefore a detailed description of this detector will be discussed later.

ALICE is the only dedicated heavy ion experiment at the LHC. The heart of the ALICE detector is the Time Projection Chamber (TPC) surrounded by the solenoid magnet with the magnetic field up to 0.5 T. The purpose of the TPC is to reconstruct tracks and

to identify particles via specific energy loss $\mathrm{d}E/\mathrm{d}x$. The TPC together with the Time Of Flight (TOF) detector and the Inner Tracking System (ITS) provides particle identification and separates charged pions, kaons and proton down to low p_{T} at mid-rapidity ($|\eta| < 0.9$). The primary goal of the ALICE experiment is to study QCD matter at extreme conditions of high temperatures and energy densities, where quarks and gluons are no longer confined inside hadrons, the so-called quark-gluon plasma. It is expected that this state of matter was present in the earliest stages of the universe shortly (10^{-5} s) after the Big Bang and that it can be experimentally recreated in high energy heavy-ion collisions [85].

The last experiment, not mentioned yet, is the LHCb experiment, a single arm forward spectrometer. It is a general-purpose detector in forward direction with the main focus on the precision measurement with the heavy-flavour hadrons. The main physics goal of the LHCb is to perform a measurement of charge-parity violation (CPV), to test lepton-flavour universality/violation and the spectroscopy program focused on searched QCD-allowed states including c or b quark like double charmed baryon Ξ_{cc}^{++} [86] etc.

Additionally to the major experiments, the LHC hosts three smaller experiments, namely:

- Total Elastic and Diffractive Cross Section Measurement (TOTEM) experiment [87],
- Large Hadron Collider forward (LHCf) experiment [88],
- Monopole and Exotics Detector at the LHC (MoEDAL) experiment [89].

These smaller experiments share the IP with one of the major experiments. The TOTEM and LHCf is situated in the forward direction of the CMS and ATLAS experiment, respectively and study the production of particles in the forward direction close to the beam line. While TOTEM focuses on the charged particles, the neutral particles are of particular interest of the LHCf. The MoEDAL experiment, located closed to LHCb, searches for the unknown particle carrying magnetic charge, the so-called magnetic monopole.

3.3 Accelerator Parameters

3.3.1 Beam parameters

The most important parameters characterizing the accelerator performance are the emittance, the amplitude function, and the crossing angle. These parameters are used also to calculate the instantaneous luminosity and therefore they are briefly discussed.

Emittance ϵ is the area occupied by the particle beam in $p-x$ phase space and characterizes the spread of the particle beam. The particle beam can be described by the physical dimensions (the Gaussian width σ), but they can vary with the location in the accelerator. According to Liouville's theorem [80] the emittance is conserved as the beam circulates in a synchrotron. The amplitude function $\beta(z)$ is a beam optics quantity and is determined by the accelerator magnet configuration. For the definition of luminosity, we need to know the amplitude function at the interaction point. The relation between the amplitude function $\beta(z)$ and the amplitude function at the interaction point β^* is given by

$$\beta(z) = \beta^* + \frac{z^2}{\beta^*}, \tag{3.1}$$

where z stands for the distance along the nominal beam direction. Finally, the crossing angle θ_{c} is defined as the angle between two interacting particle beams.

3.3.2 Luminosity and collision energy

The two key characteristics of an accelerator are the centre-of-mass energy and the instantaneous luminosity. The centre-of-mass energy of the collision of two identical beams is given by

$$\sqrt{s} = 2E_{\text{beam}}, \tag{3.2}$$

where the E_{beam} is the energy of the beam. In case of heavy-ion and proton-ion collisions, the centre-of-mass energy $\sqrt{s_{\text{NN}}}$ is normalized by number of nucleons inside of the ion. The design value of the centre-of-mass energy for pp collisions is 14 TeV, while the heavy ion collision using the lead ion $^{207}_{82}\text{Pb}$ can reach up to $\sqrt{s_{\text{NN}}} = 5.02$ TeV.

A second key parameter is the instantaneous luminosity $\mathcal{L}$, which summaries all quantities related to the accelerator and provides the information about the number of interaction produced per second and per surface area. The instantaneous luminosity $\mathcal{L}$ can be defined by the collider parameters as

$$\mathcal{L} = \frac{n_b N_b^2 f \gamma F}{4\pi\sqrt{\beta_x^* \epsilon_x \beta_y^* \epsilon_y}}, \tag{3.3}$$

where N_b is the number of particles per bunch, n_b stands for the number of bunched per beam, f is the collision frequency, $\gamma = E/m$ is the relativistic gamma factor, $\epsilon_{x,y}$ is the emittance of the beam, $\beta_{x,y}^*$ is the amplitude function at the interaction point and F is the geometric reduction factor related to the crossing angle θ_c. At the LHC, the crossing angle is about 300 μrad. The LHC was designed for $\mathcal{L} = 10^{34}\text{cm}^{-2}\text{s}^{-1}$, but already in 2017 the machine achieved $\mathcal{L} = 2.05 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$. The integrated luminosity over time, $\mathcal{L} = \int \mathcal{L}(t)\,\text{d}t$, is used to characterize the size of the data sample and gives us the expected number of events for a given process with known cross section σ through

$$N_{\text{events}} = \mathcal{L} \cdot \sigma. \tag{3.4}$$

Figure 3.4 shows the overview of the integrated luminosity in the LHC Run 2. In that data-taking period, the LHC delivered 162.85 fb^{-1} and the CMS experiment recorded data sample corresponding to the 150.26 fb^{-1}. An additional reduction of the size of the recorded datasets is done after certain good data quality monitoring.

3.3.3 Pileup

The recorded data can be increased by increasing the data-taking period and/or by increasing the number of particle per bunch. The second possibility results in the pileup, a phenomena when more that one pp collision happens during one bunch crossing. The number of pileup events follows a Poisson distribution with the mean μ. The mean number of simultaneous pp interactions per bunch crossing during the LHC Run 2 is shown in Figure 3.4.

3.4 LHC Performance

3.4.1 LHC Run 2

The LHC Run 2 started in April 2015 after the first two-year long shutdown, when the accelerator has undergone the improvements enabling the increase of the collision energy up to $\sqrt{\text{s}} = 13$ TeV and collision rate to 40 MHz corresponding to 25 ns bunch spacing. During the LHC Run 1, the machine operated at the lower collisions energies ($\sqrt{\text{s}} = 7$ TeV

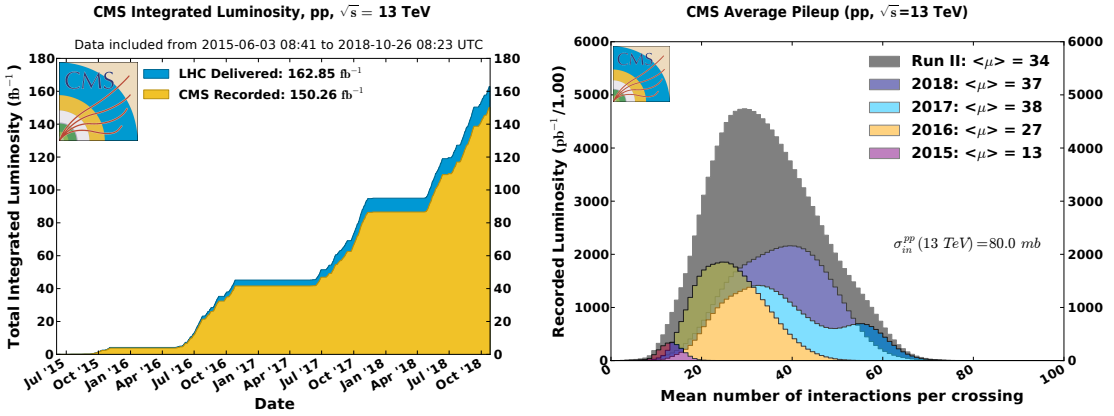

Figure 3.4 – Total integrated luminosity versus time for pp collisions at $\sqrt{s}$ = 13 TeV in the LHC Run 2 (left). The blue histogram corresponds to the delivered luminosity by the LHC, while the recorded luminosity by the CMS experiment is shown by orange histogram. Distribution of the mean number of interaction per crossing in the LHC Run 2(right).

and $\sqrt{s} = 8$ TeV) with larger bunch spacing (50 ns), therefore the Run in 2015 was used for the re-commissioning of the accelerator complex and detectors. The LHC Run 2 lasted for 4 year and successfully ended in December 2018 [90]. In total, the LHC delivered 160 fb^{-1} of the integrated luminosity to the IP1 and IP5, which house the ATLAS and CMS experiments. Such high number of delivered integrated luminosity was achieved by operating with the higher instantaneous luminosity than initially expected. The design value of instantaneous luminosity was already exceeded in 2016 and during 2018 the LHC routinely operated with the instantaneous luminosity twice larger than the design one. In addition, the stable beams when the collisions may occur were available for almost 50% of the machine time schedule. Table 3.1 summaries the LHC availability during the LHC Run 2.

Year	Stable Beam	Downtime	Operation
2015	33% (455h)	30% (426h)	37% (511h)
2016	49% (1840h)	26% (980h)	25% (919h)
2017	49% (1634h)	19% (653h)	32% (1075h)
2018	49% (1932h)	24% (943h)	27% (1069h)

Table 3.1 – The LHC availability breakdown for pp collisions [90].

The beam parameters were not constant during the LHC Run 2, but they changed and evolved as a result either of the improvement of machine performance or the influence of technical issues [90]. Already in July 2016, a new type of beam based on Bunch Compressions, Merging and Splitting (BCMS) became available and the transverse normalized emittance was reduced for the same beam intensity. This improvement helped to exceed the design value of instantaneous luminosity. Nevertheless the LHC also faced several technical inconveniences. During the winter stop in 2017 when the dipole magnet was replaced, 7 litres of air entered the beam vacuum and froze. It resulted in an abnormal background radiation and sudden beam losses. Temporary solution of this issue was the alternative beam production and new type of beam 8b4e [1] and later 8b4e-BCS was introduced. This

1. The bunch train consists of 8 bunches and 4 empty buckets.

issue was partly solved in 2018 and the LHC could continue the running with the BCMS beams. In this year, the recorded instantaneous luminosity of $2.07 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ was reached. Table 3.2 provides the overview of the main LHC Machine and beam parameters for the LHC Run 2.

Parameter	**Design**	**2015**	**2016**		**2017**		**2018**
Beam Type:	**Std**	**Std**	**Std/BCMS**	**BCMS**	**8b4e**	**8b4e-BCS**	**BCMS**
Energy [TeV]	7	6.5	6.5	6.5	6.5	6.5	6.5
Number of bunches per ring	2808	2244	2040/2076	2556	1916	1868	2556
Bunch population N_b [10^{11}p/b]	1.15	1.15	1.2	1.2	1.2	1.25	1.1
Trans. norm. emittance ε_n [mm·mrad]	3.75	3.5	3.5/2.1	2.1	2.3	1.8	2
Betatron func. at IP1 and IP5 β^* [m]	0.55	0.8	0.4	0.4	0.4/0.3	0.3	0.3/0.25
Half crossing angle [μrad]	142.5	145	185/140	150	150	150/120	160/130
Peak luminosity [10^{34} cm^{-2}s^{-1}]	1	0.55	0.83/1.4	1.74	1.9	2.06/1.5	2.1
Maximum pileup μ	∼20	∼15	∼20/35	∼45	70/60	80/60	60

Table 3.2 – Summary of the main LHC Machine and beam parameters for the LHC Run 2 and comparison with the design values. Adapted from [90].

3.4.2 Special runs

The LHC provides the possibility to study the hadronic matter under the extreme condition by colliding lead nuclei. Several special runs related to the heavy ion physics occurred during the LHC Run 2. Each run has been unique in terms of collision energy and colliding species. While the first dedicated heavy ion run during the LHC Run 2 happened in 2015, the second heavy ion run was in November 2018 [91] and ended the successful LHC Run 2 period. In both cases, the lead ions collided at the centre-of-mass energy per colliding nucleon pair of 5.02 TeV. Especially, the second heavy ion run was very successful and the LHC managed to deliver a dataset corresponding to $\sim 1800\mu\text{b}^{-1}$ to IP1 and IP5. Such large dataset enabled to study nuclear matter by measuring Z boson production [92] and the first evidence for top quark production in high-energy heavy ion collisions [93]. In addition to the lead collisions, in October 2017 the LHC performed the pilot run with the xenon ion $^{129}_{54}$Xe at a centre-of-mass energy per nucleon pair of 5.44 TeV [94].

In order to study the nuclear matter effects related to the QGP formation in heavy ion collisions, reference pp and p-Pb collisions are needed as it is being expected that the QGP is not formed in these collisions. Several low pileup pp, p-Pb and Pb-p runs at various centre-of-mass energies happened in 2016 and 2017 [95]. Reversal of the beams from Pb-p to p-Pb is needed by the ALICE and LHC*b* experiments as their detectors are asymmetric. Such runs were trying to shed lights on open questions from the LHC Run 1 and arising from the observations of several phenomena (the strangeness enhancement [96] and the collective behaviour [77]) in small collision systems which were originally proposed as a signature of QGP formation in nuclear collision [97].

The LHC is not only the most powerful hadron collider ever built, but also the most powerful light-light collider. The protons and heavy ions in the beam carry an electromagnetic field and its intensity is proportional to the atomic number Z^2. Such field can be viewed as a source of quasi free photons and photon-photon ($\gamma\gamma$) or photon-hadron (γp and γPb) interactions [98] may occur. These interactions offer an unique opportunity to study QED and QCD via photon-induced processes. The photon-hadron (γPb) interactions occurs in the so-called ultra-peripheral collisions (UPC) when the impact parameter is larger than the sum of nuclear radii. One of the interesting research area in UPC physics is the vector meson photoproduction which can be used to study saturation and shadowing effects [98].

In context of this thesis, we would like to highlight recent observation of photon-induced diboson production, $\gamma\gamma \to W^+W^-$ [99].

3.5 Compact Muon Solenoid detector

The CMS detector [100], shown in Figure 3.5, is a general multi-purpose detector built to cover the widest possible range of physics at the LHC. It is located at IP 5 of the LHC tunnel. The CMS detector has a cylindrical shape with the overall dimensions of the length of 28.7 m, a diameter of 15.0 m, and a total weight of 14 000 tonnes [100]. The multiple subdetectors are aligned layer-wise around the beam axis, symmetrically to the primary interaction point. The first layer, closest to the beam pipe, consists of the CMS tracker system. The second and third layer holds the electromagnetic and hadronic calorimeters. All these layers are surrounded by the large-bore superconducting solenoid. Finally, the muon system forms the last, fifth, layer.

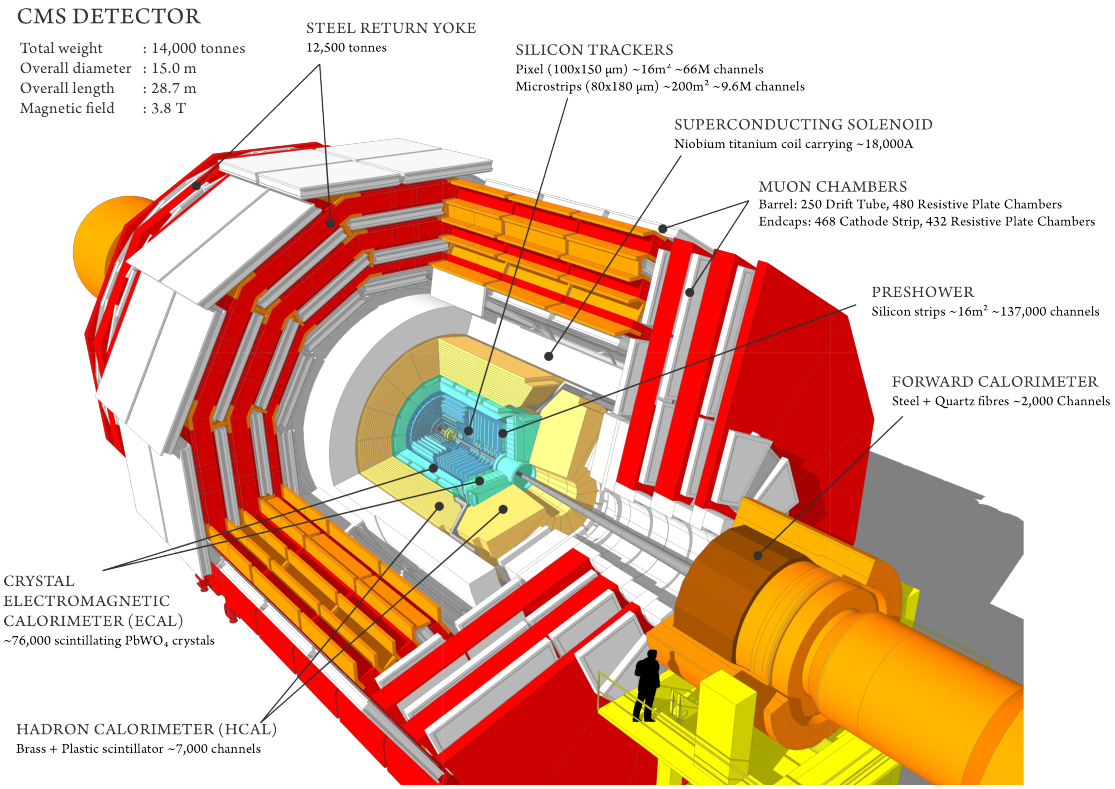

Figure 3.5 – The CMS detector with all its main components and characteristics. Taken from [100].

The CMS detector was designed with the following requirements:

- good muon identification and momentum resolution, good dimuon mass resolution and the unambiguous determination of the charge of muons with $p < 1$ TeV,
- good charged particle momentum resolution and reconstruction efficiency for efficienct offline tagging of τ's and b jets,
- good electromagnetic energy resolution, good diphoton and dielectron mass resolution,
- good missing transverse momentum and dijet mass resolution.

The coordinate system describing the positions and directions of measured objects is based on the right-handed Cartesian (x,y,z) coordinate system with the origin (0,0,0) located in the centre of the detector, where collisions occur. While the z axis is parallel to the beam in counter-clockwise direction, the x axis points to the centre of the LHC and the y axis is perpendicular to the beam direction as shown in Figure 3.6. The x and y axis form the transverse plane. However for the track measurement, it is more convenient to use the cylindrical coordinate system. The azimuthal angle ϕ, which is a angle between the x axis and the track in $x-y$ plane, is defined as zero in direction of the x axis. The angle between the measured point and the z axis is defined as the polar angle θ. This coordinate system is sketched in Figure 3.6 and can be used to define the pseudorapidity

$$\eta = -\ln\tan\frac{\theta}{2} = -\ln\frac{p+p_z}{p-p_z}, \tag{3.5}$$

where p and p_z are the absolute and z component of the particle momentum, respectively. For massless non-realistic particles, the pseudorapidity η is equal to the rapidity

$$y = -\ln\frac{E+p_z}{E-p_z}, \tag{3.6}$$

where E stands for the energy of particle.

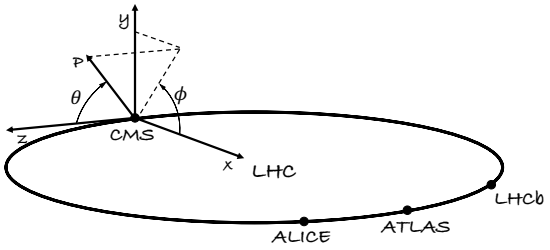

Figure 3.6 – The coordinate system used by the CMS detector.

The overview of the most important components of the CMS detectors will be briefly discussed in the following subsections.

3.5.1 Solenoid magnet

The main feature of the CMS detector is the 4 T superconducting solenoid magnet with a free bore of a diameter of 6 m and a length of 12.5 m. Even though, it is the largest solenoid magnet ever constructed, the size of the magnet system is relatively small for the magnetic field of 4T. Such a strong magnetic field is achieved by using the superconducting technology. Namely, its coil is made of superconducting NbTi conductors operating at a temperature of 4.7 K. The magnetic flux outside the magnet is returned into the detector volume by the three layers of massive iron yoke divided on the barrel and endcap parts. The muon system is embedded between the yoke layers. The main role of the iron yokes is to increase the homogeneity of the magnetic field in the detector volume. In addition, they play a role of the absorber for the muon system. The design value of the magnetic field is 4 T, but the magnet is operated at 3.8 T. As shown in Figure 3.7, outside of the detector volume, in the muon system the magnetic field has a strength of 2 T. The magnet system is necessary for the measurement of charged particle momentum p_T. When charged

particle travels through the magnetic field, its trajectory is bended by the Lorentz force. The radius R of curvature of the reconstructed track together with the information about the magnetic field strength can be then used to determine the transverse momentum of the particle as follows

$$p_{\mathrm{T}}\,[\mathrm{GeV}] = 0.3 \cdot B\,[\mathrm{T}] \cdot R\,[\mathrm{m}]\,. \tag{3.7}$$

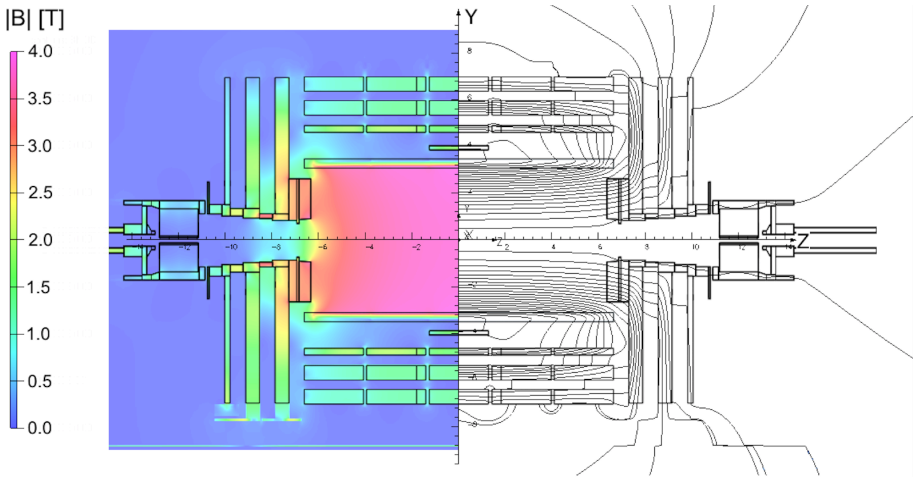

Figure 3.7 – The magnetic field strength $|B|$ (left) and field lines (right) of the CMS solenoid magnet [101].

3.5.2 Tracking system

The tracking system located closest to the interaction point is designed for the precise measurement of charged particles' momentum and the vertex reconstruction. The detection principle is based on silicon semiconductors. The charged particles passing through the silicon detectors create the hits which are recorded and then used to reconstruct the particles' trajectory, track. The reconstructed tracks point to the common points of the origin - vertex.

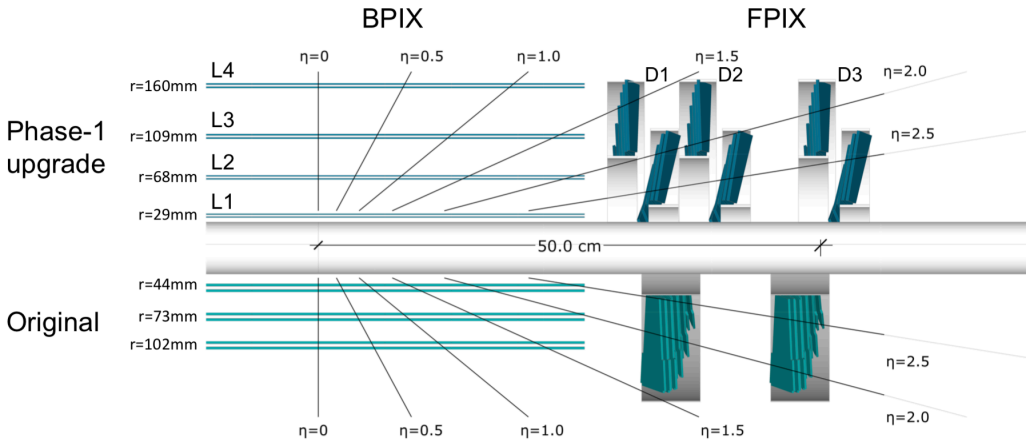

Figure 3.8 – Longitudinal layout of one quadrant of the CMS Phase-1 pixel detector compared to the original, CMS Phase-0, pixel detector layout [102].

The CMS tracking system, the so-called tracker, with a length of 5.8 m and a diameter of 2.5 m consists of the inner silicon pixel detector and the outer silicon strip detector. The

coverage of the tracker is up to $|\eta| < 2.5$. The original CMS tracker, known as Phase-0, has been replaced by the Phase-1 pixel detector between 2016 and 2017 during the extended year-end technical stop (EYETS) [103]. The conceptual differences between the Phase-0 and Phase-1 is shown in Figure 3.8. In the following paragraphs we will describe the Phase-1 pixel detector if not stated otherwise.

The inner silicon pixel detector consists of four barrel (BPIX) layers and three forward disk (FPIX) layers. The BPIX layers with the length of 54.0 cm are located at radii of 2.9, 6.8, 10.9, and 16.0 cm, while the FPIX layers are located at distances of 29.1, 39.6, and 51.6 cm from the centre of the detector [102]. In total the pixel detector is built from 1856 segmented silicon sensor modules. Each module consists of 66560 pixels with the size of $100 \times 150\,\mu\text{m}^2$ connected to 16 readout chips. Despite of adding one extra layer, the total material weight of Phase-1 pixel detector was reduced compared to Phase-0 pixel detector. The significant mass reduction was achieved by relocation of the passive material out in z direction and using an ultra-lightweight support with CO_2 cooling. During operation, the sensors are cooled below 0° C to mitigate radiation damage effects. Challenges for the inner silicon pixel arise from the high particle flux resulting from the increasing pileup and the radiation damage.

The inner silicon pixel detector is surrounded by the silicon strip detector located at the radial distance of 20 cm and 116 cm. The inner part of the silicon strip detector is made of Tracker Inner Barrel(TIB), four layers of silicon and strip pitch sensors, and Tracker Inner Disk (TID) situated at each side of the detector. Finally, the last layer of the tracking system consists of the Tracking Outer Barrel (TOB) and the Tracker End Cap (TEC).

3.5.3 Calorimetry system

The main purpose of the CMS calorimetry system is the excellent energy resolution, which is needed for the identification of hadrons, electrons, and photons. The CMS calorimetry system is formed by two detectors with different detection technique, namely the electromagnetic calorimeters (ECAL) and the hadronic calorimeters (HCAL).

The ECAL is constructed for the detection of photons and electrons by measuring the amount of the kinetic energy deposited in the absorbers. In particular, it measures the electromagnetic shower initiated by the photons or electrons when they are passing through the absorbers. The photons are typically converted in electron-positron pairs, while the electrons lose their energies by the bremstrahlung radiation. The distance needed for the photons or electrons to produce emissions is called the radiation length X_0. In case of electrons, the radiation length X_0 also corresponds to the average distance needed to reduce its initial energy E by a factor of $1/e$ due to the bremstrahlung radiation. The shape of the shower can be approximated by a cylinder and the Moliere radius characterizes the width of that cylinder containing 90% of the total shower's energy. The radiation length and the Moliere radius are the properties of the material and they are one of the main characteristics of the calorimeters.

The hadrons, heavier particles than electrons, can penetrate deeper into the calorimeter system and they can reach the HCAL. The strong interaction of hadrons with the detector materials results in a hadronic shower which is typically broader and deeper compared to the electromagnetic one and secondary particles are produced in the cascade. Consecutively, they undergo further inelastic collisions and the sequential decays of these particles can lead to an electromagnetic shower. The mean distance that is passed by the hadron before undergoing an inelastic nuclear collision is called as the nuclear interaction length λ_I.

Electromagnetic calorimeters

The ECAL is a hermetic, homogeneous electromagnetic calorimeter surrounding the tracking system. The design of this detector was driven and optimized for the decay of Higgs boson into two photons or four electrons ($H \to \gamma\gamma$ and $H \to ZZ^* \to 4e^{\pm}$). The great energy resolution and fine granularity of the ECAL detector enabled the discovery and later the precision measurement of Higgs boson in these decay channels. The main building block of the ECAL is the lead tungstate ($PbWO_4$) crystal. The choice of this type of crystal was motivated by its extremely short radiation length ($X_0 = 0.85$ mm) and small Moliere radius ($R_M = 2.19$ cm). Last but not least, it has very fast response - 80% of the light (signal) is emitted within 25 ns. Since the design bunch spacing in the LHC is 25 ns as discussed early, this is a very important feature of the lead tungstate crystals. On the other hand, the disadvantage of these crystals is a relative low light yield. Therefore the signal needs to be amplified by the silicon avalanche photodiodes or vacuum phototriodes.

The detector is formed by the central barrel (EB) closed by two endcaps (EE) at the each side. While the EB provides the coverage in pseudorapidity up to $|\eta| < 1.48$, the forward region up to $|\eta| < 3.0$ is covered by the EE as shown in Figure 3.9. In addition, better photon-π^0 separation is achieved by a preshower detector (ES) which is placed in front of the endcaps at $1.65 < |\eta| < 2.6$. The ES is built of lead absorber and silicon strips sensors and its thickness corresponds to $\sim 3X_0$.

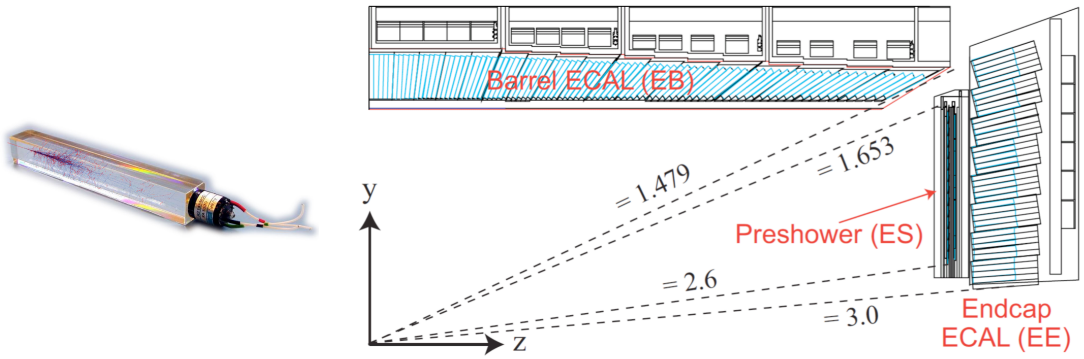

Figure 3.9 – Lead tungstate ($PbWO_4$) crystal with the read-out electronics (right) [104]. Longitudinal layout of one quadrant of the electromagnetic calorimeters (left) [105] and its individual components: barrel ECAL (EB), endcap ECAL (EE), and preshower detector (ES).

The crystals in EB with the overall length of 230 mm ($\sim 26X_0$) have a front face cross section of $2.2 \times 2.2\,cm^2$, while the EE crystals have slightly different dimensions - they are 220 mm ($\sim 25X_0$) long and the front face cross section is of $2.86 \times 2.86\,cm^2$. One of the crystal is shown in Figure 3.9. Thanks to the dimensions of the crystals, all photons and electrons are stopped and detected in the ECAL. In total, the ECAL is made of 75,848 crystals. The crystals act as absorbers as well as scintillators and they are transparent to entire emission spectrum. However, the coloured centres, created due to the irradiation, reduce the transparency of the crystals and lead to lower response. Even so the crystal transparency can be partially recovered due to self-annealing, it needs to be monitored for the ECAL calibration. The crystal transparency is examined by the laser system every ~ 30 minutes during the data-taking and dedicated corrections are derived for taking into account the transparency loss. Validation of these corrections relies on *in situ* measurement

of the reconstructed mass resolution of η mesons decaying into photon pairs ($\eta \to \gamma\gamma$) and the energy resolution determined from $Z \to e^+e^-$. Figure 3.10 shows the results from the measurement of the relative response of the crystals to laser light. As one can see there is a clear degradation of the crystal transparency, especially for the higher pseudorapidity η region.

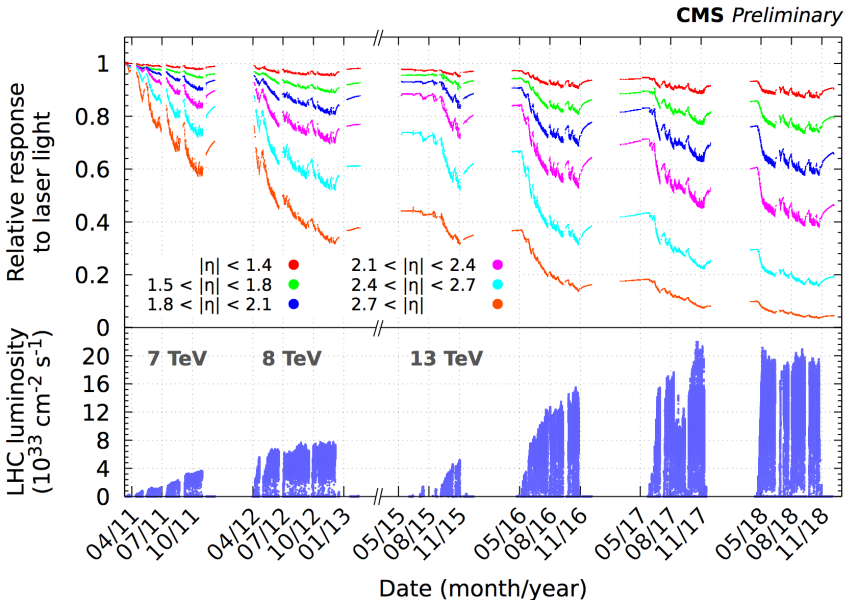

Figure 3.10 – Relative ECAL crystal response to laser light versus time in LHC Runs 1 and 2 [106], for several bins of pseudorapidity η.

The energy resolution is typically parametrized by

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2, \tag{3.8}$$

where S is the stochastic term, N is the noise term and finally C is the constant term. The electronics and digitization noise together with the pileup contributes to the noise term, while the fluctuations in the energy deposited in the preshower absorber and event-by-event fluctuations in the lateral shower containment are encoded into the stochastic term. The constant term includes the information mainly about the intercalibration errors. The ECAL energy resolution has been measured with the test electron beams. The electron energy was reconstructed by summing the signal from 3×3 crystal array around a centrally hit crystal and the energy resolution has been found to be

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E}}\right)^2 + \left(\frac{12.8\%}{E}\right)^2 + (0.3\%)^2\,. \tag{3.9}$$

Here E is measured energy in GeV.

Hadronic calorimeters

The HCAL is a sampling calorimeter, designed for precise measurement of the hadronic jet energy. In addition, thanks to pseudorapidity coverage up to $|\eta| < 5.2$ it is very important for the accurate measurement of the missing transverse energy. It consists of

layers of brass, steel and plastic scintillator. While the brass and steel are utilized as the absorber, the plastic scintillator is the active detector material. As shown in Figure 3.11, the HCAL is composed of four subdetectors mounted in different regions. The HCAL barrel (HB) surrounds the EB and covers the pseudorapidity up to $|\eta| < 1.3$. It consists of 14 layers of brass plate enclosed by two steel layers and 17 layers of plastic scintillators. The segmentation of plastic scintillators provides the final division of the HB into the so-called towers with the size of 0.087×0.087 in $(\Delta\phi, \Delta\eta)$. The depth of the HB expressed in units of the nuclear interaction length λ_I depends on pseudorapidity η and ranges from $5.8\,\lambda_\mathrm{I}$ for $|\eta| \sim 0$ to $10.6\,\lambda_\mathrm{I}$ for $|\eta| \sim 1.3$.

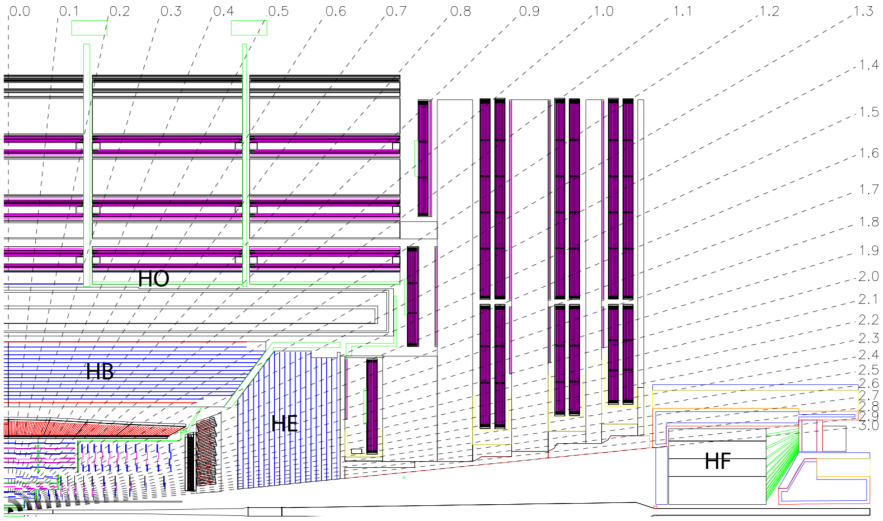

Figure 3.11 – Longitudinal layout of one quadrant of the hadronic calorimeters (HCAL) [83]. The HCAL consists of the HCAL barrel (HB) and HCAL endcap (HE), both of them located inside the solenoid magnet. The HCAL outer (HO) and forward hadron calorimeters (HF) surround the solenoid magnet.

The HCAL endcap (HE) is located inside the solenoid and extends the pseudorapidity coverage up to $|\eta| < 3.0$. Similarly to the HB, the HE is made of 16 layers of the absorbers and 17 layers of plastic scintillators. In the pseudorapidity coverage of $1.6 < |\eta| < 3.0$, each tower of the HE covers 0.17×0.17 in $(\Delta\phi, \Delta\eta)$. Due to limited space inside the solenoid magnet, two remaining subdetectors of the HCAL are mounted outside the magnet. The HCAL outer (HO), covering the $0.4 < |\eta| < 1.6$, works as a tail-catcher, in order to cover all hadronic showers leaking from the HB. The massive solenoid magnet acts as absorbers. By prolonging the effective depth of the hadron calorimetry system in the midrapidity $|\eta| \sim 0$ up to $11.8\,\lambda_\mathrm{I}$, the HO improves the energy resolution in the central region. The energy resolution of single pions measured during the test beam corresponds to

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{85\%}{\sqrt{E}}\right)^2 + (7\%)^2\,. \tag{3.10}$$

Compared to the ECAL, the response and resolution of the HCAL is worse, since most of the hadron start to lose their energy already in the ECAL.

The precise measurement of the missing transverse energy requires a hermetic calorimeter system as much as possible. Thanks to the forward hadron calorimeters (HF), the pseudorapidity coverage is increased up to $|\eta| < 5.2$, where a high particle flux is expected.

Therefore the active material of the HF is the quartz-fiber scintillator, which is well known for its radiation hardness. The quartz-fiber scintillators together with the layers of steel used as absorbers are arranged in 900 towers with length of 165 cm which are installed parallel to the beam line at a distance of 11.2 metres from the centre of the detector. In contrast to the other HCAL subdetectors, the HF produces a narrower and shorter shower which is well suited for a high particle flux environment. In addition, using of two read-out channels in different depth of the HF enables to detect both electromagnetic and hadronic showers. Unfortunately, the energy resolution of the HF is even worse than the rest of the HCAL system. The stochastic term for the electromagnetic and hadronic shower is ca 200% and 300%, respectively, while the constant term is about 10% for both type of showers.

3.5.4 Muon system

Since the muons are able to pass the whole inner detector (the tracker, the ECAL, the HCAL) with minimal energy loss, the muon system can be mounted as the outermost layer of the CMS detector. It surrounds the solenoid magnet as shown in Figure 3.12 and is interlaced with the three massive steel flux-return yokes, which return the magnetic flux into the detector volume and in addition act as absorbers to stop particles other than muons. The main objective of the muon system is the muon identification together with the measurement of their momenta and charge. The detection principle is based on gaseous detectors. When a charged particle, in our case a muon, is travelling through the detector, it ionizes the gas and initializes the electron avalanches, signals. The muon system incorporates three type of gaseous detectors. In the barrel region, where the low rate of the muon is expected, the drift tubes (DT) chambers are located. They are arranged in four stations, covering the pseudorapidity $|\eta| < 1.2$, and complemented by five layers of the resistive plate chambers (RPC). Three innermost stations contain twelve drift tubes, while the fourth station is built only of eight drift tubes. Moreover, the last station is designed only for the measurement in $r - \phi$ direction. The muon system in the forward region is made of cathode strip chambers (CSC) and the RPC. Thanks to the faster response, fine segmentation and the harder radiation hardness, the CSCs are well suited for the forward region, where the muon flux is larger and the magnetic field is stronger and more inhomegeneous. The CSCs extend the pseudorapidity coverage of the muon system up to $|\eta| < 2.4$. In total, the muon system covers more than 25000 m^2 by detection planes.

As CMS's acronym implies, the precise muon reconstruction is of particular interest for the CMS detector. Compared to the RPC, the DT and CSC provide better spacial resolution of the muon, while the faster response and timing of the RPC is beneficial for the muon triggering. In addition to the muon system, the information from the tracker system is incorporated to improve the muon reconstruction.

The performance of the muon system is studied with events including $\mathrm{J}/\psi \rightarrow \mu^+\mu^-$ and $\mathrm{Z} \rightarrow \mu^+\mu^-$. Overall, the reconstruction efficiency is above 95% for muon with the transverse momentum larger than a few GeV. The spacial resolution of the muon track was found to be $80 - 120\,\mu$m for the DT and $40 - 150\,\mu$m for the CSC. The relative p_{T} resolution ranges from 1.3% to 2.0% for muons in the barrel and is better than 6.0% in the endcap region. Without the tracking system, the relative resolution is a bit worse, but still better than 10.0% in the barrel region [107]. Excellent performance of the CMS muon system in LHC Run 2 enabled the first evidence for Higgs boson decay to a pair of muons [108].

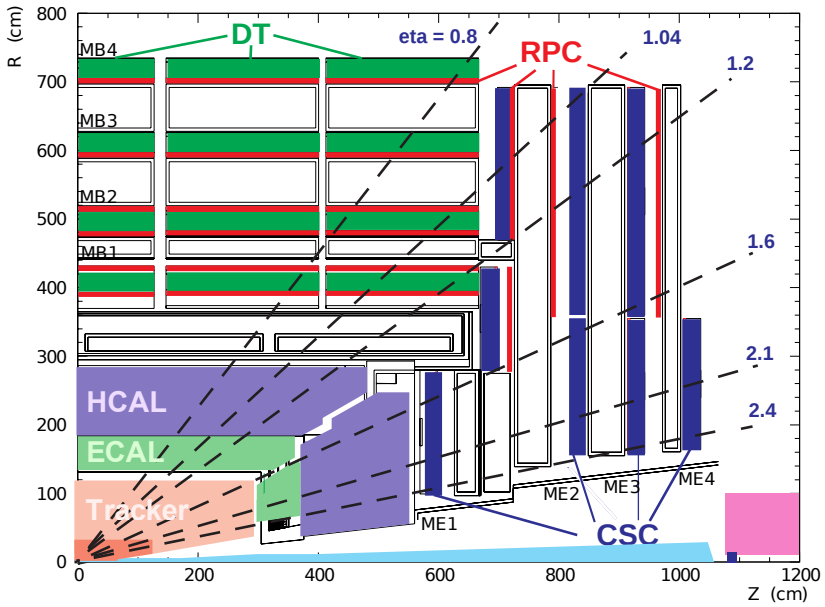

Figure 3.12 – Longitudinal layout of one quadrant of the muon system [107]. The four drift tubes (DT) stations in the barrel region, four cathode strip chambers (CSC) in the forward region, and the resistive plate chambers (RPC) plates embedded into the muon system.

3.5.5 Trigger system

The design luminosity of the LHC would lead to 10^9 of proton-proton collisions per second corresponding to the dataset of size of hundreds of TB/s. It is clear that it is impossible to record, store and further analysis such large amount of data. In addition, only a small fraction of these collisions contain events of interest for the CMS physics programme. It is the crucial job of the CMS trigger system to select events of particular interest for offline analysis and reduce the amount of the data at least by a factor of 10^6. The resulting data output is ~ 100 MB/s. The CMS trigger system uses a two-tiered system, the first level (L1) and second level (high-level trigger, HLT).

The L1 trigger is a fully hardware system with a latency of $4\mu s$ [109]. A schematic layout of the L1 trigger is shown in Figure 3.13. It uses information from the ECAL, HCAL and the muon system to construct the trigger primitives (TPs), which are further processed before an event is accepted or rejected. The calorimeter trigger path reconstructs the e/γ candidates, finds τ leptons and jets. In addition, it calculates the global quantities such as the missing transverse momentum or the scalar sum of the jet p_{T}. The muon trigger path combines the information from all muon subdetectors, namely from the DT, CSC, and RPC, and reconstructs the four most energetic muons. The trigger objects received from the calorimeters and muon paths are then sent to the L1 global trigger (L1 GT) system, the final step of the L1 trigger system. Based on the input informations, the L1 GT decides to reject or accept an event for subsequent evaluation by the HLT. The L1 trigger does not include only the triggers for the event selection, but also technical triggers developed the monitoring and calibration purposes of the CMS subdetectors.

The HLT is the second level of the CMS trigger system. Compared to the L1 trigger, the physics object reconstruction at the HLT is done in a more precise way and the

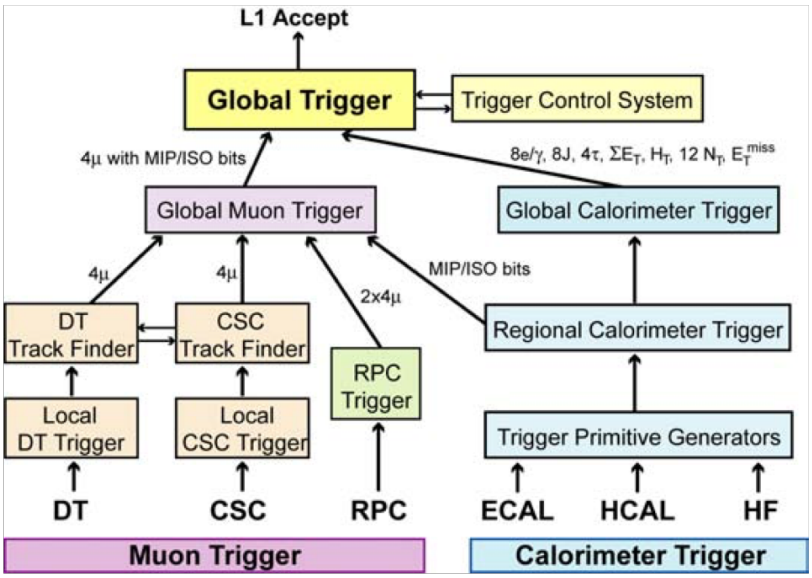

Figure 3.13 – Sketch of dataflow for the L1 trigger system [83].

employed algorithms and software are similar to that used for the offline processing and data analysis. However, the HLT system still needs to be very fast, hence there are some simplification in the reconstruction algorithms which result in a worse precision of the reconstructed objects. The HLT is structured around the various trigger paths, the so-called HLT paths. The trigger path is a set of algorithms which reconstruct the physics objects and make a decision of the event selection based on given condition. An event is stored only if at least one trigger path accepts this events. The output rate of the event accepted by the HLT is typically around $\sim$ 100Hz. For a certain physics processes, which are very abundant thanks to very high cross section, a further reduction is still needed. As example of such process, we can mention the QCD processes with the hadronic jets in the final states. An additional event reduction is achieved by the trigger pre-scale, i.e. only a certain fraction of events are in the end kept. The accepted events are stored locally on disk or transferred to the CMS Tier-0 computing farms, when the offline reprocessing of the events and permanent storage is performed.

4

Object reconstruction and identification

When particles are passing through the detector, they create signals in the different subdetectors depending on their type. Figure 4.1 shows the signals produced by particles in the detector. The trajectory of charged particles is bended by the magnetic field and they create hits in the tracking system. The hadrons are detected mainly as deposits energy in the ECAL and HCAL, while electrons and photons leave the signal only in the ECAL. The muons, most penetrating particles, are able to traverse through the whole inner detector and create hits in the muon system mounted as the outermost layer of the CMS detector. The readout signals, hits and the clusters of deposited energy, from all of the subdetectors of the CMS experiment are then used by the reconstruction algorithm to build the physics objects.

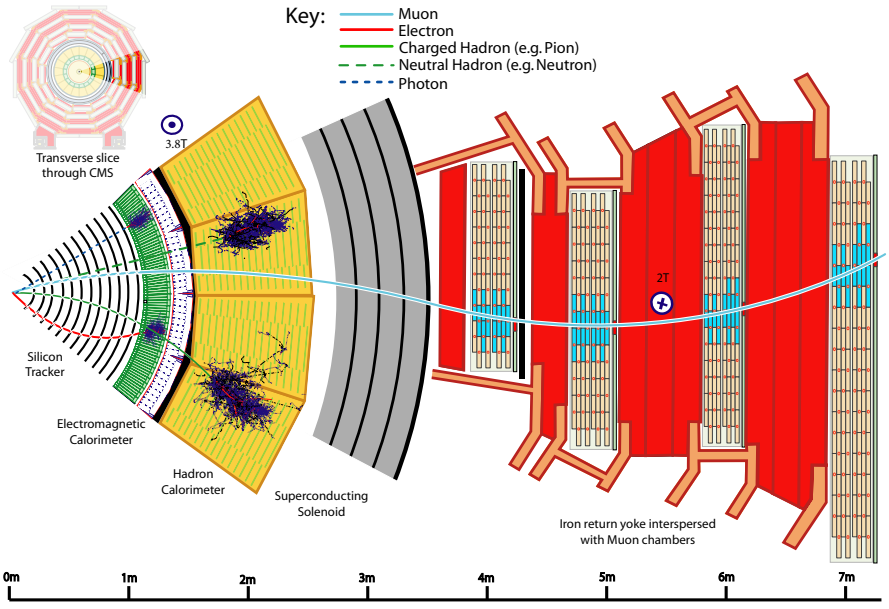

Figure 4.1 – Transverse slice of the CMS detector and illustration of specific interaction of particles with the different subdetectors [110].

4.1 CMS Particle-Flow algorithm

The CMS experiment utilizes the particle-flow (PF) event reconstruction algorithm originally developed and for the first time used by the ALEPH experiment at LEP [111]. Despite the fact that the PF algorithm was designed mainly for e^-e^+ collisions, the properties of the CMS detector like the precise tracking system, the fine-grained ECAL and hermetic HCAL, the strong magnetic field together with the muon system enables to use the PF algorithm even for pp collisions. For the event reconstruction, the CMS PF algorithm combines and correlates the signals of the all subdetectors to reconstruct and identify the final-state particles as schematically shown in Figure 4.2. The basic elements of the PF algorithm are the trajectories of the charged particle in the tracker system, the clusters of deposited energy in the calorimeters and the muon tracks from the muon system.

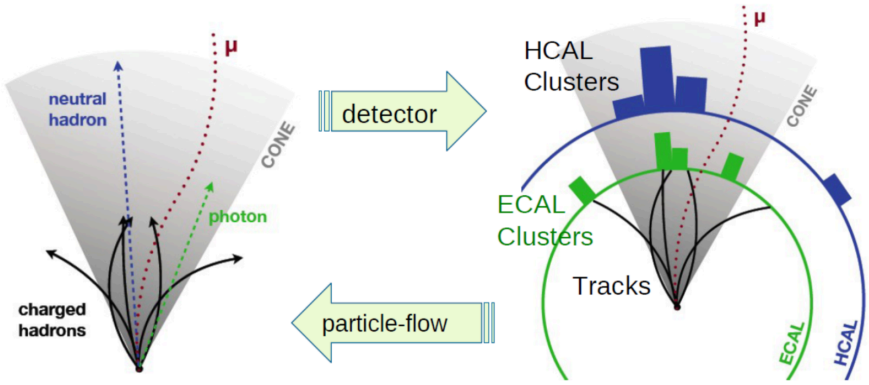

Figure 4.2 – Illustration of Particle Flow algorithm [110]. Each particle create signals in the different sub-detectors (right) and the Particle Flow algorithm combines and correlates the signal in order to reconstruct and identify the particles (left).

The PF algorithm consists of three consecutive steps. First, information from all subdetectors is analysed and processed to create charged particle tracks, muon tracks and clusters of deposited energy. The charged particle tracks are reconstructed using a Combinatorial Track Finder [112] based on a Kalman filter [112]. It is an iterative approach starting with the tracks made by particles with large transverse momentum, since they are easier to reconstruct. The same reconstruction approach is also employed for the muon track in the muon system. The reconstruction of calorimeter clusters is performed separately for each sub-detector of the calorimetry system enabling to fully exploit the fine granularity of the system. The main objective of the clustering algorithm is to detect and measure energy and direction of stable neutral particles. In addition, the separation of these neutral particles from charged hadron energy deposits helps with the measurement of charged hadrons with larger transverse momentum. The calorimeter clusters are built in two steps. Firstly, the cells with an energy larger than the given threshold and larger than the energy of the neighbouring cells are identified as cluster seeds. The neighbouring seeds are then iteratively added to the cluster seeds only if the resulting energy is twice larger than the noise level. In the end of this iterative procedure, topological clusters are formed. The profile of topological clusters are fitted to the Gaussian-mixture model which helps to determine the energy induced by the individual particles.

The reconstructed PF elements, tracks and clusters, are then connected by the link algorithm. In order to reduce the computing time, only the nearest neighbours of possible

pairs of elements in the $\eta - \phi$ plane are considered. A key part of this procedure is to avoid double counting. If several tracks are linked to one cluster, only the combination with the lowest distance is kept. Finally, the created PF candidates are used for the identification and reconstruction of the following physics objects.

Muons

The muon identification [113] benefits from the combination of available information from the muon system and the tracking system. According to the subdetectors used for the muon identification, we can group the reconstructed muons into the following categories. When only information from the muon system is used, then the reconstructed muon is called as a *standalone-muon*. The trajectory of the reconstructed track by the tracking system can be extrapolated to the outer muon system and if such track matches at least with one muon segment in the muon system, the reconstructed objects is identified as a *tracker-muon*. Contrary to the tracker-muon, the reconstruction of *global muon* is done in the opposite direction. The track from the outer muon system is extrapolated to the tracking system and the global muon is formed by the matching track from the muon and tracking system. The PF candidates that make up the reconstructed muons are not any more used for further physics object identification.

Electrons

The electron identification [114] relies on the information provided by the tracking system and the ECAL. The fact that these two subdetectors are mounted inside the strong magnetic field leads to an energy loss of the electron by emitting the bremsstrahlung photons. In addition, the interaction of electrons with the detector material causes additional emissions of bremsstrahlung photons. For these reasons, the deposited energy of electrons in the ECAL is more spread in the $\phi-$direction making the electron identification more challenging. The electron energy together with the possible bremsstrahlung photons is grouped in a supercluster (SC) which is reconstructed around the direction of the electron in a small region of η, but relatively large region of ϕ. In contrast with other charged particles, electron tracks are reconstructed by the Gaussian-sum filter (GSF) [115] as it is a more accurate algorithm for sudden energy losses along the electrons' trajectory caused by bremsstrahlung radiation. The GSF tracks are linked with the SCs and the electron momentum is determined.

Hadrons and photons

The remaining PF candidates are considered as hadrons from jet fragmentation and hadronization. While the linked tracks to the HCAL clusters make up the charged hadrons, the neutral hadrons are built up from the isolated HCAL clusters. In case of the charged hadrons, the energy measured by the calorimeters needs to be compatible with the momentum determined from the tracking information. Outside of the tracker acceptance ($|\eta| > 2.5$), the charged and neutral hadrons cannot be distinguished. Finally, the ECAL clusters not linked to any tracks are identified as photons. Depeding on the identification criteria, we distinguish between isolated and non-isolated photons like those from π^0 decays.

After the final-state particles are identified and reconstructed, the more complex physics objects like jets and missing transverse energy will be determined. However, before doing

so, the primary vertex is reconstructed and the contribution from the pileup needs to be removed.

4.2 Primary vertex

A primary vertex (PV) refers to the point in the space where the hard scattering of parton occurs. Due to the PU, one event during the LHC Run 2 typically includes multiple vertices. And all of them need to be reconstructed as they look indistinguishable from the primary vertex. The vertex reconstruction[116] is based on the information of the tracks. Only high quality prompt tracks (applying certain requirements on the track quality and number of hits in the inner tracker) are used. Tracks close to each other and which appear to originate from the same vertex are clustered by using a deterministic annealing algorithm [117]. After that the candidate vertices with at least two tracks are refitted by an adaptive vertex fitter [118] and the vertex parameters like the space coordinates and parameters related to the quality of the reconstructed vertex are calculated. In this algorithm, each track is assigned a weight w between 0 and 1 corresponding to the likelihood of its origin from the given vertex. Number of degrees of freedom in the adaptive vertex fitter is defined as

$$n_{\text{dof}} = -3 + 2 \sum_{i=1}^{\text{N}_{\text{tracks}}} w_i. \tag{4.1}$$

The value of n_{dof} is correlated with the number of tracks originating from the vertex and is used as one of the selection criteria of the primary vertex identification. In most of the CMS analyses and analyses presented in this thesis, the primary vertex is required to be reconstructed at least from four tracks, i.e. $n_{\text{dof}} > 4$, and the position of vertex must satisfy $|z| < 24\,\text{cm}$ along the beam pipe from the nominal centre of the detector and $|\rho| < 2\,\text{cm}$ in the transverse plane. The vertex satisfying those criteria with the largest value of $\sum p_{\text{T}}^2$ of physics objects associated to it is then selected as the primary vertex. The remaining reconstructed vertices are referred as PU vertices.

4.3 Pileup mitigation

To study one specific hard parton scattering viewed as pp collision, it is necessary to minimize the contribution coming from the pileup. The successive increase of instantaneous luminosity during the LHC operation goes hand in hand with larger probability of multiple pp collisions to occur during one bunch crossing leading to larger pileup contribution. The importance of more sophisticated pileup mitigation techniques is arising from the needs to reduce the impact of pileup contribution on the physic objects reconstruction and performance.

4.3.1 Charged hadron subtraction

In the previous section, we have depicted the main idea behind the PV reconstruction. As stated, also PU vertices are known once the PV is found. Then the charged hadrons are discarded from the list of PF objects for further physics object reconstruction, if they are associated to PU vertices. This mitigation technique is referred as charged hadron subtraction (CHS) and is widely used by CMS. The CHS technique is applicable only

to the charged hadrons reconstructed within the tracker acceptance ($|\eta| < 2.5$). Neutral hadrons and charged hadrons not associated to any verteces are not rejected. Hence, the CHS does not mitigate the whole contribution from the PU. In case of jets, the remaining contribution is corrected by dedicated jet-area-based correction acting on the jet four-momentum. Therefore the jet substructure variables can be still affected by the PU contributions. To overcome these limitations, more advance technique is needed.

4.3.2 Pileup per particle identification

The pileup per particle identification (PUPPI) [119] is new method built on the CHS technique targeting to suppress the contribution of any kind of hadrons originated from pileup. During LHC Run 2, this method was very extensively validated and commissioned by CMS [120] and it will become a new standard for LHC Run 3. By combining the global event information as well as local information about the particle distribution, the PUPPI algorithm calculates the so-called PUPPI weight for rescaling the four-momentum of particles. These PUPPI weights range from 0 to 1 and correspond to the likelihood of the particle to originate from PV or PU. The value of 1 is assigned to particles considered to come from the PV, while the particles from the PU are weighted by factor of 0. For charged hadrons, the PUPPI algorithm works identically as the previously discussed CHS algorithm. In addition, the charged particles not associated with the PV nor PU vertices, but with the distance of closest approach to the PV along the z axis smaller than 0.3 cm carries the weight of 1.

In case of neutral hadrons and particles outside the tracker acceptance, the PUPPI weight is assigned based on the local shape variable α which characterises the neighbourhood of given a particle. As a result of the collinear structure of a QCD parton shower, the particles produced by the shower are typically close to other particles from the same parent process. On the other hand, the PU particles are uncorrelated with the particles from the PV and can be distributed more homogeneously. Defining local shape variables α_i for a given particle i as

$$\alpha_i = \log \sum_{j \neq i, \Delta R_{ij} < R_0} \left(\frac{p_{\mathrm{T},j}}{\Delta R_{ij}} \right)^2 \begin{cases} \text{for } |\eta_i| < 2.5,\ j \text{ are charged particles from PV,} \\ \text{for } |\eta_i| > 2.5,\ j \text{ are all kinds of reconstructed particles,} \end{cases} \tag{4.2}$$

we can differentiate between particles with different neighbourhood. By construction, the particles close to the particles from the PV gets larger α. Outside the tracker acceptance, the larger value of α indicates a highly energetic particle in neighbourhood of a given particle. The charged particles from the PU vertices are used to calculate the expected PU distribution and then a median $\bar{\alpha}_{\mathrm{PU}}$ and root-mean-square $\alpha_{\mathrm{PU}}^{\mathrm{RMS}}$ are extracted. Since the tracking information is available only for the detector region corresponding to the $|\eta| < 2.5$, the $\bar{\alpha}_{\mathrm{PU}}$ and $\alpha_{\mathrm{PU}}^{\mathrm{RMS}}$ are extrapolated to the higher region of $|\eta|$ by a transfer function based on MC simulations. The need of a transfer function is arising from a variation of α with η as a result of different granularity of sub-detectors in various η region. Afterwards, the α_i of each neutral particle and particle outside the tracker region can be translated into a probability by using a signed χ^2 approximation defined as

$$\text{signed } \chi_i^2 = \frac{(\alpha_i - \bar{\alpha}_{\mathrm{PU}})\,|\alpha_i - \bar{\alpha}_{\mathrm{PU}}|}{\left(\alpha_{\mathrm{PU}}^{\mathrm{RMS}}\right)^2}. \tag{4.3}$$

Finally, the PUPPI weight is given by the cumulative distribution function of the signed χ^2 distribution with one degree of freedom. Formally written as

$$w_i = F_{\chi^2,\mathrm{NDF}=1}\left(\text{signed } \chi^2_i\right). \tag{4.4}$$

In the end, the particles coming most likely from the PU, i.e. with weight of 0.01, are rejected. In order to remove the residual dependency of the jet energy on the reconstructed number of vertices N_{vertices}, the particles fulfilling the condition on $w_i \cdot p_{\mathrm{T},i} < (A + B \cdot N_{\mathrm{vertices}})$ GeV are also removed. Overall, the PUPPI algorithm includes four tunable parameters: A, B and 2 transfer functions. The tuning of these parameters is done separately in three η regions, namely for $|\eta| < 2.5$, $2.5 < |\eta| < 3.0$, and $3.0 < |\eta| < 5.0$. As optimal values of these parameters are such values for which the jets built from the PUPPI-weighted PF candidates, so-called PUPPI jets, in a region of $|\eta| < 3.0$ have a jet response close to unity and PU independent. The parameters for the forward region are tuned by optimizing the resolution of missing transverse energy.

Two main parameters of the PUPPI algorithm are shown in Figure 4.3. The separation power of the PUPPI algorithm is shown by the comparison of the α distribution for the charged particles from the PV and PU. Most of the charged particles from the PU have a significantly lower value of α, while the double peak structure can be observed for the distribution made of the charged particles from the PV. The first peak located at larger α corresponds to the collimated particles inside the jets. The isolated particles create the second peak at lower α. The right plot shows the distribution of the PUPPI weights.

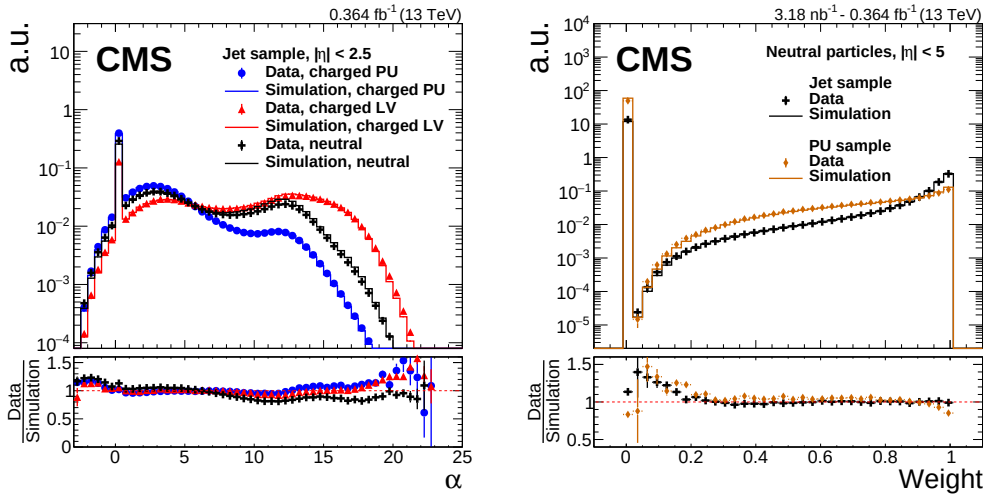

Figure 4.3 – The comparison of the α distribution for the charged particles from the PV and PU [120] (left) and the distribution of the PUPPI weights measured in Jet and PU sample [120](right).

4.4 Jet reconstruction

In the early days, jets in hadron-hadron collisions have been reconstructed only from the calorimetry information. The CMS Experiment benefits from the PF algorithm and as input to jet reconstruction uses all reconstructed particles. As shown in Ref. [110], it results in a significant improvement in jet angular and energy resolution with respect to

calorimeter jets mainly because of the more accurate measurement of the jet charged-hadron momentum in the PF algorithm. By default in CMS, jets are reconstructed by the anti-k_T clustering algorithm with the jet distance parameter of $R = 0.4$ and $R = 0.8$, respectively. According to the applied pileup mitigation technique, the reconstructed jet can be divided into three groups:

- PF jets - all particles reconstructed by the PF algorithm are used for the jet reconstruction,
- PF+CHS jets - charged hadrons associated to the PU vertices are removed from the jet clustering,
- PF+PUPPI jets - particles' four-momenta are rescaled by the PUPPI weight.

The default jet collections used by CMS in LHC Run 2 consists of the anti-k_T PF+CHS jets with $R = 0.4$ and anti-k_T PF+PUPPI jets with $R = 0.8$. When using MC simulations, we can also construct the simulated particle level (ptcl) jets by applying the same clustering algorithm to all stable particles excluding undetectable neutrinos. In case of CMS, as stable particles are considered those particles with the proper lifetime $c\tau > 1$ cm.

The small cone-size jets are widely used in most CMS analyses and therefore dedicated calibrations like the jet energy corrections and b tagging scale factors for these jets are centrally provided by the corresponding physics objects groups. Large cone-size jets are used mainly for the measurements with the Lorentz boosted heavy SM particles and in studies focused on the jet substructure. In contrast to the small cone-size jets, the PUPPI algorithm for the PU mitigation is preferred for the large cone-size jets. Figure 4.4 shows the comparison of the jet energy resolution and the median soft drop jet mass for the different pileup mitigation techniques. The jet energy resolution is defined as Gaussian width of the ratio between matched reconstructed and particle-level jet energy parametrized as function of the particle-level jet p_T. The pileup contribution has an impact mainly on the low p_T jets, and therefore the PF jets have the worse resolution, while the jet energy resolution for the PF+CHS and PF+PUPPI jets is competitive over a wide range of jet p_T. The main advantage of the PF+PUPPI jets is pronounced once we start to study jet substructure variables. Figure 4.4 left shows the median soft drop jet mass for the PF+CHS and PF+PUPPI jets as the function of number of vertices. The used MC sample is enriched with W boson jets, therefore the median soft drop jet mass is expected to be around the actual W mass and the amount of PU contribution is proportional to the number of reconstructed vertices. Since the CHS technique does not mitigate the whole contribution from PU and additional jet-area-based correction acts on the jet four-momentum, the soft drop jet mass or any other jet substructure variable is still affected by PU. As one can see, the median soft drop jet mass for the PF+PUPPI jets is more stable as function of the number of vertices.

4.5 Jet energy calibration

The measured energy of the reconstructed jets can differ from the true energy of the parton initiating the jets for various reasons. As already discussed, the PU, but also other detector-related effects like non-linear detector response or just the noise of the electronics can influence the energy of the reconstructed jets. Therefore the energy of measured jets needs to be calibrated. However the complexity of jets as the physics objects consisting of numerous reconstructed objects and sensitive to various detector effects makes the precise measurement of jet energy corrections (JEC) a challenging task. The CMS collaboration uses a factorized approach, where various corrections are derived in consecutive steps in a

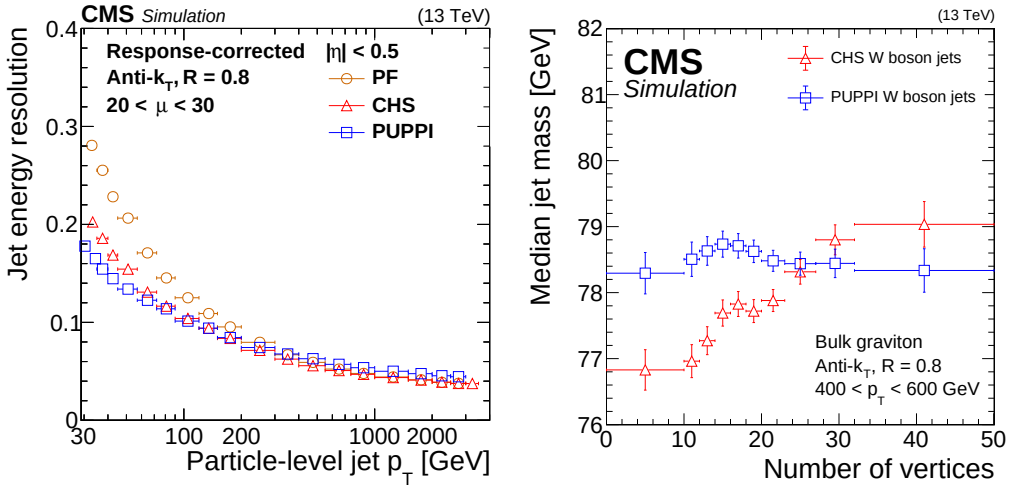

Figure 4.4 – Jet energy resolution as a function of particle-level jet p_{T} for three different type of jets, namely PF, CHS, and PUPPI [120](left). The median jet mass of W boson jets for CHS and PUPPI jets as a function of number of vertices [120](right).

given order. Figure 4.5 illustrates these individual steps from the beginning, reconstructed jets, to the end, where the jets are fully calibrated. While the first two corrections, the pileup offset correction and simulated response corrections are used in data and MC simulations, the residual corrections are applied only to the jets in data. The JEC are parametrized as a function of jet p_{T}, η, area A, and the offset energy density ρ.

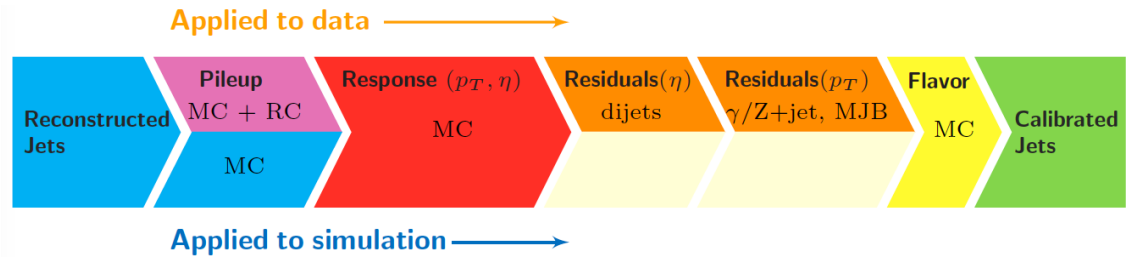

Figure 4.5 – Illustration of the factorized approach of derivation of jet energy corrections [121].

In addition, the jet p_{T} resolution is also measured. The difference between data and MC simulations is corrected by dedicated scale factors.

Having precisely calibrated jets are one of the key ingredients for the most of CMS analyses including the measurement presented in this thesis. Therefore, in the following subsections each of the steps of the derivation of the jet energy calibration will be discussed. More details and comprehensive description can be find in Ref. [121].

The notation used in the next subsections is following: the square brackets [] denote the binning variables, while the averaging within those bins for the variables that are used to parametrized the correction is denoted by the angle brackets $\langle\rangle$. If not stated otherwise, we discuss the corrections for the anti-k_{T} PF+CHS jets with $R = 0.4$ as they are the most widely used within the CMS.

4.5.1 Pileup offset corrections

The pileup offset corrections are important for jet calibration in hadron-hadron collisions, and of particular interest at the LHC, where the very high instantaneous luminosity results in very high PU. In case of the CMS detector, we distinguish between in-time pileup (IT PU) and out-of-time pileup (OOT PU). In the first case, the IT PU contribution results of two or more pp collisions within one bunch crossing, while the OOT PU is more related to the detector parameters and the finite signal decay time in calorimeters. It is possible that two different pp collisions happen in two subsequent bunch crossings and both of them create signals in the sub-detectors in the same time window as the primary collision. Despite of using the CHS technique for the PU removing, some remaining PU contribution is still present and leads to an additional energy offset of the reconstructed jet. The amount of PU contribution is proportional to the number of reconstructed vertices and the diffuse offset energy density ρ. For the pileup offset corrections, the latter variable is preferred as the vertex reconstruction can be affected by tracking inefficiencies. The offset energy density ρ is evaluated as a median of the energies calculated in a grid of $\eta - \phi$ cells. Alternatively and as done in LHC Run 1, ρ can be also defined in a given event as the median of jet p_{T} divided by their area $A \approx \pi R^2$ formally written as $\rho = \mathrm{median}\,(p_{\mathrm{T},\,i}/A_i)$. Using the median instead of mean makes ρ insensitive to high energetic jet. The former evaluation of ρ is simpler and does not require the jets clustering any more, therefore is nowadays preferred.

From two MC simulations containing exactly the same events with and without pileup overlay, we can calculate the offset caused by pileup by a very straightforward method. The particle-level offset is then defined as the average difference in the reconstructed jet p_{T} between matched jets in simulations with and without pileup overlay. Formally written as

$$\left\langle p_{\mathrm{T,\,offset\,ptcl}} \right\rangle \left(\langle \rho \rangle \,, [\eta] \,, \langle p_{\mathrm{T,\,uncorr}} \rangle \right) = \langle p_{\mathrm{T,\,with\,PU}} - p_{\mathrm{T,\,without\,PU}} \rangle \left[\mu_{\mathrm{PU}}, \eta, p_{\mathrm{T,\,ptcl}} \right]. \quad (4.5)$$

The offset measurement is done as a function of the average number of pileup interactions μ and jet $p_{\mathrm{T,\,ptcl}}$ and η. Jets are considered to be matched, if $\Delta R < R/2$, where R is the jet distance parameter. As consequence of PU, the larger fraction of unmatched jet with low p_{T} is present in the MC sample with PU overlay. Figure 4.6 shows the average offset versus jet p_{T} for various μ. As one can see, the larger PU naturally leads to a larger offset. For a given μ, the offset slightly increases with the jet $p_{\mathrm{T,\,ptcl}}$.

Variables like μ or jet $p_{\mathrm{T,\,ptcl}}$ are available only for the MC simulations and if we want to apply these corrections also to data, the measured offset needs to be parametrized by an appropriate function of offset density ρ and jet η, $p_{\mathrm{T,\,uncorr}}$ and area A. Then the correction formula for the pileup offset $C_{\mathrm{hybrid}}\left(p_{\mathrm{T,\,uncorr}}, \eta, A, \rho\right) = 1 - \left\langle p_{\mathrm{T,\,offset\,ptcl}} \right\rangle / p_{\mathrm{T,\,uncorr}}$ is used as multiplicative factor for the uncorrected jet p_{T}. This approach is known as the hybrid jet area method.

A possible difference in offset between data and MC simulation is corrected by the dedicated offset data/simulation scale factor measured from zero-bias data by a random cone (RC) method. As zero-bias data does not contain or contains only very little contribution from a hard scattering, the energy deposits comes mainly from the detector noise and pileup. The average sum of the p_{T} of the PF candidates in a randomly placed cone in $\eta - \phi$ space is therefore equal to the pileup offset. Figure 4.6 (right) shows the average

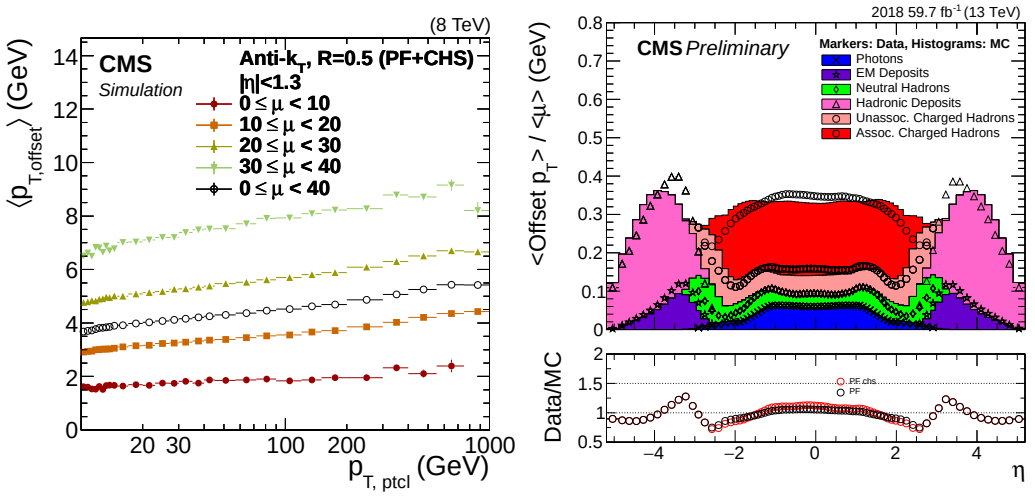

Figure 4.6 – The average particle-level offset as a function of particle-level jet p_{T} for various pileup interaction μ (left) [121]. The average offset per additional pileup interaction μ in data and in simulation as a function of jet η (right) [122].

offset normalized by the average number of pileup interaction μ for each PF candidate type. The ratio of the measured offset for data and simulation represents the scale factor applied in the hybrid method for the data correction. The studies with PF+PUPPI jets have shown that the pileup offset corrections are close to unity for PUPPI jets.

4.5.2 Simulated response corrections

The next step in the jet energy calibration is the simulated response corrections aiming to correct the energy of the reconstructed jets to the true energy of parton initiating the jets. The derivation of this correction relies fully on the MC simulations with a realistic and accurate simulation of the CMS detector. The detector response is modelled by the GEANT4 package. The response is measured from a QCD multijet sample with a matrix element of $2 \to 2$ describing the hard process. While two jets with the highest p_{T} are coming from the ME, the additional softer jets are from the parton shower and the PU. The sample is generated with a flat p_{T} spectrum in order to efficiently cover the full phase space, but then the spectrum is reweighed by $p_{\mathrm{T}}^{-4.5}$ to have a realistic steeply falling shape. The reconstructed jets are matched with the particle-level jets if $\Delta R < R/2$ and then the response is calculated by the following formula:

$$R_{\mathrm{ptcl}}\left(\langle p_{\mathrm{T}} \rangle, \eta\right) = \frac{\langle p_{\mathrm{T}} \rangle}{\langle p_{\mathrm{T,\,ptcl}} \rangle}\left[p_{\mathrm{T,\,ptcl}}, \eta\right]. \tag{4.6}$$

The simulated jet response as a function of the reconstructed jet η for various jet p_{T} is shown in Figure 4.7. Within the barrel region, the response is stable and its size is ~ 0.95. Typically $\sim 15\%$ of the energy of the jet is carried by neutral hadrons which can be detected only in the HCAL and $\sim 60\%$ of their energy is reconstructed. Hence $\sim 5\%$ of the energy of the jet is not measured.

The drop of the response in the barrel region for jets with lower p_{T} is related to the HCAL acceptance. Neutral particles from these jets have such small p_{T} that they are not able to reach the HCAL and to be detected. The response exhibits a strong p_{T} dependence especially in the EC2 and HF regions. A significant drop of the response is observed for

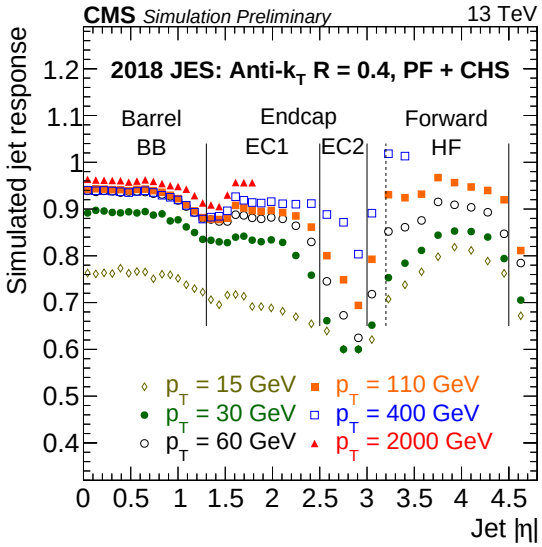

Figure 4.7 – The simulated jet response as a function of the reconstructed jet η for various jet p_{T} [122].

$|\eta| \sim 3.0$ and $|\eta| > 4.5$. The first one is related to the detector transition, while the latter one is caused by the detector acceptance in the forward direction.

The simulated response corrections are applied to the reconstructed jets in data and MC simulations. In case of the data, an additional correction is needed to account for the residual difference in the detector response, while the energy of the jets in MC simulations is now fully corrected.

4.5.3 Residual corrections for data

The main objective of these corrections is to mitigate the remaining residual disagreement between the data and MC simulations. In contrast to the MC-based corrections, the residual corrections for data are time dependent in order to take into account changes in response due to the detector degradation or different detector performance caused by varying conditions during the data-taking periods. The determination of residual corrections utilizes various topologies like dijets, Z/γ+jets and multijet events in order to cover the full phase space from a p_{T} of around 30 GeV to 1 TeV. In all these topologies, the transverse momentum balance between the reconstructed jets and the reference objects is probed and the observed imbalance can come from the jet energy scale different from the unity. The jet response is studied by the p_{T} balance and MPF (missing transverse momentum projection fraction) method. The definition of response for these two methods is given by:

$$R_{\mathrm{jet},\,p_{\mathrm{T}}} = \frac{p_{\mathrm{T,\,jet}}}{p_{\mathrm{T,\,ref}}}, \tag{4.7}$$

$$R_{\mathrm{jet,\,MPF}} = 1 + \frac{\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} \cdot \vec{p}_{\mathrm{T,\,jet}}}{\left(p_{\mathrm{T,\,ref}}\right)^2}. \tag{4.8}$$

While the p_{T} balance method uses information only about the reference object and the

jet, the MPF method is probing the whole hadronic activity in the event recoiling against the reference objects. One of the natural source of imbalance of the studied system is the radiation of an extra jet by initial or final state radiation. Therefore all these measurements are done in bins of additional jet activity α. In case of dijets events, α is defined as a ratio between the p_{T} of the 3rd most energetic jet in the events and the average p_{T} of the dijets system. For Z/γ+jets events, the extra jet activity is quantified as

$$\alpha = p_{\mathrm{T,\,2nd\,jet}}/p_{\mathrm{T,\,Z/\gamma}}. \tag{4.9}$$

The residual corrections consist of the relative η-dependent corrections and the absolute corrections.

Relative η-dependent corrections

Thanks to the high cross section, dijets events are utilized to determine the relative η-dependent correction covering a large p_{T} region and a full pseudorapidity coverage of the CMS detector. By using the tag-and-probe method, where the tag jet has $|\eta| < 1.3$ and the pseudorapidity of the probe jet is unconstrained, the jet response is corrected relatively to the response of jet in the region of $|\eta| < 1.3$. Figure 4.8 shows the relative η-dependent correction measured as a function of jet $|\eta|$. In the central region of the detector, the size of the correction is below 1%, while the largest corrections are measured in the EC2 ($2.0 < |\eta| < 3.0$) and in the HF ($|\eta| > 3.0$). The derivation of this correction will be discussed in more detail in the next chapter.

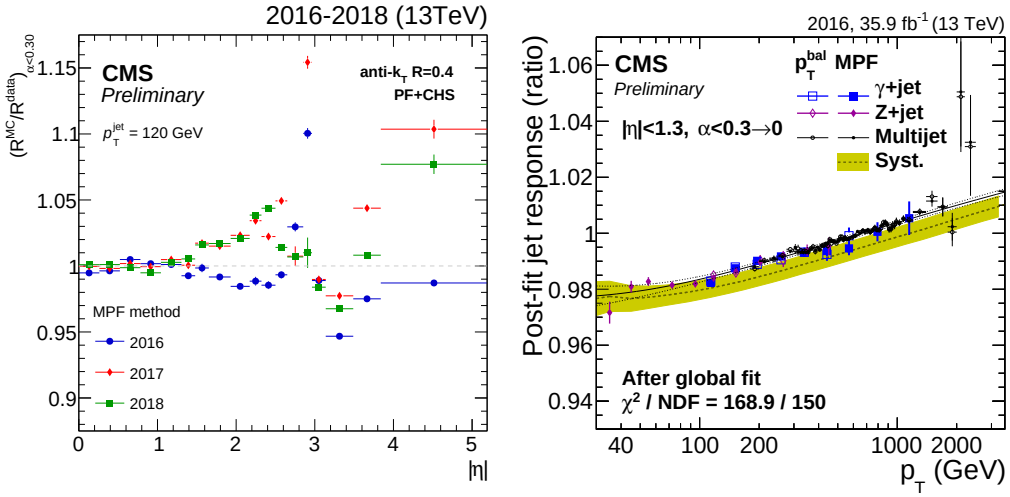

Figure 4.8 – The relative η-dependent corrections measured as a function of jet $|\eta|$ for different data taking period (left) [122]. The combination of various channels to determine the absolute residual correction is shown as a function of jet p_{T} (right) [122].

Absolute corrections

The last remaining step in the derivation of the jet energy correction is to the determination of the absolute scale of corrections. For that, the topologies like Z ($\rightarrow e^+e^-/\mu^+\mu^-$)+jets and γ+jets are used as the jet is balanced by well and precisely measured objects, Z boson or photon. The jet is required to have $|\eta| < 1.3$ and the extra jet activity in events is controlled by cut on α to be smaller than 0.3. The correction is extrapolated to the

$\alpha \to 0$ in order to address only genuine jet energy response effects. By combining different channels in a global fit of the absolute corrections, we explore various regions of p_{T}. While the Z+jets probe a lower region of a p_{T} of around 40 GeV to a 400 GeV, the intermediate region is well covered by γ+jets events. The high p_{T} region can be controlled by multijet events, where the leading jet is balanced with the recoil system consisting of two or more jets. Such leading jet is required to be have $|\eta| < 1.3$. The ratios of response measured in data and MC-simulation by different methods and various topologies are simultaneously fitted in the global fit and the absolute correction is determined as a function of jet p_{T} as shown in left plot in Figure 4.8. In the fit, several nuisance parameters related to the lepton and photon energy scale, ISR+FSR uncertainty and PU effects are considered.

For completeness, let us also mention the dedicated JEC flavour correction aiming to mitigate the difference in response of jets with different flavour. While dijets events are enriched by the gluon jets, the recoiling jets in Z/γ+jets are typically jets initiated by quarks. In consideration of the fact that gluon jets have more constituents and a broader structure than quark jets, a dependence of the response on the jet flavour can be expected. The jet response for different flavour is derived in a similar way as the previously discussed response corrections. In addition, the flavour correction for b jets is validated in Z + b jet events. Such corrections are of particular interest for measurements with a Higgs boson or top quark as their most likely decay channel includes b jets.

4.5.4 Systematic uncertainties

Various steps in the derivation of JEC are sensitive to different effects, which can introduce possible bias on the final calibrations. Precise evaluation of the systematic uncertainties is therefore required. In total, there are upto 26 individual sources which can be divided into 6 main group:

- Absolute scale - given by the global fit and related to uncertainties (lepton energy scale, resolution) of reference objects and correction for ISR+FSR
- Relative scale - estimated by varying the jet energy resolution scale factor in measurement of relative η-dependent corrections
- Pileup - estimated by 5% variation of the data/simulation scale factor measured by RC method and p_{T} dependence of the pileup offset correction
- Method/Sample - given by the difference between p_{T} balance and MPF method and various channels (dijets vs Z/γ+jets vs multijets)
- Jet flavour - based on variance in response of different flavour jet response measured with Pythia and Herwig
- Time stability - given by differences between the luminosity-weighted average of the residual corrections for data and the full dataset corrections.

Figure 4.9 shows the total JEC uncertainty and its six main sources. The left plot shows the JEC uncertainty for the jet in the central region as a function of its p_{T}. The uncertainty ranges from 1% for jet with $p_{\mathrm{T}} \sim 500$ GeV upto 5% for the low p_{T} jets. The JEC uncertainty also depends on the jet η and reaches a minimum in the central region and then gradually increases with larger η.

For many high precision SM measurements, the JEC uncertainty is one of the dominant sources of the experimental uncertainties. As an example we can mention the inclusive jet cross-section measurement. As the jet spectra are steeply falling distributions approximated by p_{T}^{-5}, 1% uncertainty in the jet energy calibration is translated into an uncertainty on the cross section of 5%.

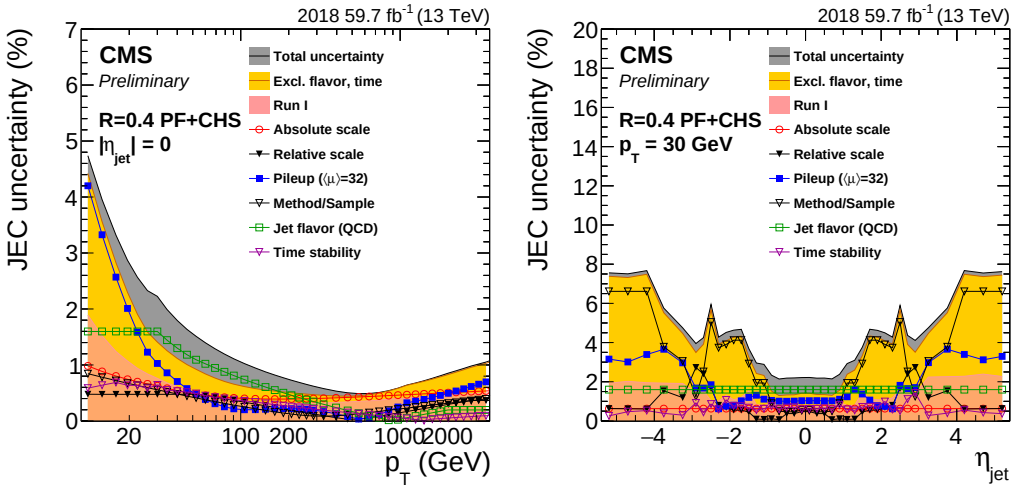

Figure 4.9 – Jet energy correction uncertainties with six main source as a function of jet p_{T} for the jet in the central region (left) and as a function of jet η for the jet with $p_{\mathrm{T}} = 30$ GeV [122].

4.5.5 Jet p_{T} resolution

Once the jets are calibrated, the jet energy resolution (JER) can be measured in MC simulations and data. The observed differences need to be corrected by data/MC scale factors (SF). In contrast to other physics objects like electrons, muons or photons, the JER is rather poor and the determination of JER and JER SF is therefore important to avoid any possible bias arising from the inappropriate smearing of steeply falling jet spectra.

In MC simulation, the JER is defined as Gaussian width of the ratio between the reconstructed and particle-level jet p_{T}. Jets are required to fulfil the $\Delta R < R/2$ matching condition. Then the relative resolution can be parametrized by

$$\frac{\sigma_{p_{\mathrm{T}}}}{p_{\mathrm{T}}} = \sqrt{\frac{\mathrm{sgn}\,(N)\,N^2}{p_{\mathrm{T}}^2} + \frac{S^2}{p_{\mathrm{T}}} + C^2}. \tag{4.10}$$

This formula, sometimes known as NSC function, was already introduced in the context of the energy resolution of the calorimeters.

In data, the measurement of JER extends the technique used for the determination of residual corrections for data. Namely the p_{T} balance method in dijet and Z/γ+jets events is utilized for the JER measurement. For the residual corrections, we are interested in the mean of the distribution describing the jet response and trying to correct for effects which can shift it. Now we study the width of this distribution and effects, which can widen the distribution. The derivation of JER will be briefly described in following two paragraphs. Firstly, we will discuss the measurement in Z/γ+jets topology, later the technique for dijet events is introduced. Different topologies cover various p_{T} range, therefore, the obtained results are complementary to each other. The notation used in next paragraphs is following: the symbols $\oplus$ and $\ominus$ indicated quadratic sum and subtraction, respectively.

Z/γ+jets topology

The jet response for the p_{T} balance method defined by Eq. 4.7 can be expanded as

$$\mathcal{B} = \frac{p_{\mathrm{T,\,jet}}}{p_{\mathrm{T},\,\gamma}} = \frac{p_{\mathrm{T,\,jet}}}{p_{\mathrm{T,\,jet\ ptcl}}} \frac{p_{\mathrm{T,\,jet\ ptcl}}}{p_{\mathrm{T},\,\gamma\ \mathrm{ptcl}}} \frac{p_{\mathrm{T},\,\gamma\ \mathrm{ptcl}}}{p_{\mathrm{T},\,\gamma}}, \tag{4.11}$$

where the $p_{\mathrm{T,\,jet}}$ is the reconstructed and JEC corrected jet transverse momentum with corresponding transverse momentum $p_{\mathrm{T,\,jet\ ptcl}}$ at the particle-level. Similarly, the $p_{\mathrm{T},\,\gamma}$ stands for the photon transverse momentum and the corresponding momentum at the particle-level is $p_{\mathrm{T},\,\gamma\ \mathrm{ptcl}}$. For widths σ of Eq. 4.11 we can write [1]

$$\sigma_{\mathcal{B}} = \sigma\left(\frac{p_{\mathrm{T,\,jet}}}{p_{\mathrm{T,\,jet\ ptcl}}}\right) \oplus \sigma\left(\frac{p_{\mathrm{T,\,jet\ ptcl}}}{p_{\mathrm{T},\,\gamma\ \mathrm{ptcl}}}\right) \oplus \sigma\left(\frac{p_{\mathrm{T},\,\gamma\ \mathrm{ptcl}}}{p_{\mathrm{T},\,\gamma}}\right). \tag{4.12}$$

The first and third term on the right-hand side of Eq. 4.12 are the jet energy resolution and the photon p_{T} resolution. The second term, where the particle-level jet p_{T} and photon p_{T} are compared, is sensitive to the ISR and FSR. However, by extrapolation of the extra jet activity to $\alpha \to 0$, the impact of ISR and FSR is eliminated and the second term just represents the particle-level imbalance (PLI) caused by the UE, out-of-cone showering and the presence of neutrinos. Then Eq. 4.12 is rewritten as

$$\sigma_{\mathcal{B}} k_{\mathrm{rad}} = \sigma_{\mathrm{JER}} \oplus \sigma_{\mathrm{PLI}} \oplus \sigma_{\gamma}, \tag{4.13}$$

where k_{rad} stands for $\sigma_{\mathcal{B}}\left(\alpha \to 0\right)/\sigma_{\mathcal{B}}$. Finally the jet energy resolution reads as

$$\sigma_{\mathrm{JER}} = \sigma_{\mathcal{B}} k_{\mathrm{rad}} \ominus \sigma_{\mathrm{PLI}} \ominus \sigma_{\gamma}. \tag{4.14}$$

In practise, the measurement is performed in bins of the extra jet activity α and the value of σ_{PLI} and σ_{γ} are taken from MC simulations. Then the JER is obtained by the extrapolation of $\alpha \to 0$.

Dijet topology

A similar idea of a factorization approach is also behind the determination of JER in dijet events. In contrast to the previous case, in dijet events there is no reference object with known p_{T} resolution so the JER appears once for each jet. The possible η dependence of JER results in the requirement of having two jets with similar η. Thus both jets have very similar p_{T} and share the same JER. The dijet asymmetry is defined as

$$\mathcal{A} = \frac{p_{\mathrm{T,\,1st\ jet}} - p_{\mathrm{T,\,2nd\ jet}}}{p_{\mathrm{T,\,1st\ jet}} + p_{\mathrm{T,\,2nd\ jet}}}. \tag{4.15}$$

Here $p_{\mathrm{T,\,1st\ jet}}$ and $p_{\mathrm{T,\,2nd\ jet}}$ stand for the transverse momentum of the leading and subleading jets ordered according to the transverse momentum. Then the resolution of the asymmetry distribution reads

$$\sigma_{\mathcal{A}} k_{\mathrm{rad}} = \frac{\sigma_{\mathrm{JER,\,probe}}}{2} \oplus \frac{\sigma_{\mathrm{JER,\,tag}}}{2} \oplus \sigma_{\mathrm{PLI,dijet}}. \tag{4.16}$$

1. The right-hand side of Eq. 4.11 can be rewritten in the following way: firstly we will take logarithms, then apply $\log\left(AB\right) = \log\left(A\right) + \log\left(B\right)$ and $\log\left(1+X\right) \approx X$ for $X \ll 1$. Finally by applying the rules that for a sum of independent and identically distributed random variables, the quadratic summation of widths applies [121].

Thanks to the condition that both jets have similar η, the JER can be obtained as

$$\sigma_{\text{JER}} = \sqrt{2}\left(\sigma_{\mathcal{A}} k_{\text{rad}} \ominus \sigma_{\text{PLI,dijet}}\right). \tag{4.17}$$

On the other hand, the requirement on jets being in the same region is quite strict and reduces very significantly number of events, especially for the case, when we want to measure the JER in the forward direction. Hence, once the JER in the central region is determined, the so-called forward extension can be used to measure the JER in the forward direction with better precision. The forward extension requires one jet to be in the central region and the other jet in the forward direction. For such configuration, the JER of the forward jet is given by

$$\sigma_{\text{JER, forward}} = 2\sigma_{\mathcal{A}} k_{\text{rad}} \ominus 2\sigma_{\text{PLI,central−forward}} \ominus \sigma_{\text{JER, central}}. \tag{4.18}$$

Similarly to the measurement of JER in Z/γ+jets topology, the particle-level imbalance $\sigma_{\text{PLI,dijet}}$ and $\sigma_{\text{PLI,central−forward}}$ are extracted from MC simulation and the JER is determined from the measured width of the dijet asymmetry in bins of α after extrapolation of $\alpha \to 0$

The jet energy resolution data/MC scale factor (JER SF) measured in dijet events as a function of the jet $|\eta|$ for various datasets from LHC Run 2 is shown in Figure 4.10. As one can see, the JER SF is around 1.15 with a 10-20% uncertainty up to the detector region of $|\eta| < 2.5$. The largest value of JER SF is observed for jets with $2.5 < |\eta| < 3.0$. The larger value of JER SF in this region is arising from the detector mis-calibration (cf. Subsection 3.5.3) and inappropriate detector simulation of the transition region.

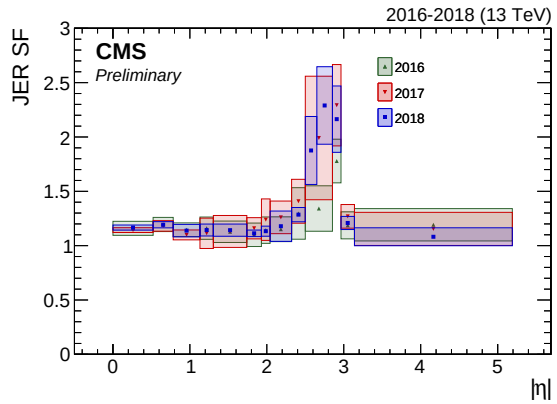

Figure 4.10 – The jet energy resolution data/MC scale factor measured in dijet events as a function of the jet $|\eta|$ for various datasets from Run 2 [122].

4.6 Missing transverse momentum

Once the jets and all visible final-state particles are reconstructed, we can try to estimate the presence of weakly interacting neutral particles by measuring the momentum imbalance in the transverse plane. In case of SM measurements, a momentum imbalance indicates the presence of neutrinos in the events. An example is the leptonic decay of the W-boson or invisible decay of the Z-boson. However, the momentum imbalance can simply also

come from the detector miscalibration etc. Therefore the precise measurement of the momentum imbalance in the transverse plane is crucial.

The measured momentum imbalance in the transverse plane is referred as the missing transverse momentum $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, sometimes called as MET. The shortcut is coming from the alternative name, the missing energy transverse. Its magnitude is then $p_{\mathrm{T}}^{\mathrm{miss}} = \left|\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}\right|$. Formally written the PF MET is defined as

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} = -\sum_{i=1}^{\mathrm{N_{PF\ cand}}} \vec{p}_{\mathrm{T},i}. \tag{4.19}$$

Similarly to the jet reconstruction, we can also define the PF+CHS MET and PF+PUPPI MET [123] according to the employed PU mitigation method. In all cases, the PF candidates can be further categorized in two groups; the clustered PF candidates in jets and the unclustered ones. Thus the sum over PF candidates can be rewritten as

$$\sum_{i=1}^{\mathrm{N_{PF\ cand}}} \vec{p}_{\mathrm{T},i} = \sum_{i=1}^{\mathrm{N_{jets}}} \vec{p}_{\mathrm{T},i} + \sum_{i=1}^{\mathrm{N_{PF\ uncl}}} \vec{p}_{\mathrm{T},i}. \tag{4.20}$$

By the propagation of the jet energy calibration to jets, the corresponding correction factor $\vec{C}^{\,\mathrm{Type-I}}$ for the MET arises as

$$\vec{C}^{\,\mathrm{Type-I}} = \sum_{i=1}^{\mathrm{N_{jets}}} \left(\vec{p}_{\mathrm{T},i}^{\,\mathrm{corr}} - \vec{p}_{\mathrm{T},i}\right), \tag{4.21}$$

where the sum is running over jets with $p_{\mathrm{T}} > 15$ GeV, within the CMS Collaboration very often referred as Type-I correction. Then the "jet-energy-corrected" MET, the most popular and widely used, is given as

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss,\ Type-I}} = \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} - \vec{C}^{\,\mathrm{Type-I}}. \tag{4.22}$$

The propagation of additional corrections like JER SF is possible, but their importance is more analysis-dependent.

5

Determination of relative residual jet energy corrections

Precisely calibrated jets are one of the key ingredients for the most analyses including hadronic final state. In a previous chapter, the factorized approach of the jet energy calibration was discussed and the main idea of each step was explained. Here we will show in more detail the determination of relative residual jet energy corrections. After applying the pileup and simulated response corrections, the jets in the MC simulations are fully corrected. If the MC simulation fully describes the data, no additional corrections for data is needed. However, the observed differences in the jet responses lead to the need of residual jet energy corrections. They are typically time dependent in order to take into account changes in response caused by the detector degradation and different detector performance over the data-taking period. As already discussed, various topologies like dijet, Z/γ+jets, multijet events are utilized for the data-driven estimation of the residual jet energy corrections. The dijet topology is used for the measurement of η dependent relative residual jet energy corrections. The work presented in this chapter was done by the author of this thesis.

5.1 Methodology and jet response definition

The ideal dijet topology consists of two jets with balanced transverse momenta thanks to the momentum conservation principle. The azimuthal separation between jets is given by

$$\Delta\phi_{\text{dijet}} = \min\left(|\phi_1 - \phi_2|\,, 2\pi - |\phi_1 - \phi_2|\right). \tag{5.1}$$

Here ϕ_1 and ϕ_2 are the azimuthal angles of the first and second jet, respectively. At the LHC energies, the ideal dijet topology is modified due to ISR and FSR. The high energetic incoming or outgoing partons typically radiate partons which carry part of energy and introduce an imbalance between the two highest jets. The extra jet is typically softer. An illustration of the dijet topology with one additional soft jets is shown in Figure 5.1.

The determination of the relative residual jet energy corrections in a dijet topology is based on the tag-and-probe method. The leading and subleading jets ordered according the jet transverse momenta are selected as tag and probe. The tag jet is required to be in the barrel region, i.e. $|\eta| < 1.3$, while the pseudorapidity of the probe jet is unconstrained. If both jets are in the barrel region, then the choice of the tag and probe jets is done

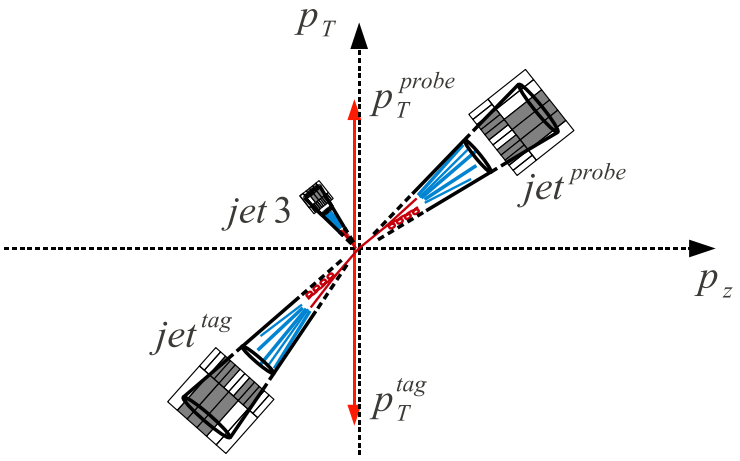

Figure 5.1 – Sketch of the dijet topology in back-to-back configuration with one extra soft jet [122].

randomly. The relative residual jet response is measured by the probe jets with respect to the tag jets lying in the central region.

The average transverse momentum of the dijet system, i.e. tag and probe jets, is

$$\overline{p}_{\mathrm{T}} = p_{\mathrm{T}}^{\mathrm{ave}} = \frac{p_{\mathrm{T}}^{\mathrm{probe}} + p_{\mathrm{T}}^{\mathrm{tag}}}{2}. \tag{5.2}$$

The extra jet activity is quantified as

$$\alpha = \frac{p_{\mathrm{T}}^{\mathrm{3rd\ jet}}}{\overline{p}_{\mathrm{T}}} \tag{5.3}$$

A smaller value of α indicates a configuration more similar to the ideal dijet topology, while α closer to unity corresponds more to the trijet topology as the transverse momenta of all three jets are similar. The jet response is explored by the p_{T}-balance and MPF method. The dedicated correction is then derived from the ratio of data and MC simulation.

5.1.1 Dijet p_{T}-balance method

This method is based on the momentum conservation. The p_{T}-balance between the tag and probe jets is quantified by the asymmetry $\mathcal{A}$ defined as

$$\mathcal{A} = \frac{p_{\mathrm{T}}^{\mathrm{probe}} - p_{\mathrm{T}}^{\mathrm{tag}}}{p_{\mathrm{T}}^{\mathrm{probe}} + p_{\mathrm{T}}^{\mathrm{tag}}} = \frac{p_{\mathrm{T}}^{\mathrm{probe}} - p_{\mathrm{T}}^{\mathrm{tag}}}{2\overline{p}_{\mathrm{T}}}. \tag{5.4}$$

The asymmetry distribution $\mathcal{A}$ is studied in the bins of $\overline{p}_{\mathrm{T}}$, η^{probe}, and the extra jet activity α. The relative response $R_{\mathrm{rel}}^{p_{\mathrm{T}}}$ with respect to the tag jet is then given as

$$R_{\mathrm{rel}}^{p_{\mathrm{T}}}\left(\overline{p}_{\mathrm{T}}, \eta^{\mathrm{probe}}, \alpha\right) = \frac{1 + \langle\mathcal{A}\rangle}{1 - \langle\mathcal{A}\rangle}, \tag{5.5}$$

where the symbol $\langle \mathcal{A} \rangle$ denotes the average value of the distribution $\mathcal{A}$. For the fine binning in $\overline{p}_{\mathrm{T}}$, this equation can be rewritten as

$$R_{\mathrm{rel}}^{p_{\mathrm{T}}}\left(\overline{p}_{\mathrm{T}}, \eta^{\mathrm{probe}}, \alpha\right) = \frac{\left\langle p_{\mathrm{T}}^{\mathrm{probe}} \right\rangle}{\left\langle p_{\mathrm{T}}^{\mathrm{tag}} \right\rangle}. \tag{5.6}$$

Here one can see similarity with the absolute residual jet energy correction studied in Z/γ+jets topologies, where the jet response is measured with the respect to the transverse momentum of the reference objects.

5.1.2 Missing transverse energy projection fraction method

The alternative technique which allows to determine the jet response is the MPF method. This method utilizes the fact that the topologies used for the determination of the residual jet energy corrections do not contain the genuine missing transverse momentum. The dijet system is balanced in the transverse plane. Formally written

$$\vec{p}_{\mathrm{T,\,true}}^{\,\mathrm{tag}} + \vec{p}_{\mathrm{T,\,true}}^{\,\mathrm{recoil}} = 0, \tag{5.7}$$

where $\vec{p}_{\mathrm{T,\,true}}^{\,\mathrm{recoil}}$ stands for the vectorial sum of the transverse momenta of the probe jet and additional softer jets coming mainly from ISR and FSR and other event activities like the UE activity, OOC showering and so on. The presence of $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ in the dijet topology can be explained only by jet energy miscalibration. Thus, Eq. 5.7 can be rewritten as

$$R_{\mathrm{tag}}\,\vec{p}_{\mathrm{T,\,true}}^{\,\mathrm{tag}} + R_{\mathrm{recoil}}\,\vec{p}_{\mathrm{T,\,true}}^{\,\mathrm{recoil}} = -\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}, \tag{5.8}$$

where R_{tag} and R_{recoil} stand for the relative response for the tag jet and the recoil system. TEquation 5.8 and 5.7 can be solved to express R_{recoil} as

$$R_{\mathrm{recoil}} = R_{\mathrm{tag}} + \frac{\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} \cdot \vec{p}_{\mathrm{T,\,true}}^{\,\mathrm{tag}}}{\left(\vec{p}_{\mathrm{T,\,true}}^{\,\mathrm{tag}}\right)^2}. \tag{5.9}$$

In order to extract the probe jet response from the MPF method, additional steps are required. First, as we are interested in the measurement of the relative correction, the response R_{tag} for the tag jet is set as unity. Thus, $\vec{p}_{\mathrm{T,\,true}}^{\,\mathrm{tag}} = \vec{p}_{\mathrm{T}}^{\,\mathrm{tag}}$. To set R_{recoil} as R_{MPF}, the recoil system needs to be identified as the probe jet. That means, that additional event activity is required to be reduced as much as possible. This is achieved by applying cuts on the extra jet activity α and the jet response for MPF method reads as

$$R_{\mathrm{probe}} \approx R_{\mathrm{MPF}} = 1 + \frac{\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} \cdot \vec{p}_{\mathrm{T}}^{\,\mathrm{tag}}}{\left(\vec{p}_{\mathrm{T}}^{\,\mathrm{tag}}\right)^2}. \tag{5.10}$$

In case of dijet events, the asymmetry distribution $\mathcal{B}$ is then defined as

$$\mathcal{B} = \frac{\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} \cdot \left(\vec{p}_{\mathrm{T}}^{\,\mathrm{tag}} / p_{\mathrm{T}}^{\,\mathrm{tag}}\right)}{2\overline{p}_{\mathrm{T}}}. \tag{5.11}$$

Identically to the p_{T}-balance method, the relative response measured by the MPF method in the bins of $\overline{p}_{\mathrm{T}}$, η^{probe} and α is expressed as

$$R_{\mathrm{rel}}^{\mathrm{MPF}}\left(\overline{p}_{\mathrm{T}}, \eta^{\mathrm{probe}}, \alpha\right) = \frac{1 + \langle \mathcal{B} \rangle}{1 - \langle \mathcal{B} \rangle}. \tag{5.12}$$

The p_{T}-balance and MPF method are two complementary techniques which can be utilized to measure the relative jet response. The former one is well known and an established technique since the SPS era [124], however it has a certain disadvantage. As the jet p_{T} spectrum is steeply falling, the measurement is sensitive to the jet energy resolution and a possible bias related to the jet energy resolution can arises. Despite of the choice of the $\overline{p}_{\mathrm{T}}$ as binning variable in order to minimize the bias, the resolution bias is not completely eliminated. The accurate determination of the relative residual correction requires to have the same jet energy resolution in data and MC simulation. On the other hand, utilizing $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ in the MPF method in a way as is done, the sensitivity of this method to ISR and FSR is reduced. In addition, the resolution bias is believed to have a reduced impact.

5.2 Dijet event selection and reconstructions

5.2.1 Data and MC simulations

The determination of the relative residual jet energy corrections is performed by using proton-proton collisions at a centre-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ recorded by the CMS experiment in 2018, corresponding to the integrated luminosity of $59.8\,\mathrm{fb}^{-1}$. The recorded data sample is split in four run periods. Within one run period the data-taking conditions are assumed to be stable. The first three run periods of 2018 data have been reprocessed and more precise detector calibrations were used while the last run period, namely Run2018D, uses the so-called *prompt reconstruction*, which is usually less accurate. The data samples with the information about the run numbers and the integrated luminosity are listed in Table 5.1.

Sample	Run Period	Reconstruction version	Run ranges	Int. lumi. / fb^{-1}
JetHT	Run2018A	17Sep2018-v1/MINIAOD	315252 – 316995	14.02
JetHT	Run2018B	17Sep2018-v1/MINIAOD	317080 – 319310	7.06
JetHT	Run2018C	17Sep2018-v1/MINIAOD	319337 – 320065	6.89
JetHT	Run2018D	PromptReco-v*/MINIAOD	320673 – 325175	31.74
				total: 59.7

Table 5.1 – Sample, Run period, reconstruction version, run ranges together with corresponding integrated luminosity of data set used for the determination of the relative residual jet energy corrections.

The MC samples of QCD events used in this study are simulated by Pythia8 using CP5 tune [125] to model the UE activity. The MC simulation is produced in bins of the transferred momentum of the primary interaction $\hat{p}_{\mathrm{T}}$ in order to achieve good statistical precision over the jet p_{T} spectrum. The MEs include only $2 \to 2$ processes, therefore the additional jets are coming from ISR and FSR. The used MC samples with the corresponding cross section are listed in Table 5.2.

Dataset/channel/phase space		MC generator	Cross-section/pb
QCD	$30 < \hat{p}_{\mathrm{T}} \leq 50$	PYTHIA8 + CP5 TUNE	$1.069 \cdot 10^{8}$
	$50 < \hat{p}_{\mathrm{T}} \leq 80$		$1.576 \cdot 10^{7}$
	$80 < \hat{p}_{\mathrm{T}} \leq 120$		$2.341 \cdot 10^{6}$
	$120 < \hat{p}_{\mathrm{T}} \leq 170$		$4.071 \cdot 10^{5}$
	$170 < \hat{p}_{\mathrm{T}} \leq 300$		$1.033 \cdot 10^{5}$
	$300 < \hat{p}_{\mathrm{T}} \leq 470$		$6.826 \cdot 10^{3}$
	$470 < \hat{p}_{\mathrm{T}} \leq 600$		$5.526 \cdot 10^{2}$
	$600 < \hat{p}_{\mathrm{T}} \leq 800$		$1.566 \cdot 10^{2}$
	$800 < \hat{p}_{\mathrm{T}} \leq 1000$		$2.632 \cdot 10^{1}$
	$1000 < \hat{p}_{\mathrm{T}} \leq 1400$		7.500
	$1400 < \hat{p}_{\mathrm{T}} \leq 1800$		$6.479 \cdot 10^{-1}$
	$1800 < \hat{p}_{\mathrm{T}} \leq 2400$		$8.715 \cdot 10^{-2}$
	$2400 < \hat{p}_{\mathrm{T}} \leq 3200$		$5.242 \cdot 10^{-3}$
	$3200 < \hat{p}_{\mathrm{T}}$		$1.349 \cdot 10^{-4}$

Table 5.2 – MC simulatations with cross section used for the determination of the relative residual jet energy corrections. The transferred momentum of the primary interaction $\hat{p}_{\mathrm{T}}$ is quoted in GeV.

5.2.2 Trigger selection

The events are collected online by dedicated HLT dijet triggers with different requirements on the minimal $\overline{p}_{\mathrm{T}}$. The measurement of the relative residual jet energy corrections employed two sets of triggers. The first group, the so-called central triggers, requires the dijet events without any constrain on the pseudarapidity of the jets. As the jet spectrum steeply falls not only with the increasing p_{T} but also with the increasing $|\eta|$, dedicated triggers are required for the forward region in order to collect a sufficient amount of events with the jets in the forward direction. The forward dijet triggers require one jet to be located in the forward region, while the other jet is in the central direction.

As the trigger decision needs to be done relatively fast, the online reconstruction is usually less precise. Thus, the jet p_{T} is typically underestimated. The physics analyses are performed with the objects reconstructed offline. In order to properly select events, the measurement of the trigger efficiency as the function of offline $\overline{p}_{\mathrm{T}}$ is required. The trigger is considered as fully efficiency if it reaches 99% efficiency. This measurement is performed by the tag-and-probe method and it will be discussed in more detail later. The measured offline thresholds for the central and forward dijet triggers are summarized in the Table 7.5.

5.2.3 Dijet event selection

The determination of the residual JEC is done by default for the AK4 PF+CHS jets. In our case, the jets have to be corrected by the pileup offset and MC simulated response corrections. Then, the dijet events are selected in the following steps:

- Candidate dijet events are required to be triggered by one of the dijet triggers and the offline $\overline{p}_{\mathrm{T}}$ needs to be above the offline threshold of the given triggers.
- Reconstructed jets need to pass certain quality criteria, known as Jet ID.
- To ensure the proper dijet topology, two leading jets are required to be in a back-to-back configuration, i.e. $\Delta\phi > 2.7$. In addition, for the determination of the relative

Central trigger	Offline threshold / GeV
HLT_DiPFJetAve40	67
HLT_DiPFJetAve60	93
HLT_DiPFJetAve80	118
HLT_DiPFJetAve140	189
HLT_DiPFJetAve200	257
HLT_DiPFJetAve260	325
HLT_DiPFJetAve320	391
HLT_DiPFJetAve400	478
HLT_DiPFJetAve500	585

Forward trigger	Offline threshold / GeV
HLT_DiPFJetAve60_HFJEC	93
HLT_DiPFJetAve80_HFJEC	118
HLT_DiPFJetAve100_HFJEC	143
HLT_DiPFJetAve160_HFJEC	210
HLT_DiPFJetAve220_HFJEC	279
HLT_DiPFJetAve300_HFJEC	379

Table 5.3 – Offline thresholds of the central and forward dijet triggers.

residual correction one jet must be reconstructed in the barrel region, $|\eta| < 1.3$.

- Events with p_{T}-balance asymmetry $|\mathcal{A}| > 0.7$ are rejected as a large asymmetry can indicate not properly reconstructed jets.

In order to remove suspicious events, an additional event rejection is done by applying the MET filters, *hot-zone* maps etc. Simulated QCD sample require dedicated selection criteria to remove possible bias arising from mismodelling of the PU contribution. Namely, the events containing the reconstructed leading jet with the p_{T} larger than 1.5 of the p_{T} of the leading particle-level jet are removed. In this case, the large fraction of the PU contribution is clustered into the jet. Furthermore, the p_{T} of the leading particle-level jets or the reconstructed jet is required to be smaller than 1.5 of the transferred momentum of the primary interaction, $\hat{p}_{\mathrm{T}}$.

The analysis is performed in the bins of $\overline{p}_{\mathrm{T}}$, $\left|\eta^{\mathrm{probe}}\right|$, and α. The bin edges for binning in $\overline{p}_{\mathrm{T}}$ is given by the offline trigger threshold in such way that only one trigger contributes to one specific bin. The used η binning, in total 18 bins, is listed in Table 5.4. The bin edges correspond to the physical boundaries of the various HCAL and ECAL sectors of the CMS detector.

bin Nº	$\lvert\eta\rvert$ range	bin Nº	$\lvert\eta\rvert$ range
1	0.000 - 0.261	10	2.172 - 2.322
2	0.261 - 0.522	11	2.322 - 2.500
3	0.522 - 0.783	12	2.500 - 2.650
4	0.783 - 1.044	13	2.650 - 2.853
5	1.044 - 1.305	14	2.853 - 2.964
6	1.305 - 1.479	15	2.964 - 3.139
7	1.479 - 1.653	16	3.139 - 3.489
8	1.653 - 1.930	17	3.489 - 3.839
9	1.930 - 2.172	18	3.839 - 5.191

Table 5.4 – Bins of $|\eta|$ used in the measurement.

The binning in α is following: 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, and 0.45 . Finally, event reweighting of the MC simulation is done in such way that the simulated PU profile and $\overline{p}_{\mathrm{T}}$ distribution match the observed distributions in data.

The reconstructed jets in data have worse jet energy resolution than in simulation and as discussed earlier the methods for the determination of the relative residual jet energy

correction are sensitive to this difference. Hence, jets in MC simulations need to be smeared. This is done by the hybrid smearing method. If a matched particle-level jet is found, the scaling method is preferred, otherwise the stochastic smearing is applied. The scaling method is based on particle-level jet matching with the reconstructed jet. The requirements for particle jet - reconstructed jet matching are following:

$$\Delta R < R_{\text{cone}}/2,\ \left|p_{\text{T}} - p_{\text{T}}^{\text{ptcl}}\right| < 3\sigma_{\text{JER}} p_{\text{T}}, \tag{5.13}$$

where R_{cone} stands for the jet cone size and σ_{JER} is the relative p_{T} resolution measured in simulation. If those criteria are satisfied, then the reconstructed jet is smeared with a factor

$$c_{\text{JER}} = 1 + (s_{\text{JER}} - 1)\,\frac{p_{\text{T}} - p_{\text{T}}^{\text{ptcl}}}{p_{\text{T}}}, \tag{5.14}$$

where s_{JER} is the jet energy resolution scale factor. If a particle-level jet is not found, the stochastic smearing has to be used. In such case, the jet four-momentum is rescaled with a factor

$$c_{\text{JER}} = 1 + \mathcal{N}\,(0,\sigma_{\text{JER}})\,\sqrt{\max\left(s_{\text{JER}}^2 - 1,0\right)}, \tag{5.15}$$

where $\mathcal{N}\,(0,\sigma_{\text{JER}})$ stands for a random number from a normal distribution with a mean equal to zero and variance σ^2. In both cases (scaling method and stochastic smearing), the scaling factor c_{JER} is truncated at zero, i.e. if it is negative, it is set to zero.

The effect of the JER smearing needs to be also propagated into the $\vec{p}_T^{\,\text{miss}}$. In this analysis, the PF+CHS $\vec{p}_T^{\,\text{miss}}$ with the modified Type-1 correction is employed.

5.3 Residual correction determination

The determination of the relative residual correction relies on the measurement of the relative response $R_{\text{rel}}^{p_{\text{T}}}$ and $R_{\text{rel}}^{\text{MPF}}$ via asymmetry distributions of $\mathcal{A}$ and $\mathcal{B}$, respectively, in data and MC simulations. The residual correction, that mitigates the difference in jet response in data and MC simulations, is given as a response ratio. Using the ideal dijet topology, i.e. without any additional jets, would be statistically limited, therefore the response ratio is evaluated for a given α and then an additional correction is applied for the interpolation to the ideal dijet configuration. The nominal working point for the Run 2 is $\alpha < 0.3$. Formally written, the relative residual correction is given as

$$C\left(\eta^{\text{probe}}\right) = \left\langle \frac{R^{\text{MC}}}{R^{\text{Data}}} \right\rangle_{\alpha<0.3} \cdot k_{\text{FSR}}, \tag{5.16}$$

where R stands for $R_{\text{rel}}^{p_{\text{T}}}$ or $R_{\text{rel}}^{\text{MPF}}$. Here k_{FSR} is the radiation correction defined as

$$k_{\text{FSR}} = \lim_{\alpha' \to 0} \frac{\left\langle \frac{R^{\text{MC}}}{R^{\text{Data}}} \right\rangle_{\alpha<\alpha'}}{\left\langle \frac{R^{\text{MC}}}{R^{\text{Data}}} \right\rangle_{\alpha<0.3}}. \tag{5.17}$$

It removes the remaining bias related to the extra jet radiation and transforms the relative correction $\left\langle \frac{R^{\mathrm{MC}}}{R^{\mathrm{Data}}} \right\rangle_{\alpha<0.3}$ from the working point $\alpha < 0.3$ to a ideal dijet configuration without any additional jet ($\alpha = 0$). This is the general strategy of the derivation of the residual correction. In the following two subsections, the step-by-step derivation with the exemplary distribution will be shown.

5.3.1 Relative response and interpolation

The asymmetry distribution of $\mathcal{A}$ and $\mathcal{B}$ is measured triple-differentially in bins of $\overline{p}_{\mathrm{T}}, \eta^{\mathrm{probe}}$ and α. Examples of the asymmetry distributions measured by the p_{T}-balance and MPF method are shown in Figure 5.2.

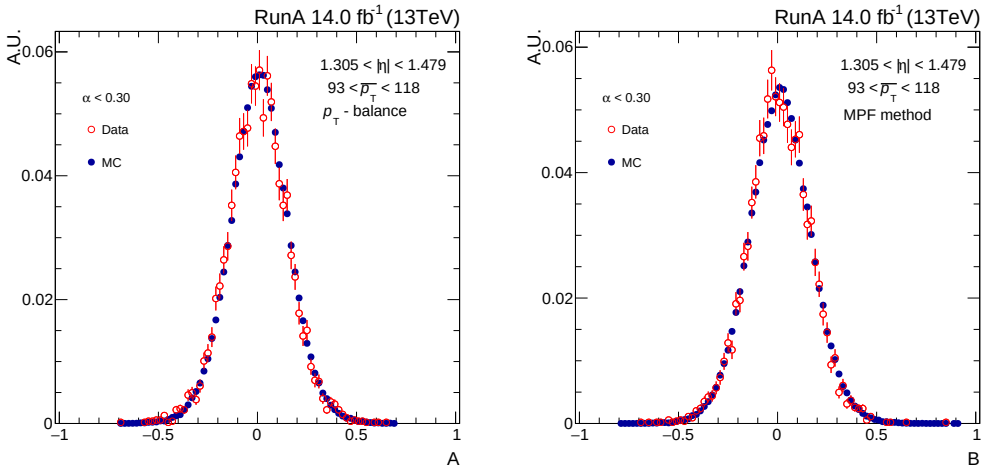

Figure 5.2 – The asymmetry distribution for p_{T}-balance and MPF method for jets with p_{T} in range from 93 to 118 GeV in 6th η-bin.

The arithmetic mean of the distribution is then calculated. Apart from the arithmetic mean, the mean of the Gaussian fits can be used, but the difference is negligible. The former method is preferable because of practical purpose. In Figure 5.3, the arithmetic means of the p_{T}-balance and MPF method as a function of $\overline{p}_{\mathrm{T}}$ are shown. For illustration purpose, the distributions from two detector regions, namely the barrel and endcap, are given. The jet responses measured in data and MC simulations agree in the central region. It implies relatively small residual corrections. The larger disagreement is typically observed in the endcap region, namely for $2.5 < |\eta| < 3.2$. There are various reasons for that. First, this region is out of the tracker coverage, therefore the CHS PU subtraction method is not applicable here. In addition, the degradation of the crystal in the endcap ECAL detector (cf. Subsection 3.5.3) together with the modelling of the transition region makes this region quite challenging for the jet energy calibration.

The relative residual corrections are assumed to be η-dependent, however a certain p_{T} dependence is observed. This is especially true for the measurement in Run 2. In order to estimate the possible bias arising from the p_{T} dependence, the response ratio is fitted with a constant function or a function which depends on $\overline{p}_{\mathrm{T}}$ as

$$f\left(\overline{p}_{\mathrm{T}}\right) = a + b \log\left(\overline{p}_{\mathrm{T}}\right). \tag{5.18}$$

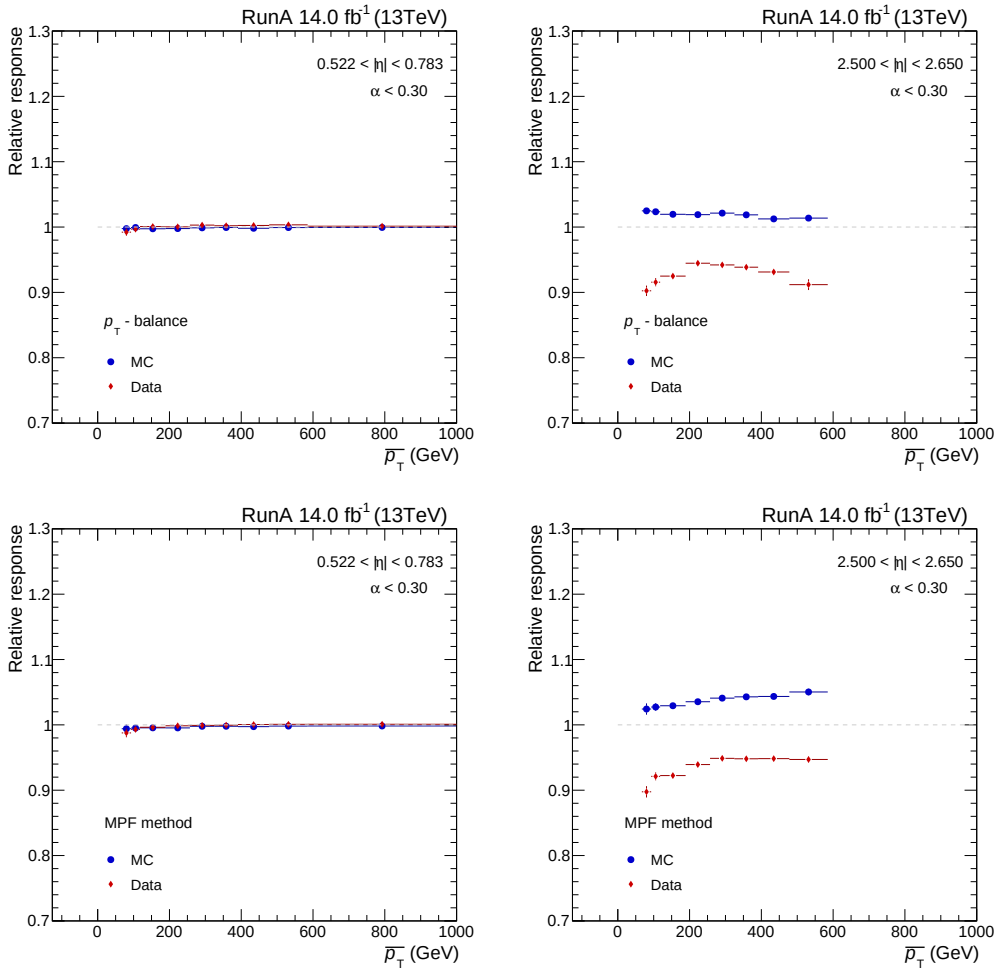

Figure 5.3 – The arithmetic means of the p_{T}-balance (upper) and MPF method (bottom) as a function of $\overline{p}_{\mathrm{T}}$ in the barrel (left) and endcap (right) regions.

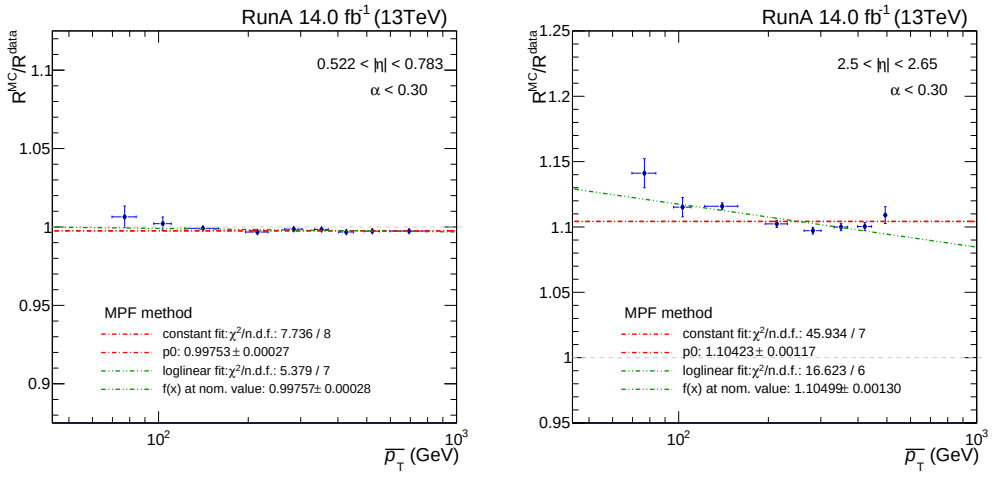

Figure 5.4 – The ratio of the relative responses measured in MC simulation and data as a function of $\overline{p}_{\mathrm{T}}$ in the barrel (left) and endcap (right) regions. Fits to the ratio is represented by red (constant function) and green (log-linear function) lines.

Here a and b are the free parameters of the fit. Figure 5.4 shows the ratio of the jet responses in data and MC simulations. While the ratio in the central region is rather flat with negligible $\overline{p}_{\mathrm{T}}$ dependence, the forward region features a stronger $\overline{p}_{\mathrm{T}}$ dependence. For illustration, the results of the constant and log-linear function are displayed. The difference between fitting these functions will be taken as a systematic uncertainty.

5.3.2 Radiation correction of the methods

The last step in the derivation of the relative residual correction is the determination of the radiation corrections of the methods. The main objective of them is to remove possible bias related to soft radiation of jets by ISR and FSR, that can introduce some imbalance of the dijet system. By applying radiation corrections, the so called k_{FSR} factor, the relative correction $\left\langle R^{\mathrm{MC}}/R^{\mathrm{Data}}\right\rangle_{\alpha<0.3}$ is projected from the working point $\alpha < 0.3$ to the ideal dijet configuration without any additional jet ($\alpha = 0$).

The radiation corrections are extracted from the measurement of the double response ratio, i.e. $\left\langle R^{\mathrm{MC}}/R^{\mathrm{Data}}\right\rangle_{\alpha<\alpha'} / \left\langle R^{\mathrm{MC}}/R^{\mathrm{Data}}\right\rangle_{\alpha<0.3}$ for various α'. By construction, all detector effects cancel and the double ratio manifests mainly the sensitivity of each method to ISR and FSR. Examples of the double ratios as a function of α for various $\overline{p}_{\mathrm{T}}$ bins are shown in Figure 5.5.

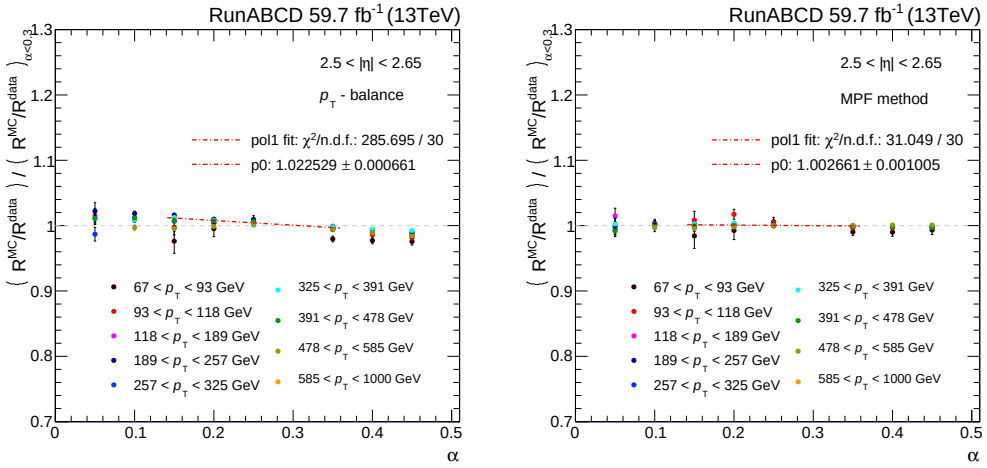

Figure 5.5 – Double ratio of the relative response of the p_{T}-balance and MPF method as a function of extra jet activity α for different p_{T}-bins.

The k_{FSR} factor is obtained by the linear extrapolation of $\alpha \to 0$. For the extrapolation only events with α in range of 0.15 and 0.35 are used, while all $\overline{p}_{\mathrm{T}}$ bins are considered. The obtained values as a function of $\left|\eta^{\mathrm{probe}}\right|$ for both methods are shown in Figure 5.6. The k_{FSR} factor for the MPF method is found to be close to unity. It confirms that this method is less sensitive to ISR and FSR. On the other hand, the k_{FSR} factor for the p_{T}-balance varies with increasing η^{probe}. In the central region, it is close to unity and then gradually increases with larger η^{probe}. For both methods, the k_{FSR} factor is parametrized by

$$f\left(\left|\eta^{\mathrm{probe}}\right|\right) = a + \frac{b \cdot \cosh\left(\left|\eta^{\mathrm{probe}}\right|\right)}{1 + c \cdot \cosh\left(\left|\eta^{\mathrm{probe}}\right|\right)}, \tag{5.19}$$

where a, b, and c are free parameters of the fit.

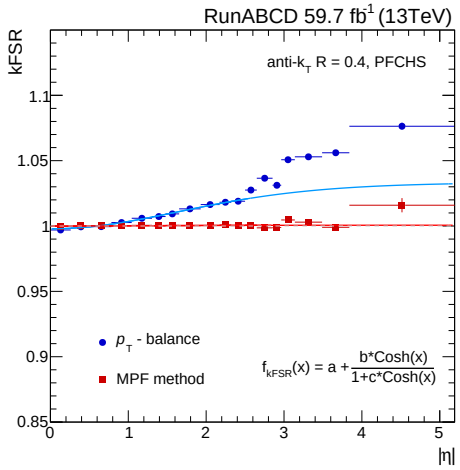

Figure 5.6 – The radiation correction factor k_{FSR} for p_{T}-balance and MPF method as a function of jet $|\eta|$. Fits to the correction factor are shown by blue and red lines.

5.4 Results

The results on the determination of the relative residual jet energy corrections using proton-proton collisions at a centre-of-mass energy of 13 TeV recorded by the CMS experiment in 2018, corresponding to the integrated luminosity of 59.8 fb^{-1} are shown in Figure 5.7. The relative corrections, measured by the p_{T}-balance and MPF method, are shown as a function of $\left|\eta^{\mathrm{probe}}\right|$. In addition, for a comparison, the results obtained by the constant and log-linear functions are provided. Within the barrel region, the relative corrections are of order 1-2%. One can see, that the corrections exhibit a different behaviour for the run period RunABC and RunD. In the endcap region, the corrections for RunABCD are above unity. The η bin where the maximum of the corrections ($\sim 20\%$) is hit, corresponds to the transition endcap-forward region. Opposite to these run periods, the corrections for RunD reach the minimum in this region. The contrasting behaviour in this region arises from the different version of the data reconstruction. Finally, the example of the p_{T} dependence of the determined residual corrections is shown in Figure 5.8. The corrections exhibit the large p_{T} dependence in the endcap region, while it is negligible in the barrel region. The results obtained by the MPF method and log-linear fits are used as a central results.

5.5 Contribution to the jet energy scale uncertainty

The derivation of the relative residual correction is sensitive to various effects, that can introduce a possible bias on the final results. The most important ones that are also included in the total jet energy scale uncertainty are described in the next subsections.

5.5.1 Method of derivation

Two methods have been used to determine the relative residual jet energy corrections. Both methods should lead to the same corrections, therefore the difference is counted into the systematic uncertainties. Moreover, in order to incorporate a possible bias related to the observed p_{T} dependence, the difference between the results obtained by the constant and log-linear fits is assigned as the uncertainty.

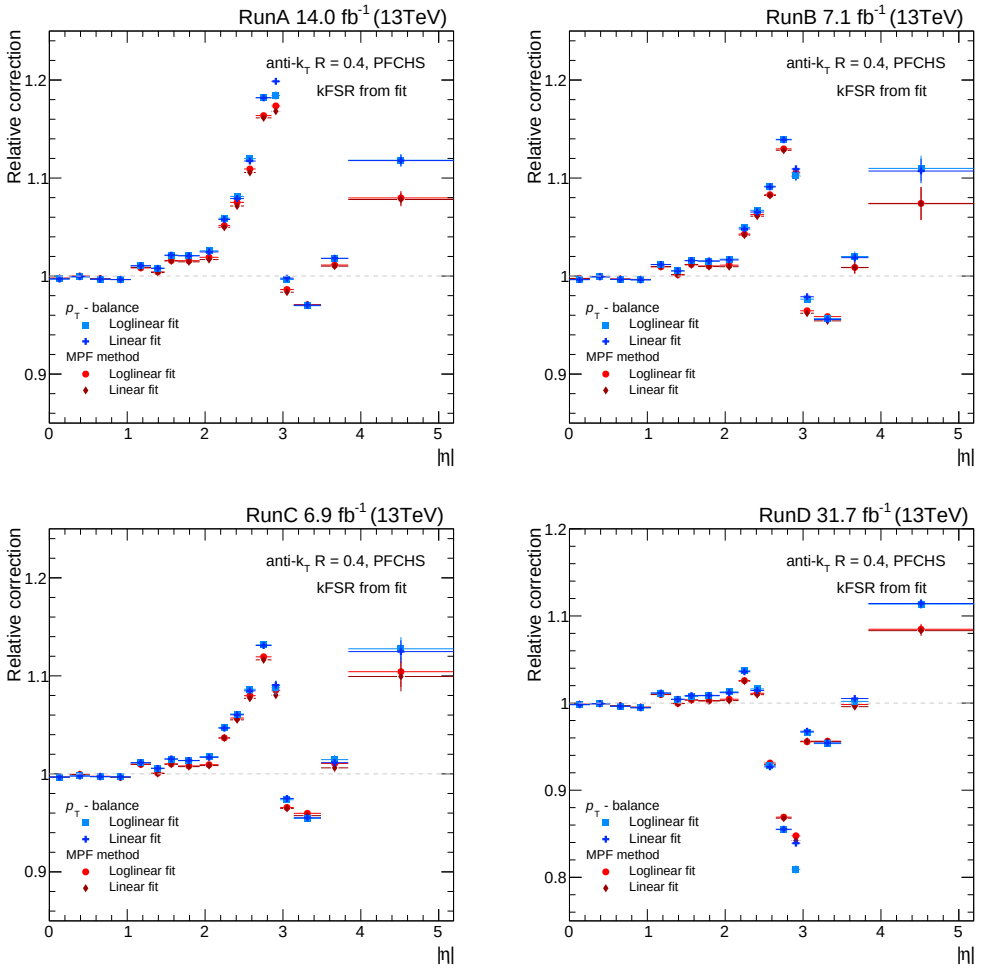

Figure 5.7 – The relative residual jet energy corrections as a fuction of jet $|\eta|$ for various data-taking period. The red and blue markers indicate the results obtained by the p_{T}-balance and MPF method, respectively.

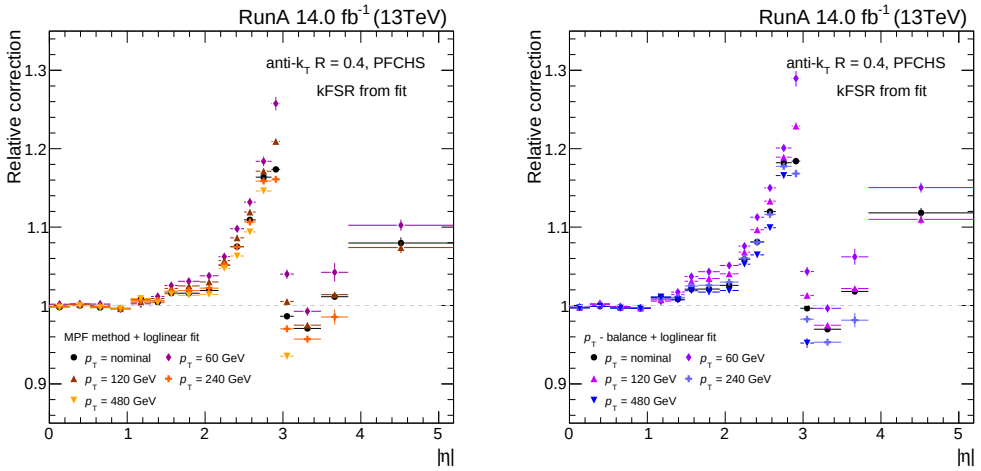

Figure 5.8 – The p_{T} dependence of the relative residual jet energy corrections as a fuction of jet $|\eta|$ for the MPF method (left) p_{T}-balance (right).

5.5.2 Time stability

To take into account the changes of the jet response due to the detector degradation or different detector performance caused by varying conditions during the data-taking, the residual corrections are derived as run period dependent. Then the uncertainty related to the time stability of the corrections is assigned as the difference between the luminosity-weighted average of the residual corrections for data and the full dataset corrections.

5.5.3 Jet energy resolution scale factor uncertainty

As the jet energy resolution differs between the data and MC simulations, the jets momenta in the MC simulations need to be smeared. The JER SF arrives with a certain uncertainty, which must be propagated into the measurement. The JER in the MC simulation is varied within ± 1 standard deviation and its impact is evaluated by comparing the derived relative corrections. The largest impact of the JER SF is typically observed in the endcap region.

5.5.4 Modelling of the PS

The impact of the modelling of the PS can be evaluated by comparing the relative corrections using MC samples with the various PS, for example Pythia8 and Herwig++/Herwig7. Because of the missing MC simulations based on Herwig++/Herwig7, this uncertainty is not updated.

5.6 Summary

In this chapter, the derivation of the relative residual corrections for data recorded by the CMS experiment in 2018 was presented. The relative residual corrections are part of the jet energy corrections and aim to mitigate the residual difference of the jet response in the data and the MC simulation. The obtained results are included in the official JEC for 2018 data used by CMS collaboration in all analyses using the jets.

As an outlook of this study, let us mention the expected improvement and the reduction of the systematic uncertainties by the so-called Legacy reprocessing of Run 2 data set during the long shutdown between LHC Run 2 and Run 3. More accurate detector calibration and simulation lead to more precise jet reconstruction. The final Run 2 Legacy jet calibration is still advancing, but the first results exhibit a better time stability and a reduce p_{T} dependence. The comparison of the relative residual jet energy correction for the current data sample and Legacy data sample is shown in Figure 5.9.

In addition, the Run 2 Legacy 2018 data sets were utilized to study the relative corrections for various jet cone sizes and the pileup subtraction methods. Namely, the residual corrections were derived for AK4 PF+CHS, AK8 PF+CHS, AK4 PF+PUPPI, and AK8 PF+PUPPI. The results are shown in Figure 5.10. Such studies have been done for the first time in Run 2 and in future can help to assign more accurate uncertainties for various jet collections. Last, but not least, the study with the PF+PUPPI jets is of particular interest as it will be new default jet collection for LHC Run 3.

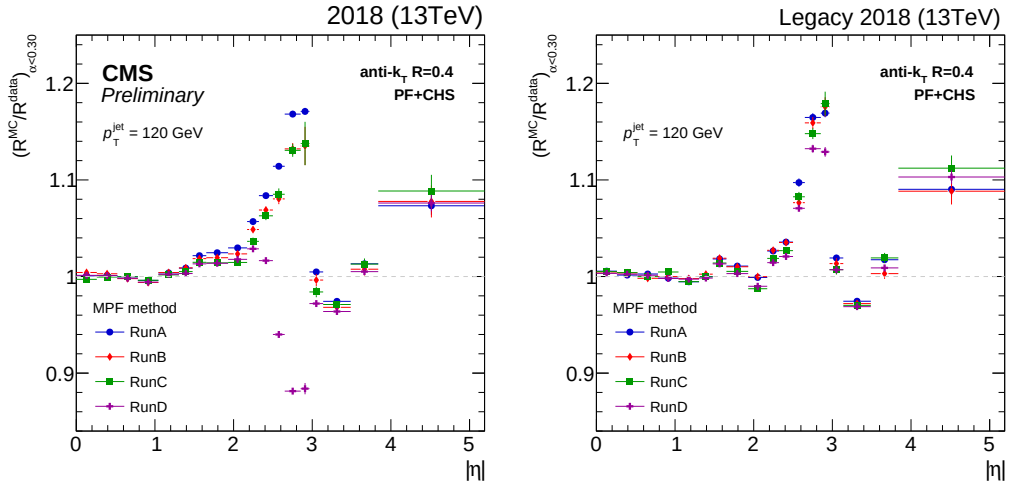

Figure 5.9 – Ratio of the relative responses of MPF method for jet with p_T = 120 GeV as a function of jet $|\eta|$ for the end-of-processing (left) and the Legacy data reprocessing. Published in Ref. [122].

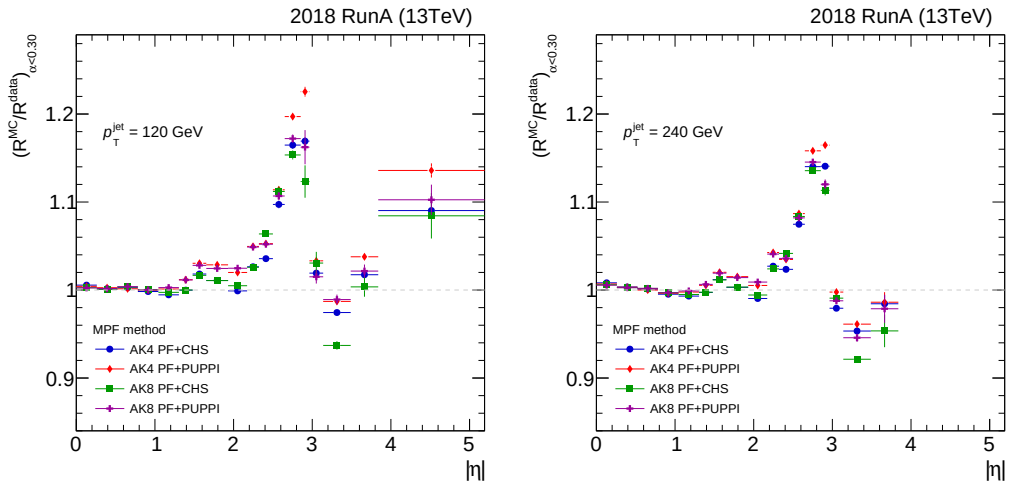

Figure 5.10 – Ratio of the relative response of MPF method for various jet collection as a function of jet $|\eta|$ for jet with p_T = 120GeV (left) and jet with p_T = 240GeV (right).

Part II

Physics analysis

6

Analysis strategy

The main objective of the analysis presented in this thesis is the measurement cross section of hadronically decaying W and Z bosons in the high transverse momentum region in proton-proton collisions. In this kinematic region, the decay product of W and Z bosons are collimated and boosted in the direction of the mother particles. Thus, the vector bosons are reconstructed as large cone-size jets with characteristic substructure. The analysis is aiming for the differential measurement in bins of the jet p_{T}. Formally written, the differential cross section in bins of the transverse momenta p_{T} is given as:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}p_{\mathrm{T}}} = \frac{N_{\mathrm{V-tagged\,jets}}}{\epsilon\,\mathcal{A}\,\mathcal{L}\,\Delta p_{\mathrm{T}}}, \tag{6.1}$$

where $N_{\mathrm{V-tagged\,jets}}$ stands for the number of V-tagged jets. The detector effects are corrected by the acceptance $\mathcal{A}$ and the efficiencies ϵ. The size of the used dataset corresponds to the collected luminosity $\mathcal{L}$. Finally, Δp_{T} stands for bin width.

6.1 Physical conditions

Several experimental challenges have to be addressed in order to perform this measurement. The nominal final state consists of two or more large cone-size jets with the high transverse momenta. One of them can be identified as W and Z bosons, while the other ones are the recoiling jets produced in order to balance the system in the transverse plane. These jets are typically initialized by light-flavour quarks.

Despite the total inclusive vector bosons production cross section is rather large compared to other processes as illustrated in Figure 1.9, the main challenge of this analysis arise from using the hadronic decay channels. A similar final state is typical for jet production described by QCD. The measurement of the inclusive jet cross-section with jet $p_{\mathrm{T}} > 100\,\mathrm{GeV}$ and $|y| < 3.0$ has been recently performed [126] and the fiducial production cross section was found to be $\sigma = 1845 \pm 4\,(\mathrm{stat})^{+119}_{-120}\,(\mathrm{syst})$ nb [126]. In contrast, the fiducial cross section corrected for branching fraction of the associated production of Z bosons and jets, namely Z+ $\geq$ 1jets measured with lepton pairs in the final state, have been found to be $\sigma = 116 \pm 0.3\,(\mathrm{stat}) \pm 10\,(\mathrm{syst})$ pb [127]. Such large difference between cross-sections makes the inclusive jets the dominant background processes for this analysis. Hence, a detailed study of jet substructure is needed for background reduction. In addition, the proper modelling of the jet mass distribution is required for the signal extraction. The physical measurement conditions can be summarized in following aspects:

- W and Z boson tagging: The jet mass and substructure seem to be suitable variables for the tagging of jets originating from the hadronic decay of W and Z bosons. In particular, the two-body decay of the bosons leads to the two-prong structure. This can be utilized to discriminate the signal from the background. As discussed earlier, special efforts need to be made during the tagging in order to avoid possible mass sculpting.
- QCD modelling: The associated production of vector bosons and jets has an identical final state as the multijets and dijets events. Thanks to high production cross-section of the inclusive jets, the QCD jets are the dominant contribution. By utilizing jet substructure, the QCD jets that are typically initiated by light-flavour quarks and gluon can be partially suppressed as outlined above.
- Irreducible background: Apart from the QCD jets, the contribution of the $t\bar{t}$, tW and single top needs to be considered as it results in similar final states. That means that the boosted W bosons can also originate from the decay of a t quark or associated production with a single t. This contribution will be referred as an irreducible background. Currently, the only possible way for its subtraction relies on the MC simulations.
- Jet mass resolution: Regrettably, the jet exhibits a relatively bad mass resolution. The typical mass resolution measured in the semileptonic $t\bar{t}$ events is around $\sim$ 10 GeV. Therefore, without any dedicated heavy flavour tagging, the jets originating from the hadronic decay of W and Z bosons cannot be separated.

6.2 Measurement strategy

These physical conditions have a direct impact on the measurement strategy of the presented analysis. The measurement strategy can be divided into three parts, namely event selection & corrections, vector boson tagging, and signal extraction. All parts of the analysis are briefly discussed in the following subsections. A more detailed description is then given in the corresponding sections.

6.2.1 Event Selection & Corrections

Events selections include all essential steps to properly identify proton-proton collisions of interest that will be further analysed. Chapter 7 summarises the cuts used for the event selections applied in data and MC simulations. Then the physics objects selection criteria are described. The selection criteria on the physics objects follow the central recommendations provided by the corresponding Physics Object Groups (POG) of the CMS experiment. It enables us to use the centrally measured and derived scale factors which aim to mitigate the residual disagreement between the reconstruction and identification inefficiencies observed in data and MC simulations.

6.2.2 Vector boson tagging

For the purpose of this analysis, a dedicated mass decorrelated vector boson tagger was constructed. The tagger is based on the N_2^1 variable that is well suited for tagging objects with two-prong structure like jets originating from W and Z bosons decay. The mass decorrelation is achieved by fixed efficiency regressions. Before using the tagger in the analysis, the validation of the tagger together with the measurement of the scale factor to correct for the difference in the performance in data and MC simulations is performed.

This is possible by using tag-and-probe in the semileptonic $t\bar{t}$ events. In more detail, the vector boson tagging is described in Chapter 8.

6.2.3 Signal extraction

Chapter 9 describes the background subtraction and signal extraction, respectively. First, the MC simulations used for the subtraction of the irreducible background is validated in the dedicated control region. The signal extraction relies on a data-driven method which utilizes the control region dominated by the background contribution and a transfer function to predict the background contribution in the signal region. The measured distributions are simultaneously fitted in order to extract the differential cross section. Several checks of the fit model are given to validate the fitting technique. The impact of all relevant systematic uncertainties is also discussed. Finally, the unfolded cross sections to the so-called stable particle level is given in Section 10. In this section, the discussion on results and its comparison to the MC simulation predictions is given.

The given measurement strategy is adopted from an analyses aiming to search for signatures beyond the SM in boosted topologies. In recent years, advanced experimental techniques were developed and successfully employed in searches in order to maximize the discovery potential. Newly, these techniques have been started to be utilized for more SM-like measurements. In contrast to the searches, the signal and its properties like mass and branching ratio are known for the SM measurement. This allows us to further optimize the selection criteria and cuts.

As an example, the techniques used for search [63] for a narrow, low-mass, scalar and pseudoscalar resonances decaying to bottom quark-antiquarks pairs can be easily modified and optimized for the search for a boosted Higgs [128]. In such a topology, the Higgs is assumed to be reconstructed as large cone-size jet with two prong structure. In addition, both subjets originate from the bottom quarks. Figure 6.1 shows the results on the measurement of boosted Higgs decaying to bottom quark-antiquarks pairs. In the signal region, i.e. passing double-b tagging, the contribution from Z boson exhibits.

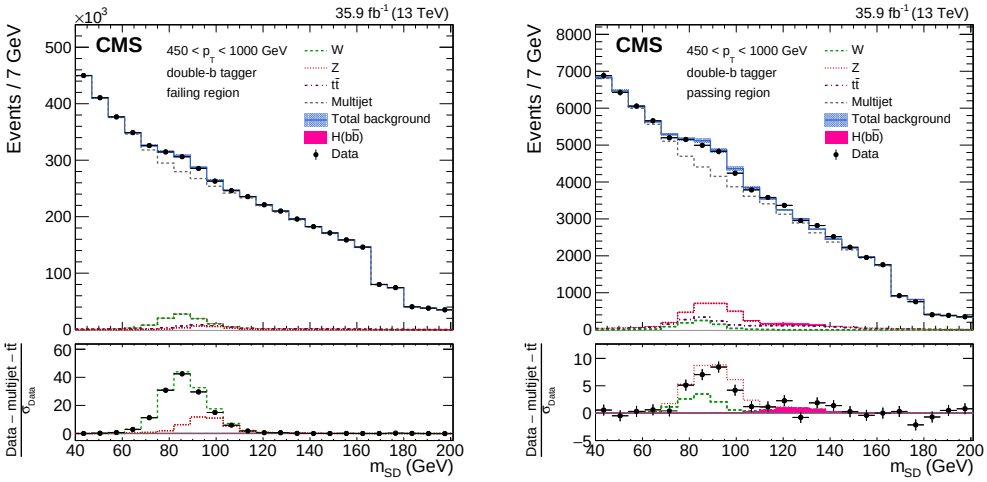

Figure 6.1 – Distribution of the soft drop jet mass in data for the failing (left) and passing (right) double-b tagger selection [128]. The background contribution from W and Z bosons is visible.

7

Event samples and physics objects selection

Event and physics object selection are the essential first steps of each analysis. For the presented measurement two different topologies, namely dijet and single muon+jet, will be utilized. While the dijet topology will be used for the differential cross-section measurement of the W and Z bosons in the hadronic channel, the single muon+jet topology will be utilized as a control region. The proper event and object selection in this topology opens the door to the phase-space of a clean sample of events enriched of hadronically decaying W boson and one leptonically decaying W boson, both of them coming from the decay of $t\bar{t}$. This control region will be used for the validation of the custom tagger of boosted W and Z bosons. In the following chapter, we discuss the used datasets and simulated samples together with the physics object selection. The corrections applied to simulations are reviewed.

7.1 Recorded data event samples

The analysis presented in this thesis is performed by using proton-proton collisions at a centre-of-mass energy of 13 TeV recorded by the CMS experiment in years 2016, 2017, and 2018, corresponding to an integrated luminosity of 35.9, 41.5, and 59.7 fb^{-1}, respectively. The total recorded integrated luminosity is 137 fb^{-1}. The data sets are divided into run periods due to technical stops of the LHC and changing the configuration of the CMS detector. Only runs passing certain selection criteria and marked as good events for physics analysis are included in these datasets, that are stored in the so-called `MiniAOD` data format. The latest reconstructions of the data sets containing well redefined detector calibrations are used. It ensures the usage of the latest and up-to-date physics object reconstructions in the analysis. Table 7.1 summarizes the information about the used data sample together with corresponding information about the run period, reconstruction version, run ranges, and integrated luminosity.

Physics objects which are expected to be present in events of interest have an impact on the type of dataset employed in the analysis. The dijet topology consisting of at least two jets is studied by using the `JetHT` datasets. For the background studies with leptonically decaying W, the `SingleMuon` dataset is more appropriate for this topology than the `JetHT` datasets, mainly because these events are better to trigger and record by the single muon triggers.

Dataset name	Run Period	Reconstruction version	Run ranges	Int. lumi / fb^{-1}
JetHT, SingleMuon	Run2016B	17Jul2018-v*/MINIAOD	272007-275376	5.75
JetHT, SingleMuon	Run2016C	17Jul2018-v1/MINIAOD	275657-276283	2.57
JetHT, SingleMuon	Run2016D	17Jul2018-v1/MINIAOD	276315-276811	4.24
JetHT, SingleMuon	Run2016E	17Jul2018-v1/MINIAOD	276831-277420	4.03
JetHT, SingleMuon	Run2016F	17Jul2018-v1/MINIAOD	277772-278808	3.10
JetHT, SingleMuon	Run2016G	17Jul2018-v1/MINIAOD	278820-280385	7.58
JetHT, SingleMuon	Run2016H	17Jul2018-v1/MINIAOD	280919-284044	8.65
	Run2016			total: 35.9
JetHT, SingleMuon	Run2017B	31Mar2018-v1/MINIAOD	297046 – 299329	4.79
JetHT, SingleMuon	Run2017C	31Mar2018-v1/MINIAOD	299368 – 302029	9.63
JetHT, SingleMuon	Run2017D	31Mar2018-v1/MINIAOD	302030 – 303434	4.25
JetHT, SingleMuon	Run2017E	31Mar2018-v1/MINIAOD	303824 – 304797	9.31
JetHT, SingleMuon	Run2017F	31Mar2018-v1/MINIAOD	305040 – 306462	13.54
	Run2017			total: 41.5
JetHT, SingleMuon	Run2018A	17Sep2018-v1/MINIAOD	315252 – 316995	14.02
JetHT, SingleMuon	Run2018B	17Sep2018-v1/MINIAOD	317080 – 319310	7.06
JetHT, SingleMuon	Run2018C	17Sep2018-v1/MINIAOD	319337 – 320065	6.89
JetHT	Run2018D	PromptReco-v*/MINIAOD	320673 – 325175	31.74
SingleMuon	Run2018D	22Jan2019-v*/MINIAOD	320673 – 325175	31.74
	Run2018			total: 59.7

Table 7.1 – Summary of data samples used in the analysis. The information about the run period, reconstruction version, run ranges, and integrated luminosity is given.

7.2 Simulated event samples

Each measurement relies on the comparison with the simulated event samples to shed lights on the physics of the studied processes. Last, but not least, the simulated samples are used to develop and optimize the analysis technique. Simulated samples for each year of data taking used in this work are centrally produced by the CMS Collaboration and include full detector simulation done by GEANT4. For appropriate comparison with the data, the simulated events need to be reconstructed by using exactly the same reconstruction algorithms and detector calibration as used for the data.

Events containing W and Z bosons in association with jets are generated by MADGRAPH5_aMC@NLO at QCD LO accuracy. The hadronic decay of the vector boson is then performed within the MADSPIN framework. While the ME calculation for Z boson production is performed for up to 4 noncollinear high p_T partons, the ME for the production of W bosons includes only up to 3 partons. Different jet multiplicities are matched by the MLM method with the matching scale $q_{cut} = 19$ GeV. The dominant background of QCD multijet is simulated by PYTHIA8. The other processes containing hadronically decaying W bosons in the final state are $t\bar{t}$ and single-t production simulated by POWHEG and POWHEG v2.0, respectively. In addition, the events containing diboson production were explored, but because of the rather low cross-section, their contribution is considered to be negligible. Looser selection criteria in the single muon+jet topologies result in possible contributions from the leptonic decay of W and Z bosons. These processes are generated by MADGRAPH5_aMC@NLO.

Dataset/channel/phase space		MC generator	Cross section / pb	Label
W+jets W$\rightarrow q\bar{q}'$	$400 < H_T \leq 600$ $600 < H_T \leq 800$ $800 < H_T$	MadGraph5_aMC@NLO +MLM matching	$1.456 \cdot 10^2$ $3.388 \cdot 10^1$ $1.864 \cdot 10^1$	W($q\bar{q}$) +jets
Z+jets Z$\rightarrow q\bar{q}'$	$400 < H_T \leq 600$ $600 < H_T \leq 800$ $800 < H_T$	MadGraph5_aMC@NLO +MLM matching	$3.143 \cdot 10^2$ $6.885 \cdot 10^1$ $3.472 \cdot 10^1$	Z($q\bar{q}$) +jets
$t\bar{t}$ + jets	hadronic semileptonic leptonic	Powheg	$3.139 \cdot 10^2$ $3.009 \cdot 10^2$ 72.1	$t\bar{t}$ +jets
Single $t/\bar{t}$ inclusive	t-channel tW-channel	Powheg v2.0 + MadSpin Powheg v2.0	$1.8121 \cdot 10^2$ $6.988 \cdot 10^1$	single-t
QCD	$120 < \hat{p}_T \leq 170$ $170 < \hat{p}_T \leq 300$ $300 < \hat{p}_T \leq 470$ $470 < \hat{p}_T \leq 600$ $600 < \hat{p}_T \leq 800$ $800 < \hat{p}_T \leq 1000$ $1000 < \hat{p}_T \leq 1400$ $1400 < \hat{p}_T \leq 1800$ $1800 < \hat{p}_T \leq 2400$ $2400 < \hat{p}_T \leq 3200$ $3200 < \hat{p}_T$	Pythia8	$4.071 \cdot 10^5$ $1.033 \cdot 10^5$ $6.826 \cdot 10^3$ $5.526 \cdot 10^2$ $1.566 \cdot 10^2$ $2.632 \cdot 10^1$ 7.500 $6.479 \cdot 10^{-1}$ $8.715 \cdot 10^{-2}$ $45.242 \cdot 10^{-3}$ $1.349 \cdot 10^{-4}$	QCD
W+jets W$\rightarrow l\nu$	+1 jet + 2 jets + 3 jets + 4 jets	MadGraph5_aMC@NLO +MLM matching	$8.140 \cdot 10^3$ $2.789 \cdot 10^3$ $9.904 \cdot 10^2$ $5.488 \cdot 10^2$	W($l\nu$) + jets
Z+jets Z/$\gamma^* \rightarrow ll$ $m_{ll} > 50$ GeV	$100 < H_T \leq 200$ $200 < H_T \leq 400$ $400 < H_T \leq 600$ $600 < H_T \leq 800$ $800 < H_T \leq 1200$ $1200 < H_T \leq 2500$ $2500 < H_T$	MadGraph5_aMC@NLO +MLM matching	$1.612 \cdot 10^2$ $4.826 \cdot 10^1$ 7.030 1.742 $8.057 \cdot 10^{-1}$ $1.916 \cdot 10^{-1}$ $3.481 \cdot 10^{-3}$	Z(ll) + jets
WW WZ ZZ	hadronic inclusive inclusive	Pythia8	$7.565 \cdot 10^1$ $2.758 \cdot 10^1$ $1.215 \cdot 10^1$	VV

Table 7.2 – Summary of simulated MC samples used in the analysis. The information about the dataset, phase space cuts, decay channels decay, and the corresponding cross section is given. The transferred momentum of the primary interaction $\hat{p}_T$ ans the scalar sum of transverse jet momenta H_T are quoted in GeV.

The PYTHIA8 event generator is used for parton shower, hadronization and the underlying event simulation. The NNPDF3.0 (NNPDF3.1) set of parton distribution functions together with the CUETP8M1 (CP5) underlying event tune is used for the MC sample for the 2016 (2017 and 2018) data sets. For the modelling the top quark production for 2016 data sets the dedicated tune (CUETP8M2T4) that was developed to optimize the colour reconnection used in PYTHIA8. A comprehensive summary of simulated processes considered in this analysis together the phase space cuts and decay channels is given in Table 7.2.

The generated samples need to be normalized to their expected contributions in order to accurately describe the data. The event weight for a simulated sample with N processed events of the physics process with the cross section σ is defined as

$$\omega = \frac{\mathcal{L} \cdot \sigma}{N}. \tag{7.1}$$

Here $\mathcal{L}$ corresponds to the integrated luminosity of the collected datasets with which these simulated samples are compared.

7.3 Event and physics objects selection

The event and physics objects selection defines the phase-space of the measurement. The appropriate selection helps to suppress the background contribution, while at the same time it should maximize the signal contribution. By following the central recommendations for the physics objects selection by various POGs of the CMS experiment, the simulated events should reasonable well describe the data. In the following subsections, the event selections along with the physics objects selections with cuts applied on the physics quantities are presented. The selection criteria are applied in recorded data as well as in simulated events, if not stated otherwise.

7.3.1 Vertex selection and MET filters

Only events with the reconstructed PV as defined in Section 4.2 are considered for further analysis steps. In addition, the events need to pass selection criteria related to $p_{\mathrm{T}}^{\mathrm{miss}}$. The large value of $p_{\mathrm{T}}^{\mathrm{miss}}$ can indicate the presence of undetectable objects like neutrino in the events or it could be considered as a sign of the BSM. On the other hand, the origin of abnormally large $p_{\mathrm{T}}^{\mathrm{miss}}$ can be simply just the electronic and detector noise, cosmic rays, detector miscallibration, beam halo and so on. These kind of events should be rather rare and dedicated algorithms with binary decision have been developed to reject them. They are known as MET filters. As an example, the main objective of Flag_HBHENoiseFilter is to reject events with the anomalous signals within the calorimeters, that most likely arise from noise in the read-out electronics. The Flag_BadPFMuonFilter aims to reject events, where the high p_{T} tracks are not used for muon tracking as they are identified as low quality muons. But they still enter in the calculation of $p_{\mathrm{T}}^{\mathrm{miss}}$ and typically result in a large $p_{\mathrm{T}}^{\mathrm{miss}}$. The MET filters applied in data and simulations are listed in Table 7.3.

7.3.2 Jet selection

Two types of the jet collections are considered in this work. The boosted vector bosons are identified with AK8 PF+PUPPI jets. As grooming technique, by default CMS uses

MET filter name	Applied in data 2016/2017/2018	Applied in MC 2016/2017/2018
Flag_goodVertices	✓/✓/✓	✓/✓/✓
Flag_globalSuperTightHalo2016Filter	✓/✓/✓	✓/✓/✓
Flag_HBHENoiseFilter	✓/✓/✓	✓/✓/✓
Flag_HBHENoiseIsoFilter	✓/✓/✓	✓/✓/✓
Flag_EcalDeadCellTriggerPrimitiveFilter	✓/✓/✓	✓/✓/✓
Flag_BadPFMuonFilter	✓/✓/✓	✓/✓/✓
Flag_eeBadScFilter	✓/✓/✓	- / - / -
Flag_ecalBadCalibReducedMINIAODFilter	- /✓/✓	- /✓/✓

Table 7.3 – MET filters used in the analysis.

the soft drop algorithm with parameters $z_{\text{cut}} = 0.1$ and $\beta = 0$. The soft drop jet mass is then corrected by dedicated corrections supplied by JME POG. Those corrections are derived in the context of measurements with boosted W bosons (in Bulk Graviton signal sample) and are needed to have a p_{T} and η independent W-jet soft drop mass with a mean around the real W mass. The derivation of this correction is done in two steps. First, the p_{T}-dependent correction accounts for the shift at generator level. Then the difference between the reconstructed and the generated soft drop jet mass is corrected by a p_{T} and η-dependent factor. The two corrections are applied to the uncorrected soft drop jet mass both in data and in simulation as

$$M_{\text{SD}} = M_{\text{SD,uncorr}} \cdot w_{\text{GEN}} \cdot w_{\text{RECO}}, \tag{7.2}$$

where w_{GEN} and w_{RECO} stand for the generator and reconstruction corrections, respectively and $M_{\text{SD,uncorr}}$ is the uncorrected soft drop jet mass.

In addition to the large cone-size jets, we use also AK4 PF+CHS jets in the single muon+jet topology. They are utilized to identify jets initialized by the b quark, so called b jets. The identification of them is based on the characteristic properties of b hadrons like long lifetime($c\tau = 450\mu$m), large mass (~ 5 GeV), high track multiplicity, and large semileptonic branching fraction. The b tagging technique takes advantage of these characteristic properties and is based on the track information, secondary vertex information, and the presence of non-prompt leptons can assign to each jet a probability whether it comes from b quark or not. In this work, the DeepCSV [129] algorithm for b tagging is used. This algorithm is based on the CSVv2 [129] algorithms, it uses the same jet features but extends the range of the maximum of tracks per jets and exploits modern deep neutral network architecture. As working point, we use the medium working point which correspond to $\sim 1\%$ mistagging rate of light-flavour (udsg) jets in $t\bar{t}$ events as measured by BTV POG.

Jets are required to have $|\eta| < 2.4$ and the cut on the transverse momenta varies with the topology. The cross section measurement is performed with highly boosted W and Z bosons, therefore the jets are required to have $p_{\text{T}} > 500$GeV. On the other hand, the single muon+jet topology is probed in a slightly different phase-space. This topology makes use of the AK8 PF+PUPPI and AK4 PF+CHS jets with $p_{\text{T}} > 200$ GeV and $p_{\text{T}} > 30$ GeV, respectively.

All jets are required to pass a basic selection related to the constituents of the jets and

energy fraction carried by each type of hadrons. These selection criteria are called as JetID and ensure that noise jets and fake jets, for example if muons or electrons are reconstructed as jets, are rejected. The cuts are η dependent with an average efficiency close to 99%. The selection criteria for jets within the central region, i.e. $|\eta| < 2.4$, are listed in Table 7.4.

Variable	**Cut**
Neutral hadron fraction	< 0.90
Neutral EM fraction	< 0.90
Number of constituents	> 0
Muon fraction	< 0.80
Charged hadron fraction	> 0
Charged Multiplicity	> 0
Charged EM fraction	< 0.90 for 2016 < 0.80 for 2017, 2018

Table 7.4 – Criteria for the JetID definition.

The four-momenta of both AK8 PF+PUPPI and AK4 PF+CHS jets are corrected by JECs that are centrally supplied by JME POG. As discussed earlier, as consequence of the jet energy resolution being better in MC simulation than in data, the jets' energy in MC simulation needs to be smeared. This is done by the hybrid method, explained in Subsection 5.2.3. It is important to stress that even after the jet energy is fully corrected, additional corrections namely the jet mass scale and resolution for AK8 PF+PUPPI jets will be necessary. Historically, the jet mass corrections are derived independently after the calibrating the jet energy. In contrast to that, the ATLAS Collaboration has recently started to explore the possibility to calibrate jet energy and mass for large cone-size jets simultaneously [130].

7.3.3 Missing transverse momentum selection

In this work, the PF+PUPPI $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ with Type-1 correction is used. The final state of boosted vector bosons and recoiling jets is expected not to include genuine $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, hence the cut on $p_{\mathrm{T}}^{\mathrm{miss}}$ can be utilized to suppress the background contribution with neutrinos in the final state. Examples of such background contributions are the $t\bar{t}$, and single-t productions as well as W+jets production, where W boson is decaying leptonically.

7.3.4 Lepton selection

Electrons and muons are identified via selection criteria provided by the EGamma and Muon POG, respectively. Only lepton candidates with $p_{\mathrm{T}} > 20$ and $|\eta| < 2.4$ are considered. While the muons are required to pass identification and PF isolation criteria, the electrons have to fulfil only a cut-based identification which already includes the information about the PF isolation. The PF isolation is calculated as

$$R_{\mathrm{iso}} = \left[\sum_{\substack{\text{charged}\\\text{hadrons}}} p_{\mathrm{T}} + \max(0, \sum_{\substack{\text{neutral}\\\text{hadrons}}} p_{\mathrm{T}} + \sum_{\text{photons}} p_{\mathrm{T}} - p_{\mathrm{T}}^{\mathrm{PU}}) \right] / p_{\mathrm{T}}^{l}, \tag{7.3}$$

where the sums run over the photons, charged and neutral hadrons in a cone defined by $\Delta R = 0.4$ and 0.3 around the muon and electron trajectory, respectively. The $p_{\mathrm{T}}^{\mathrm{PU}}$ denotes

the contribution of neutral particles from pile-up. The isolation criterium in the lepton selection is used for identifying events with prompt leptons. As non-prompt leptons we denote leptons originating from heavy flavour decays (e.g. from b quark decay).

For electron candidates, the PF isolation criterium depends also on the electron η. As an example we can mention loose cut-based electron identification (cutBasedElectronID-Fall17-94X-V2-loose). For this specific cut-based identification, only electron candidates with $R_{\text{iso}} < 0.112 + 0.506/p_{\text{T}}$ and $R_{\text{iso}} < 0.108 + 0.963/p_{\text{T}}$ in the ECAL barrel ($|\eta| < 1.479$) and ECAL endcap ($|\eta| > 1.479$) region, respectively, are considered as isolated.

For the measurement of boosted W and Z boson cross section, the presence of a loose lepton works as event veto. It helps us to eliminate the background contribution as our final state is believed to be without leptons. As loose electron and muon, the physics object need to pass following selection:

- $p_{\text{T}} > 20\,\text{GeV}$ and $|\eta| < 2.4$,
- for electron, identification known as cutBasedElectronID-Fall17-94X-V2-loose need to be passed,
- for muon, identification known as LooseMuonID and $R_{\text{iso}} < 0.25$ need to be passed.

In contrast, a tighter muon selection is used in the single muon+jet topology to achieve lower background contamination. Muons in this topology are required to have $p_{\text{T}} > 55\text{GeV}$ and $|\eta| < 2.1$. Futhermore they have to pass tight identification criteria, i.e. TightMuonID and $R_{\text{iso}} < 0.15$.

7.3.5 Trigger selection

This work makes use of the single jet triggers, defined only by the minimum transverse momentum of the leading AK8 jet as events of interest are assumed to contain only jets. The trigger decision needs to be done quite fast, therefore the trigger system has a very fast reconstruction algorithm, but not so precise as offline reconstruction. The energy of online reconstructed (HLT) jets are typically underestimated. Thus, the trigger efficiency must be measured and the region where the triggers become fully efficient needs to be determined in terms of the offline jets p_{T}. As fully efficient, we consider a trigger with efficiency larger than 99% and the value of p_{T}, where the 99% is reached, is called as a turn-on point.

In the presented work, the trigger efficiency was measured by the tag-and-probe method. The main principles of this technique are the following. First, we select events with the dijet final state and we require a back-to-back topology, i.e. $\Delta\phi_{12} > 2.7$. In order to suppress additional jet activity, all other jets are required to have $p_{\text{T}} < 0.3 \cdot \frac{p_{\text{T}}^1 + p_{\text{T}}^2}{2}$. The leading and subleading HLT jets are matched with the selected jets if $\Delta R\,(\text{HLT}, \text{jets}) < 0.4$. Second, the leading and subleading jets are randomly assigned as the tag and probe jets. If the HLT jets matched with the tag jet has $p_{\text{T}} > X$, where the X is the online trigger threshold, we can use the probe jet to measure the trigger efficiency as

$$\epsilon\,(p_{\text{T}}) = \frac{N\left(\text{Probe jet } p_{\text{T}} \text{ with } p_{\text{T,probe}}^{\text{HLT}} > X\right)}{N\,(\text{Probe jet } p_{\text{T}})}. \tag{7.4}$$

The measured distribution is then fitted by

Trigger name	2016		2017		2018	
	Offline threshold/GeV	Integrated lumi./fb	Offline threshold/GeV	Integrated lumi./fb	Offline threshold/GeV	Integrated lumi./fb
HLT_AK8PFJet200	258	0.085	263	0.189	271	0.202
HLT_AK8PFJet260	314	0.513	339	0.470	343	0.465
HLT_AK8PFJet320	383	1.510	410	1.228	412	1.237
HLT_AK8PFJet400	484	4.545	492	9.660	513	3.711
HLT_AK8PFJet450	550	33.182	564	9.660	578	7.373
HLT_AK8PFJet500	632	33.182	613	41.527	621	59.684

Table 7.5 – Summary of single AK8 jet triggers together with offline thresholds and recoreded integrated luminosity.

$$\epsilon\left(p_{\mathrm{T}}\right) = a + 0.5 \cdot (1-a) \cdot \left(1 + \mathrm{Erf}\left(\frac{p_{\mathrm{T}} - \mu}{\sigma}\right)\right), \tag{7.5}$$

where Erf stands for the error function defined as

$$\mathrm{Erf}\left(x\right) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t. \tag{7.6}$$

The free parameters of fit are following ones: a, μ, and σ. The trigger turn-on point will be used as offline trigger threshold for given trigger and the given trigger can be used from this value to the turn-on point of the next trigger. The offline thresholds of various single AK8 jet triggers are extracted for each year of data-taking separately and given in Table 7.5. As QCD processes are very abundant in proton-proton collisions, these triggers are pre-scaled. Only the trigger with the highest online threshold is without any pre-scale. The integrated luminosity collected by each trigger needs to be taken into account to reweight data to obtain a proper physical behaviour of corresponding observables.

7.4 Corrections to MC simulations

When comparing data to MC simulations, a possible disagreement can be observed. The source of this disagreement can be various. From modelling of proton-proton collisions by MC event generators and the simulation of the detector response and algorithm performance to a hint of new physics. While the modelling of the underlying physics of proton-proton collisions can be improved by MC tuning, the difference in the detector response and algorithm performance between data and MC simulations is mimicked by multiplicative corrections applied as additional event weight factors. The latter are usually referred to as scale factors and depend on various event variables or the properties of the physics objects in a given event. When deriving scale factors, events containing well known physics processes are explored by a data-driven technique. The scale factors then act as multiplicative factors to the nominal event weight given by Eq. 7.1. The resulting event weight is then given as

$$\omega = \omega_{\mathrm{nominal}} \prod_{i \in \mathrm{corrections}} \omega_i \tag{7.7}$$

In the following Subsections, the corrections to MC simulations that are relevant for the presented work are discussed.

7.4.1 Number of pileup interactions

The MC simulations are generated with a reasonable estimation of the pileup profile that is expected in data. As MC simulation are typically produced before the whole data are available, the modelling of the pileup can differ from the one observed in data. The pileup profile in data depends on the luminosity, beam condition etc. In order to correct the difference in pileup profile, pileup reweighting is used and applied to the MC simulation. The pileup μ, in other words said the total number of proton-proton interactions per bunch crossing can be calculated as

$$\mu = \mathcal{L}_{\text{inst}} \frac{\sigma_{\text{inel}}}{f_{\text{rev}}}, \tag{7.8}$$

where $\mathcal{L}_{\text{inst}}$ stands for an instantaneous luminosity of a given single bunch, σ_{inel} is the total proton-proton inelastic cross section and f_{rev} is the orbit frequency. In case of the LHC collider, the orbit frequency is equal to 11246 Hz. As the lumi section is the fundamental unit of the CMS luminosity calculation, the pileup can be calculated on a per-lumi section basis. The lumi section is 23.3 seconds long. The pileup which is calculated from the instantaneous luminosity is the average pileup during a single lumi section. Such obtained distribution of pileup for individual events will follow a Poisson distribution.

The nominal data pileup is estimated by the Brilcalc tool [131] and a value of 69.2 mb is used for the total inelastic cross-section in calculation. The variation of 4.6% uncertainty in the total inelastic cross-section is used to calculate the variation of ± 1 standard deviation for pileup reweighting. Figure 7.1 shows the pileup profile in data and MC simulation before any corrections. In case of data, the pileup profile is shown for the nominal value of the total inelastic cross-section as well as for $\pm 4.6\%$ variations of the nominal value. The ratio of the pileup profile for data and MC simulation is used to reweight the pileup distribution in MC simulation. After pileup reweighting, the pileup profile in data and MC simulation will match.

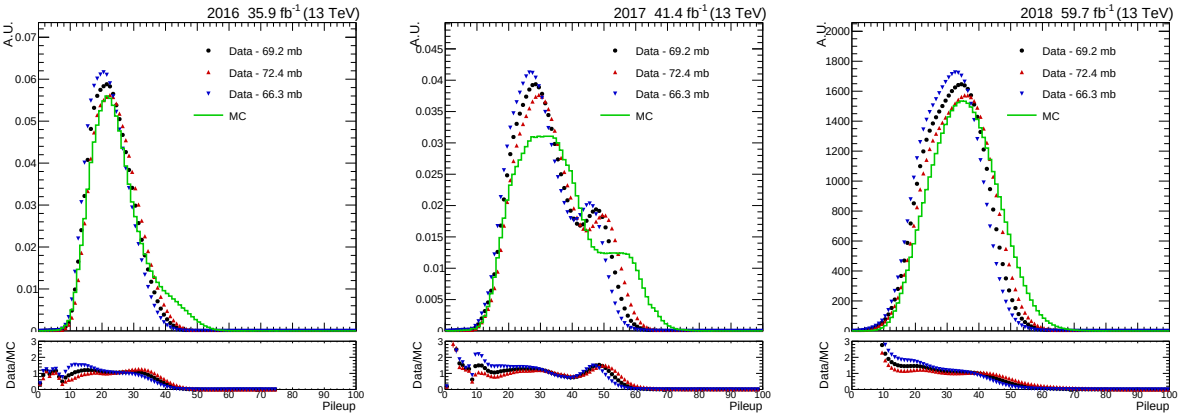

Figure 7.1 – Distributions of the number of pileup interactions in data and simulation for 2016 (left), 2017 (middle), and 2018 (right) data-taking periods.

7.4.2 Triggers

Single jet triggers

As single jet triggers reach the p_{T} region where they are more than 99% efficient, trigger emulation is not required in the MC simulation. Thus, no scale factor is needed.

Single muon triggers

For single muon + jet topology, events must contain a high p_{T} muon. Such events are rather rare, therefore they can be collected by unprescaled trigger. In this work, the trigger HLT_Mu50 are used and the reconstructed muons are required to have $p_{\mathrm{T}} > 53$ GeV. As the muon triggers do not reach a region, where they are fully efficient and in addition the trigger efficiency depends on the muon η, we need to apply the trigger emulation also in the MC simulation. The trigger efficiency in data and MC is measured by the Muon POG by the tag-and-probe method. The different trigger performance in data and MC is corrected by centrally provided scale factors. As Figure 7.2 shows, the scale factors are p_{T} and $|\eta|$ dependent.

For 2016 data the scale factor SFs were measured separately for the beginning (RunBCDEF) and the end (RunGH) of data-taking period. The final scale factor for 2016 is estimated as weighted sum of the two scale factors using the integrated luminosities $\mathcal{L}$ corresponding to the RunBCDEF and RunGH periods. Formally written as

$$SF = \frac{SF\,(\mathrm{RunBCDEF}) \cdot \mathcal{L}\,(\mathrm{RunBCDEF}) + SF\,(\mathrm{RunGH}) \cdot \mathcal{L}\,(\mathrm{RunGH})}{\mathcal{L}\,(\mathrm{RunBCDEFGH})}. \tag{7.9}$$

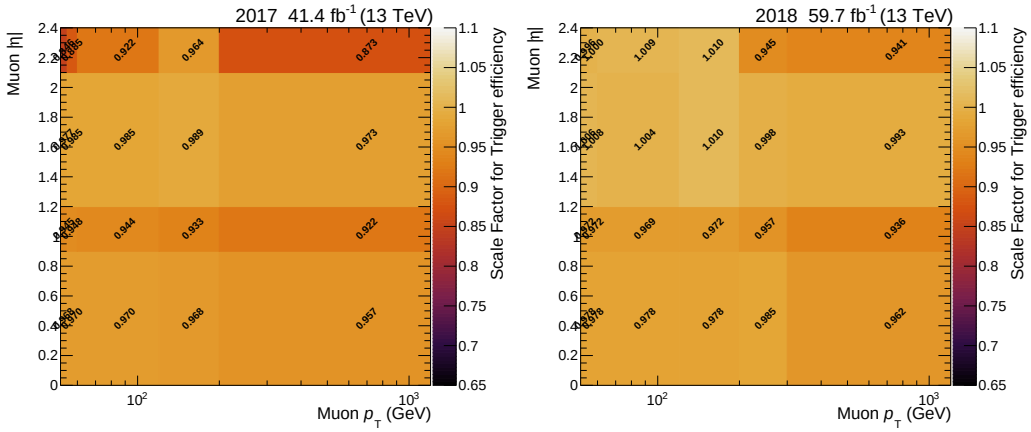

Figure 7.2 – Scale factor for the trigger efficiency measured for HLT_Mu50 trigger as a function of muon p_{T} and $|\eta|$ for 2017 (left) and 2018 (right) data-taking periods. Reproduced from [132].

Prefiring issue

The Level-1 prefiring issue refers to the timing shift of ECAL with respect to the Level-1 trigger primitives. As a consequence a significant fraction of high η trigger primitives are mistakenly associated to the previous bunch crossing. Since the Level-1 trigger rules forbid two consecutive bunch crossing to fire, the event can veto itself if a large ECAL energy is found in the region $2 < |\eta| < 3$. This effect is present in 2016 and 2017 data set

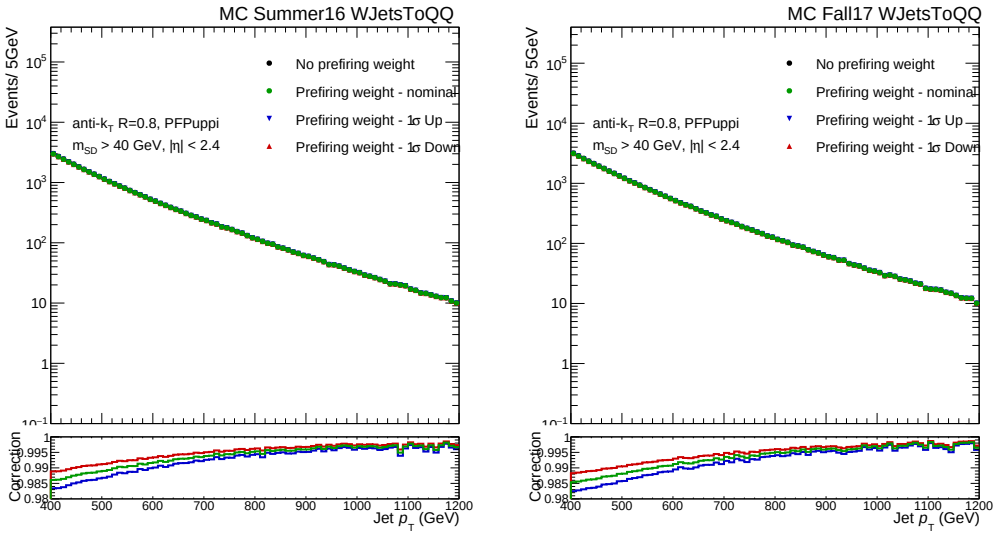

Figure 7.3 – Number of events as a function of jet p_{T} for signal samples and impact of correction on the Level-1 prefiring issue.

and is not modelled in the MC simulation. Thus, the MC simulation has to be corrected. The recipe delivered by JME POG was implemented in the analysis and the event weight is calculated as the product of the non prefiring probability of all jets in events. Formally written as

$$\omega = 1 - P\,(\text{prefiring}) = \prod_{i=\text{jets}} \left(1 - \varepsilon_i^{pref}\,(\eta, p_{\mathrm{T}})\right), \tag{7.10}$$

where $\varepsilon_i^{pref}\,(\eta, p_{\mathrm{T}})$ stands for the prefiring probability of jets in an event. The prefiring probability maps are centrally provided by JME POG. As an example of the impact of the Level-1 prefiring issue, Figure 7.3 shows the number of events as a function of jet p_{T} for signal samples with and without the Level-1 prefiring correction. Around 1.5% of events for jets with $p_{\mathrm{T}} \sim 500$ GeV are lost due to prefiring issue. The impact of prefiring issue is smaller for jets with higher jet $p_{\mathrm{T}} \sim 1$ TeV, since these jets are produced more likely closer to the central region, i.e. with smaller $|\eta|$. As an uncertainty for the Level-1 prefiring issue the maximum between 20% of the probability of prefiring and the statistical uncertainty of the probability of prefiring is taken. The resulting uncertainties related to the Level-1 prefiring issue are of the order of few per-mill.

7.4.3 Muon efficiency

The identification and PF isolation criteria of lepton reconstruction are monitored by the corresponding POG. The measured efficiency of a given procedure in data and MC simulation is used to compute the scale factors which mitigate the different performance. The scale factors for muon and electron are centrally provided by Muon and EGamma POG, respectively. The resulting scale factor ω used in the analysis is obtained as a product of the individual scale factors and reads

$$\omega = \omega_\mu^{\mathrm{ID}} \cdot \omega_\mu^{\mathrm{iso}}. \tag{7.11}$$

For illustration purpouse, the scale factors for loose muon identification and isolation criterium are shown in Figure 7.4. The scale factors are close to unity.

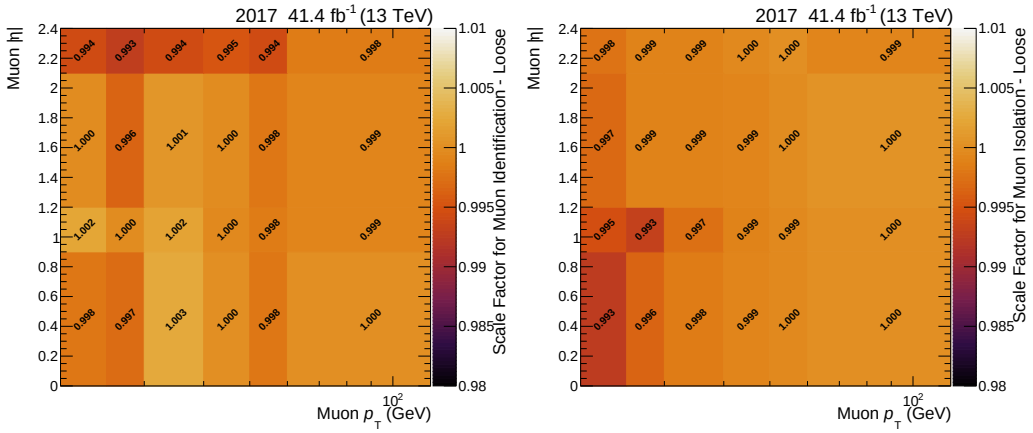

Figure 7.4 – Scale factor for the muon isolation (left) and identification (right) for loose selection criteria shown as function of muon p_{T} and $|\eta|$. Reproduced from [132].

7.4.4 b tagging efficiency

The performance of the b tagging in data and MC simulation is studied and dedicated scale factors are then derived. In general, the scale factors depend on the jet flavour, the jet p_{T} and η. The SF for the b tagging efficiency of b jets and light-flavour jets are centrally measured by the BTV POG. In the presented analysis, we will use the scale factors measured by the "QCD method" [129] which makes use of the fact that b jets can a contain non-isolated muon from the semileptonic decay of b quark.

There is variety of different methods for applying the scale factors to MC simulation. In this work, the method based on the event reweighting using scale factors and MC b tagging efficiencies is employed. This method is trying to predict the correct event yields by changing the weight of the selected MC events. The probability of given configuration of b, c, and light-flavour jets in data and MC simulation being $b-$tagged or non $b-$tagged is given by

$$P\,(\mathrm{MC}) = \prod_{i=\mathrm{tagged}} \epsilon_i \prod_{j=\mathrm{not\ tagged}} (1-\epsilon_j)\,, \tag{7.12}$$

$$P\,(\mathrm{Data}) = \prod_{i=\mathrm{tagged}} \mathrm{SF}_i\epsilon_i \prod_{j=\mathrm{not\ tagged}} (1-\mathrm{SF}_j\epsilon_j)\,, \tag{7.13}$$

where ϵ_i is b tagging efficiency measured in MC simulations. The products run over all $b-$tagged and non $b-$tagged jets in the event. The event weight ω is then given as

$$\omega = \frac{P\,(\mathrm{Data})}{P\,(\mathrm{MC})} = \prod_{i=\mathrm{tagged}} \mathrm{SF}_i \prod_{j=\mathrm{not\ tagged}} \frac{(1-\mathrm{SF}_j\epsilon_j)}{(1-\epsilon_j)}. \tag{7.14}$$

This method requires the knowledge of the scale factor SF and the tagging efficiency ϵ. While the scale factors are centrally provided by the BTV POG and are applicable for all

analyses, the tagging efficiency measured in MC simulation can differ from one analysis to another as it depends on the cut selections and the phase-space. The b tagging efficiency in MC samples is defined as

$$\epsilon_f\left(p_{\mathrm{T}},\eta\right) = \frac{N_f^{b\text{-tagged}}\left(p_{\mathrm{T}},\eta\right)}{N_f^{\mathrm{Total}}\left(p_{\mathrm{T}},\eta\right)}, \tag{7.15}$$

where $N_f^{b\text{-tagged}}\left(p_{\mathrm{T}},\eta\right)$ and $N_f^{\mathrm{Total}}\left(p_{\mathrm{T}},\eta\right)$ are the total number of $b-$tagged and the total number of jets, respectively of flavour f in the given p_{T} and η bins. The jet flavour is based on the hadron flavour definition and we can distinguish between b, c, and light-flavour jets. While $\epsilon_b\left(p_{\mathrm{T}},\eta\right)$ corresponds to the b tagging efficiency, i.e. percentage of truth b jets that are $b-$tagged, $\epsilon_c\left(p_{\mathrm{T}},\eta\right)$ and $\epsilon_{udsg}\left(p_{\mathrm{T}},\eta\right)$ provides us with the information about the mistagging rate. In other words, $\epsilon_c\left(p_{\mathrm{T}},\eta\right)$ and $\epsilon_{udsg}\left(p_{\mathrm{T}},\eta\right)$ corresponds to the contamination coming from events, when c jets or light-flavour jets are mistakenly $b-$tagged.

7.4.5 Higher order corrections

The last type of corrections, that is discussed are the higher order corrections. Since the generation of MC events together with the full detector simulation are costly in time and resources, a certain compromise between the number of generated events and the accuracy in the calculation of ME are needed. Less frequently used samples are typically produced with LO accuracy. To improve modelling, additional corrections knows as k-factors are applied. They are derived as a ratio of given observables modelled by MC simulations with NLO and LO accuracy. In this work, the correction will depend on the p_{T} of the given particle at the generator level before any decay.

NLO QCD and NLO EW corrections

The MC simulations for signal processes are generated by MadGraph5_aMC@NLO at LO QCD accuracy. As our measurement aims to probe the production of W and Z bosons in the high transverse region, higher order corrections as demonstrated in Ref. [53] will have significant impact on the measured cross section. In contrast to MC simulation with LO accuracy, the transverse momenta spectra of vector bosons are significantly softer.

The NLO QCD K-factor was derived by using centrally produced MC simulation in NLO QCD accuracy. Samples are listed in Table 7.6.

Dataset/channel/phase space		MC generator	Cross-section/pb
W+jets W$\rightarrow l\nu$	+ 0 jet + 1 jet + 2 jets	MadGraph5_aMC@NLO +FxFx matching	5333.4 984.2 362.5
Z+jets Z/$\gamma^* \rightarrow ll$ $m_{ll} > 50$ GeV	+ 0 jet + 1 jet + 2 jets	MadGraph5_aMC@NLO +FxFx matching	54549.4 9104.8 3376.8

Table 7.6 – Simulated NLO QCD sample used to derive NLO QCD correction.

As those samples contain different decay channels of vector bosons, the cross sections are multiplied by the $\mathcal{B}\left(Z \rightarrow qq'\right)/\mathcal{B}\left(Z \rightarrow ll\right)$ and $\mathcal{B}\left(W \rightarrow qq'\right)/\mathcal{B}\left(W \rightarrow l\nu_l\right)$, respectively.

These factors correct the different cross-section for different decay channels. Figure 7.5 shows the extracted k-factor. The ratio is parametrized by

$$f\left(p_{\mathrm{T}}\right) = a \cdot e^{-b \cdot p_{\mathrm{T}}} + c, \tag{7.16}$$

where a, b, and c are free parameters of fit. The red line in Figure 7.5 stands for the best fit, while the light red band represents the 68% confident interval. The larger NLO k-factors than those reported in Ref. [53], correct also for the issue in the central CMS MC production. Several MC simulations for the associated production of vector boson with jets generated by MadGraph5_aMC@NLO v2.6.2 in LO are affected by a wrong setting of `pdfwgt` parameter [133].

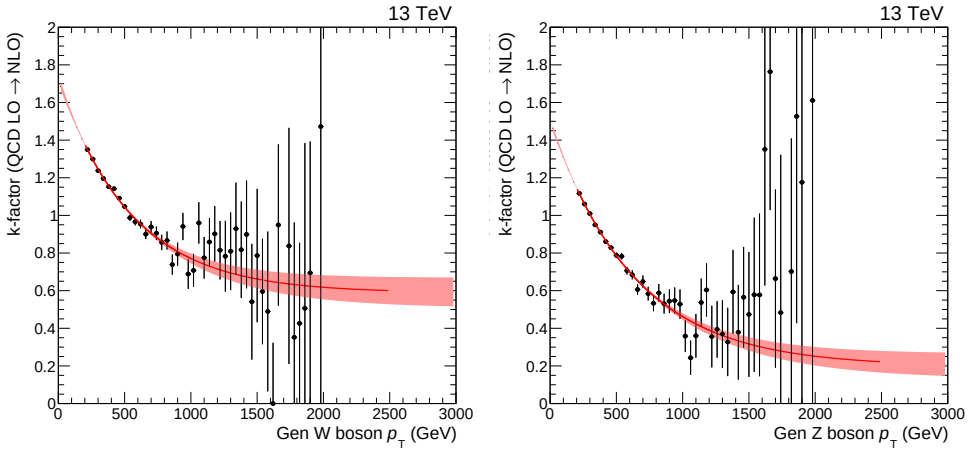

Figure 7.5 – Extracted NLO QCD k-factor as function of the generated W (left) and Z (right) bosons p_{T}. The red lines stands for the best fits.

The NLO EW corrections are extracted from [53]. As Figure 7.6 shows, NLO EW corrections at TeV scale play important role and make the p_{T} spectrum softer.

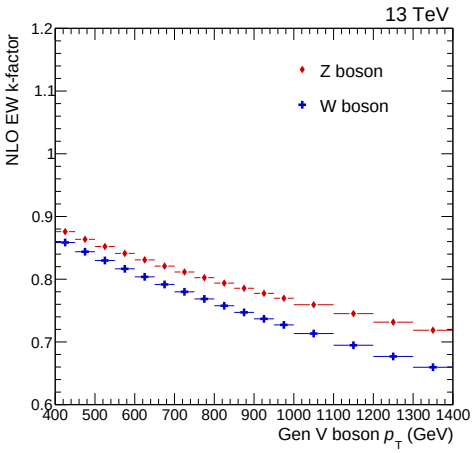

Figure 7.6 – NLO EW correction as function of boson p_{T}. Reproduced from [53].

Top p_{T} reweighting

A significantly softer top p_{T} spectrum in data compared to various MC predictions has been observed by many analyses during the LHC Run 1 and Run 2. Since many analyses rely on a reasonable description of the top p_{T} spectrum in data by MC simulation, the TOP PAG provides a recipe for the top p_{T} spectrum reweighting. The reweighing function is based on the parton level results of the top p_{T} spectrum measured in the dilepton and lepton+jet channel. The mismodelling of data by MC simulation is parametrized by

$$SF\left(p_{\mathrm{T}}\right) = e^{0.0615 - 0.0005 \cdot p_{\mathrm{T}}}. \tag{7.17}$$

The event weight will be given as $w = \sqrt{SF\left(t\right) SF\left(\bar{t}\right)}$. Since the top p_{T} reweighting is based on results from the dilepton channel, this method is validated up to $p_{\mathrm{T}}\left(t\right) < 800$ GeV. In our case, we have applied the top p_{T} reweighting on $\mathrm{t\bar{t}}$ MC simulation without any p_{T} restriction. As discussed later, an additional validation will be performed for the high p_{T} region.

7.5 Summary of event and physics objects selection

Events are considered for the measurement of differential cross section of W and Z boson if they pass the following cuts:

- At least one of the single AK8 jets trigger fired and leading jet p_{T} is above the offline threshold.
- The presence of loose muon or electron reconstructed in the tracker region is imposed as event veto to suppress the contribution from $\mathrm{t\bar{t}}$ and single-t production. The leptonic decay of W and Z bosons is also eliminated.
- Events are rejected if PF+PUPPI $p_{\mathrm{T}}^{\mathrm{miss}} > 150$GeV. Such large value of $p_{\mathrm{T}}^{\mathrm{miss}}$ indicate the presence of neutrinos in the event.

The leading and subleading jets reconstructed in the tracker region $|\eta| < 2.4$ are considered as candidates for jets originating from the hadronic decay of W or Z bosons, if their transverse momenta are larger than 500 GeV.

Figure 7.7 shows jet quantities like jet p_{T}, η, ϕ, and soft drop jet mass. All relevant corrections to MC simulations are applied. A reasonable agreement between data and MC simulations is found. The soft drop jet mass distribution shown in Figure 7.7 (4th row) exhibits a peak structure for MC simulations of various processes around $\sim$ 80 GeV as they contain hadronically decaying W bosons. Because of the dominant contribution from QCD multijet production, no peak structure is observed in data yet.

The identification of jets coming from the hadronic decay of W and Z boson will utilize the information about the jet structure. Figure 7.8 shows the following jet substructure variable: N_2^1, τ_2/τ_1, and ρ defined as $2\log(m_{\mathrm{SD}}/p_{\mathrm{T}})$. Noticeable difference in the comparison of MC simulation to data is visible for various years of data-taking periods. Despite of performing the measurement in the high p_{T} region, the impact of modelling the underlying physics encoded in the UE tune is observed. The jet substructure variables are sensitive to modelling of PS. The MC simulations compared with data from 2016 use CUETP8M1 tune based on the Monash [134] tune, while the samples for 2017 and 2018 utilize CP5 tune.

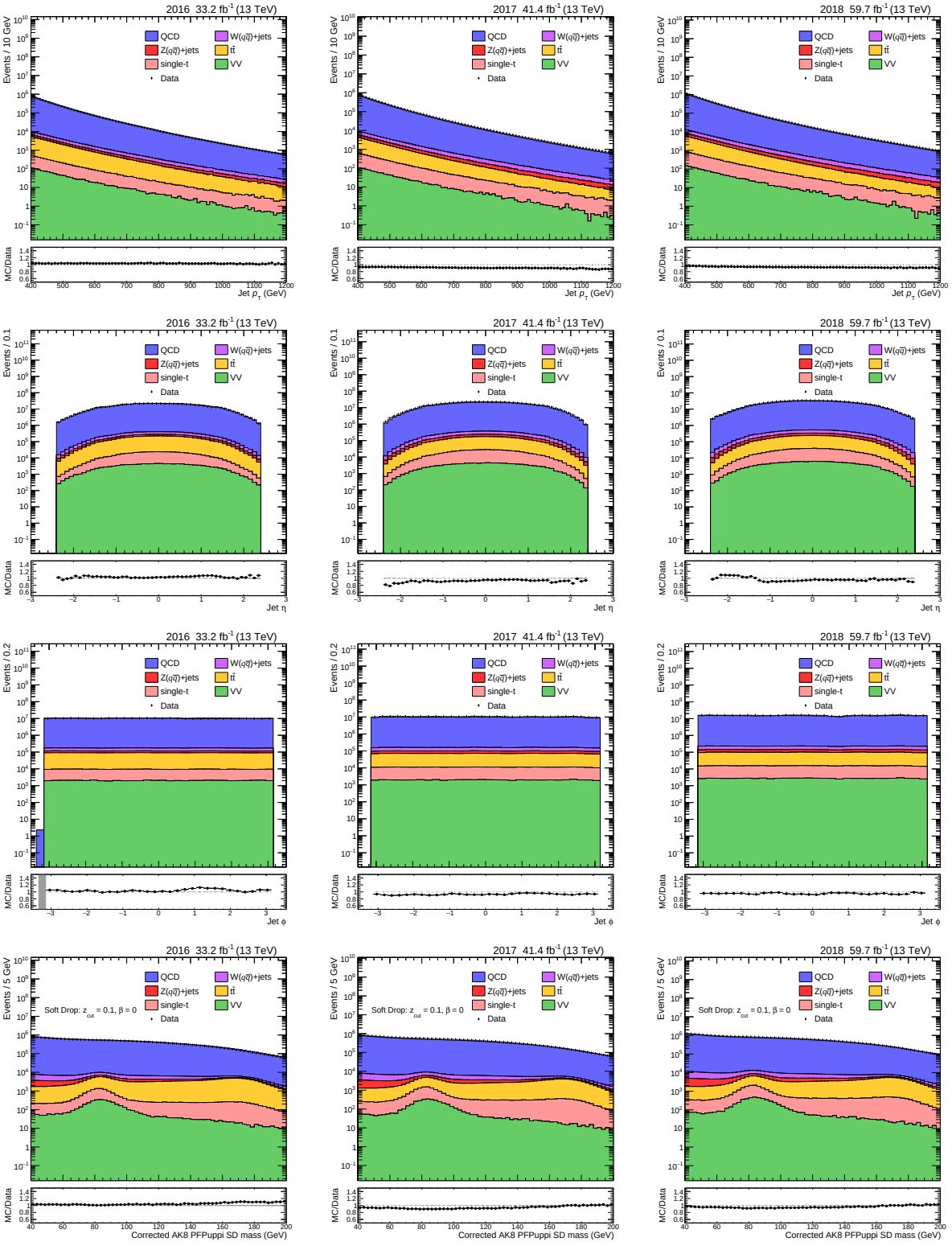

Figure 7.7 – Distributions of the jet p_{T} (upper), η (2nd row), ϕ (3rd row), and soft drop jet mass (4th row) in data and MC simulations for 2016 (left), 2017 (middle), and 2018 (right) data-taking periods.

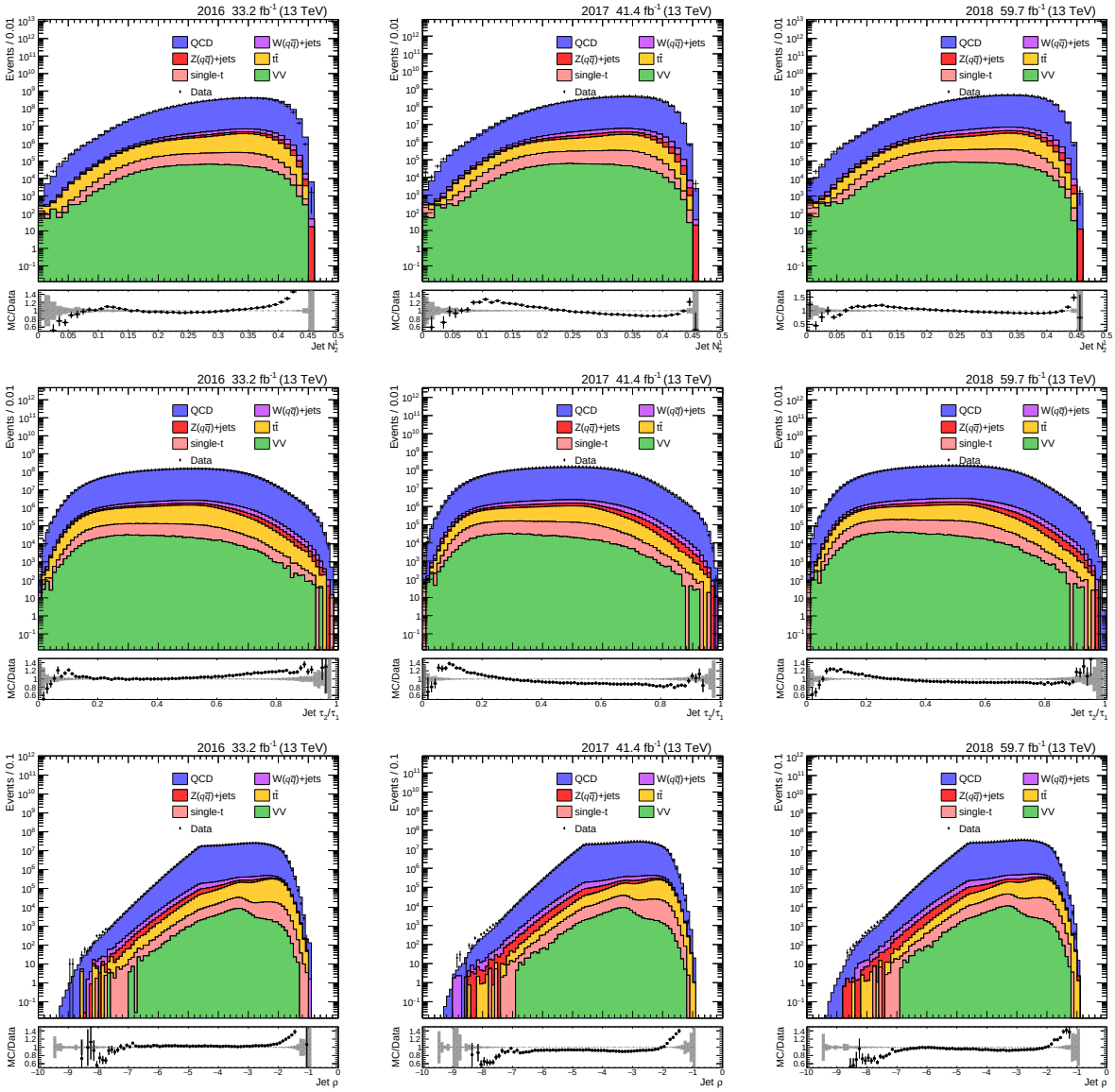

Figure 7.8 – Distributions of the jet substracture variables, namely N_2^1 (upper), $\tau 2/\tau 1$ (2nd row), jet ρ (3rd row) in data and MC simulations for 2016 (left), 2017 (middle), and 2018 (right) data-taking periods.

8

Vector boson identification

In order to identify jets originating from the hadronic decay of W and Z bosons, the soft drop jet mass and jet substructure variables are explored. For this purpose the tagger based on two-prong substructure information will be constructed. The main objective of this tagger is to preserve the steeply falling soft drop jet mass distribution of the dominant background process, the QCD multijet production. Such properties of the tagger are achieved by the mass decorrelation approach. The construction is based fully on MC simulation, therefore its validation is needed. The commissioning of the custom tagger will be done in the single muon + jet topology. As discussed in Subsection 2.3.1, using only τ_{21} or N_2^1 would introduce the mass sculpting.

8.1 Construction of mass decorrelated tagger

As the QCD multijet production is the dominant background process with much higher cross-sections than the associated vector boson production with jets, the presented measurement can be seen like a search for a bump in the jet mass distribution. In contrast to searches, where the mass of a new hypothetical heavy resonances is unknown, in our case the peak structure is expected to be around actual masses of W and Z bosons.

As main variables for vector boson tagging, N_2^1 and soft drop jet mass will be used. In contrast to τ_{21}, the definition of N_2^1 is axis free. It makes this two-prong variable more stable against the soft drop jet mass and transverse momentum. Hence, it is more suitable for a mass decorrelated tagger. In this work, the tagger is based on N_2^1 and the mass decorrelation is gained from the fixed efficiency regression (cf Subsection 2.3.1) in 2D phase-space. The decorrelation is done against the jet transverse momenta p_{T} and jet ρ. The jet ρ is a dimensionless variable defined as

$$\rho = 2 \log \left(m_{\mathrm{SD}} / p_{\mathrm{T}} \right) . \tag{8.1}$$

Here m_{SD} stands for the soft drop jet mass. Since the jet mass scales with p_{T} for QCD jets, a decorrelation of a given substructure variable against p_{T} and ρ is a well-bounded procedure. The decorrelation is performed for fixed QCD multijet efficiency. In our particular case, $\epsilon_{\mathrm{QCD}} = 15\%$ is chosen as a working point. It means that only 15% of QCD multijet events in a given bin of phase-space will pass V tagging selection according to the MC simulation. This procedure results in a 2D distribution in the phase-space of

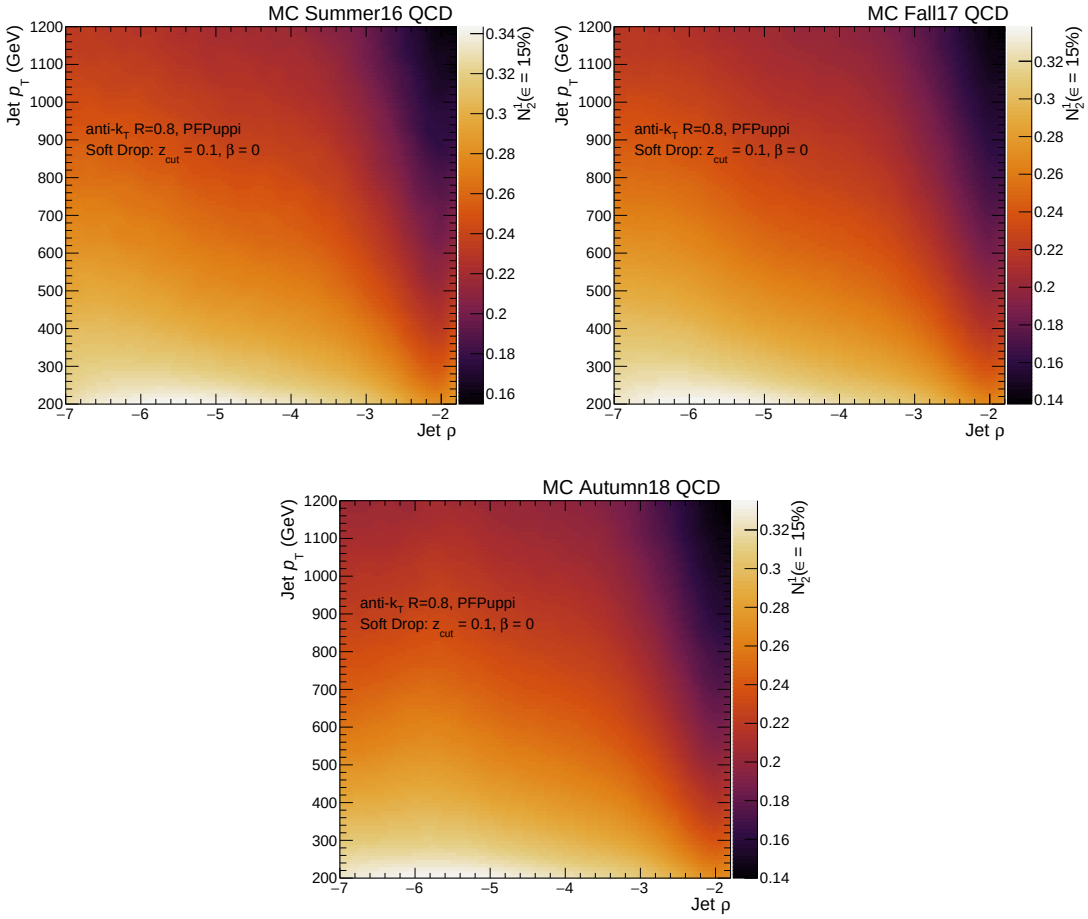

Figure 8.1 – 2D decorrelation maps as a function of jet p_{T} and ρ for V tagging with the working point $\varepsilon_{\mathrm{QCD}} = 15\%$.

the jet p_{T} and ρ. For a given point of this phase-space $N_2^1\left(p_{\mathrm{T}}, \rho;\, \epsilon_{\mathrm{QCD}} = 15\%\right)$ is known as is demonstrated in Figure 8.1.

Finally, the tagger is constructed as

$$N_2^{1,\mathrm{DDT}}\left(p_{\mathrm{T}}, \rho\right) = N_2^1 - N_2^1\left(p_{\mathrm{T}}, \rho;\, \epsilon_{\mathrm{QCD}} = 15\%\right). \tag{8.2}$$

Only jets with $N_2^{1,\mathrm{DDT}}\left(p_{\mathrm{T}}, \rho\right) < 0$ are considered as $V-$tagged jets, that means as jets possibly originating from the hadronic decay of W and Z bosons.

Furthermore, the cut on jet ρ is imposed and jets are required to have ρ in a range from -6.2 to -1.9 . The lower boundary corresponds to the non-perturbative regime of the soft drop jet mass calculation, while the upper boundary is related to the so-called finite cone-size effects. According to Eq. 2.5, the opening angle between quarks from the W boson decay is equal to 0.8, only if the transverse momentum of boson is at least 200 GeV. This condition is equivalent with $\rho \approx -1.83$. A lower transverse momentum of the boson will result in a larger opening angle and the decay products will be no longer reconstructed as one jet with the radius $R = 0.8$.

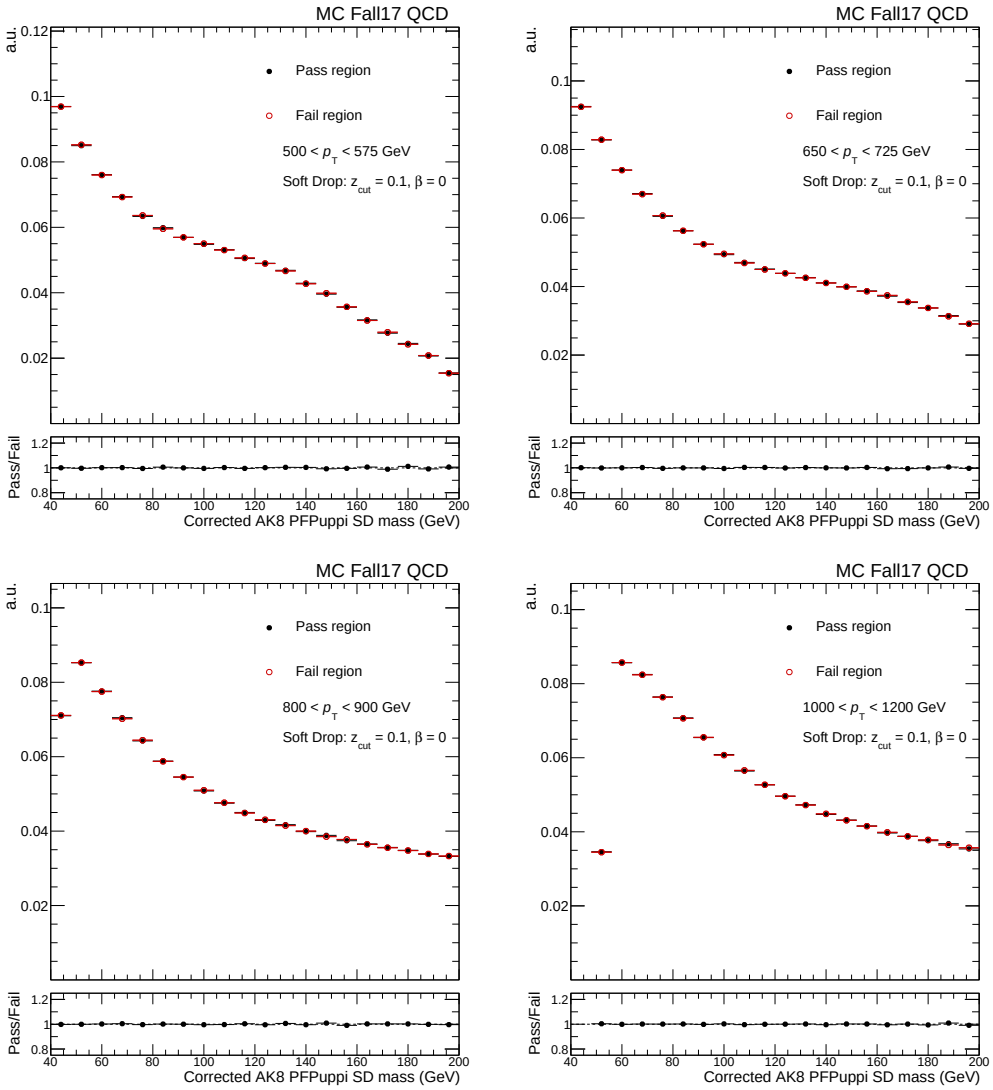

Figure 8.2 – Comparison of soft drop jet mass distribution for QCD multijets events in pass and fail region after applying V tagging.

8.1.1 Closure test

The construction of the mass decorrelated tagger relies fully on MC simulation of QCD multijet production. As a closure test of the construction, the tagger is applied to the jets from this MC simulation. Figure 8.2 shows the comparison of the soft drop jet mass distribution for the pass $\left(N_2^{1,\mathrm{DDT}}\left(p_\mathrm{T},\rho\right)<0\right)$ and fail $\left(N_2^{1,\mathrm{DDT}}\left(p_\mathrm{T},\rho\right)>0\right)$ region for four various p_T bins. The measured distribution are normalized to unity. As is shown, the shapes of the distribution in the pass and fail region agree and no artificial peak structure is observed.

8.2 Commissioning of the tagger

The construction of the mass decorrelated tagger relies only on MC simulation. The observed discrepancy of modelling the jet substructure shown in Figure 7.8 could possible

bias the measurement.

8.2.1 Methodology

In order to validate the tagger, its performance needs to be measured in a region enriched by hadronically decaying W bosons. Furthermore the phase-space used for the tagger performance study needs to be orthogonal to the phase-space where the cross-sections of hadronically decaying W and Z bosons is measured.

This can be achieved by utilizing a topology containing semileptonic decays of a $\mathrm{t\bar{t}}$ system. There are several reasons for using this topology. First, at LHC energies a good number of $\mathrm{t\bar{t}}$ events are produced. In addition, the complex topology leads to a relatively low background contamination. The signature of this topology is a lepton, a neutrino measured as $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, two AK4 b jets, and one AK8 jet. A schematic sketch of this topology is shown in Figure 8.3.

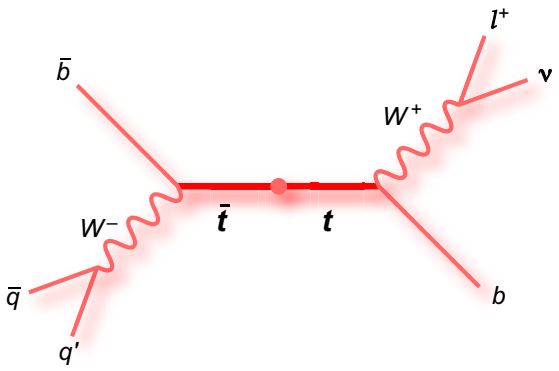

Figure 8.3 – Sketch of semileptonic decay of $t\bar{t}$ events.

The tag-and-probe method is utilized for the measurement of the V tagging efficiency. More precisely said, in this topology only the W tagging efficiency is probed. Nevertheless, no significant difference is believed to be expected for W and Z tagging efficiency, since both bosons create jets with very similar properties. The leptonic decay of the W is used as a tag, while the AK8 jets possibly coming from the hadronic decay of W bosons are considered as a probe for the examination of the custom tagger.

8.2.2 Event topology and selection

The measurement is performed with the `SingleMuon` datasets listed in Table 7.1. Events are collected by a dedicated high p_{T} muon trigger, so-called HLT_Mu50 trigger, and are considered for further processing if they fullfill the following selection criteria:

- Exactly one muon with $p_{\mathrm{T}} > 53\,\mathrm{GeV}$ passing tight ID selection.
- Events are rejected if they contain additional leptons passing the loose ID selection.
- $p_{\mathrm{T}}^{\mathrm{miss}}$ needs to be larger than 40 GeV.
- The presence of at least one b jet with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$ is required.
- AK8 PF+PUPPI jet with the highest p_{T} is considered as a candidate for W tagging. The only condition imposed on the jet is a cut on $|\eta| < 2.4$.

W bosons decaying leptonically are reconstructed from the selected muon and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$. In order to reach the boosted topology, the reconstructed W bosons are required to have

$p_T > 200$ GeV. Furthermore, the following criteria related to the angular separation of the reconstructed objects are required:

- $\Delta R\,(\text{lepton, AK8 jet}) > \pi/2$,
- $\Delta\phi\,(\vec{p}_T^{\,\text{miss}}, \text{AK8 jet}) > 2$,
- $\Delta\phi\,(\text{leptonic W, AK8 jet}) > 2$.

After the event and physics object selection, the corresponding scale factors are applied to the MC simulations. As discussed in the previous chapter, the b tagging efficiency as well as the mistagging rate of c and light-flavour jets for our phase-space has to be known for applying b tagging SF. The measurement of flavour tagging efficiency is done in bins of jet p_T and $|\eta|$ as shown in Figure 8.4. The binning scheme follows the same binning as the provided scale factor. The medium working point of b tagging corresponds to 1% mistagging rate of light-flavour jet as defined by BTV POG. A similar mistagging rate is also observed in our topology as shown in Figure 8.4(right).

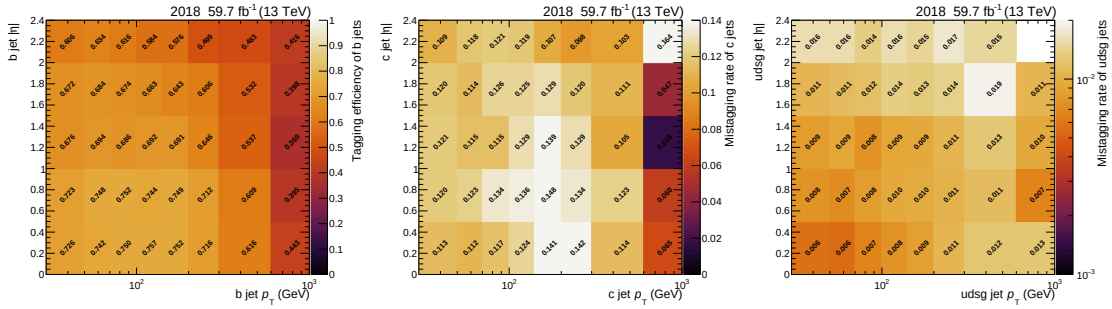

Figure 8.4 – Measured $b-$tagging efficiency(left), $c-$(middle) and $udsg-$(right) mistagging rate for 2018 data as a function of jet p_T and $|\eta|$.

Variables related to the jets are shown in Figure 8.5. A fairly good agreement between data and MC simulation is found. In contrast to the QCD multijet topology, where an enriched sample of gluon jets is being expected, jets in $t\bar{t}$ events are most likely initiated by quarks. Thus a better modelling of jet substructure variables is observed. As previously discussed, a different UE tune was used for the 2016 MC simulation (CUETP8M1) and 2017, 2018 MC simulation (CP5). The distributions for 2016 data exhibit very good agreement with the MC simulation, while some discrepancies are observed for 2017 and 2018 data.

Figure 8.6 shows the distribution of the soft drop jet mass after physics objects selection and applying the custom tagger. The main contribution comes from the signal process, semileptonic decay of $t\bar{t}$, as expected. The minor background contributions of the leptonic decay of $t\bar{t}$, inclusive sample of single top+W as well as the associated production of W boson with jets are observed. The peak structure around actual mass of W boson is observed.

To understand which part of the soft drop jet mass distribution of $t\bar{t}$ events contain *real* W jets and which are just combinatorial background, *fake* W jets, a matching between the reconstructed jets and quarks q_i from the hadronic decay of W bosons is done. As the *real* W jets those for which $\Delta R\,(\text{jet}, q_i) < 0.6$ are considered. Moreover, the condition $\Delta R\,(\text{jet}, b) > 0.6$ where b stands for a b quark from top decay, is required to reject fully merged jets which should be rather considered as top-jet. As illustrated in Figure 8.6(bottom), the soft drop jet mass distribution consists of *real* W jets and the combinatorial

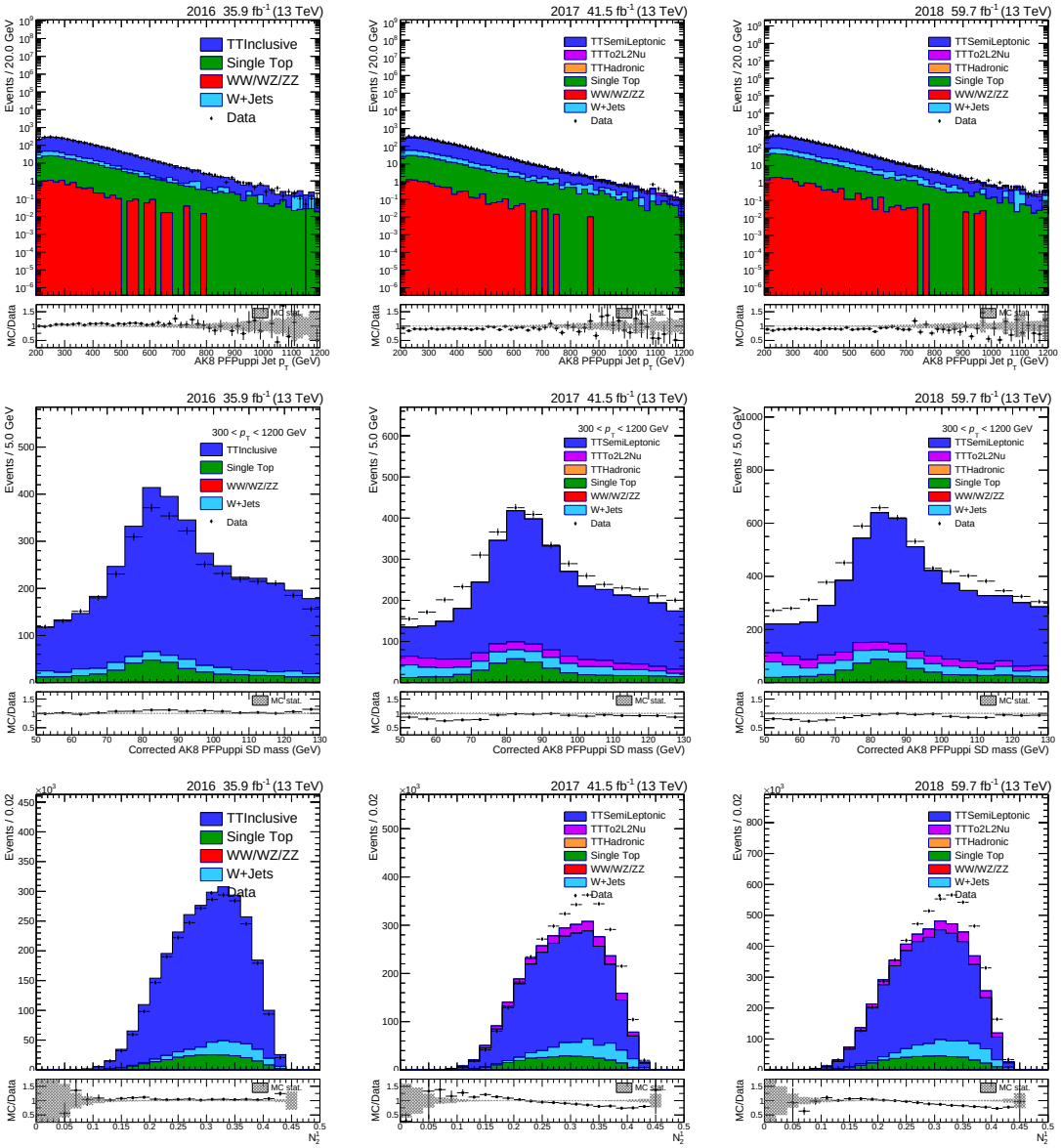

Figure 8.5 – Distribution of the reconstructed jet p_{T} (upper), soft drop jet mass (middle), and substructure variable N_2^1 (bottom) for 2016 (left), 2017 (middle), and 2018(right) data.

background. The *real* W jets follow a Gaussian distribution with a mean value around actual mass of W boson.

8.2.3 Fitting procedure and extraction of W tagging efficiency

The determination of the W tagging efficiency makes use of the leading large cone size jets in the events. Firstly, the tagger is applied to these jets and they are grouped into two categories: pass $\left(N_2^{1,\mathrm{DDT}} < 0\right)$ and fail$\left(N_2^{1,\mathrm{DDT}} > 0\right)$ regions. Then the W tagging efficiency is extracted by the simultaneous fit of the soft drop jet mass distribution in these two regions. The fitting is done iteratively in the following three steps.

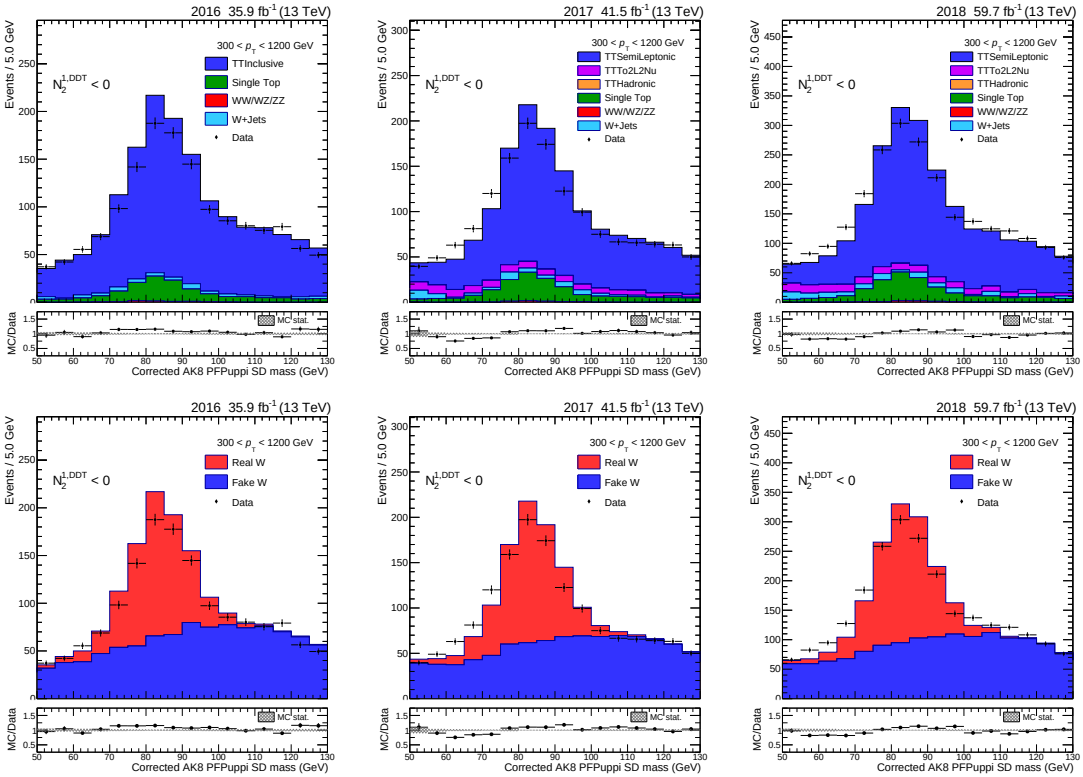

Figure 8.6 – Distribution of the reconstructed soft drop jet mass in single-μ + jet topology after physics objects selection and applying V-tagger for 2016 (left), 2017 (middle), and 2018(right) data. While the contribution of different processes is shown on the upper panel, the contribution of *real* W-jets and the combinatorial background is shown in the bottom panel.

8.2.4 Fits to signal processes

First, the jets in MC simulation for the signal sample are categorized as *real* and *fake* W jets according to the matching criteria described in the previous subsection. The distribution of the soft drop jet mass distribution of *real* W-jets is fitted by the double-sided Crystal ball function defined as

$$f^{\text{sig}}(x) = f_{\text{DSCB}}(x) = \begin{cases} A_{\text{L}}\left(B_{\text{L}} - z\right)^{-n+1} & \text{for } z < -\alpha_{\text{L}}, \\ e^{-\frac{z^2}{2}} & \text{for } -\alpha_{\text{L}} \leq z \leq \alpha_{\text{R}}, \\ A_{\text{R}}\left(B_{\text{R}} - z\right)^{-n+1} & \text{for } z > \alpha_{\text{R}}, \end{cases} \tag{8.3}$$

where z, A_i, and B_i can be expressed as

$$\begin{aligned} z &= \tfrac{x-\mu}{\sigma}, \\ A_i &= \left(\tfrac{n_i}{|\alpha_i|}\right)^n \exp \tfrac{-|\alpha_i|^2}{2}, \\ B_i &= \tfrac{n_i}{|\alpha|} - |\alpha_i|\,. \end{aligned} \tag{8.4}$$

While the core of the double-sided Crystal ball function is a Gaussian distribution $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ with mean μ and width σ, the left and right tails are parametrized by $A_{\mathrm{i}}\left(B_{\mathrm{i}}-z\right)^{-n+1}$ with 4 free parameters. The combinatorial, unmatched or *fake* W-jets, background shape in the pass and fail categories is parametrized by

$$f^{\mathrm{bkg}}\left(x\right)=f_{\mathrm{ErfExp}}\left(x\right)=\frac{1+\mathrm{Erf}\left(\left(x-a\right)/b\right)}{2}e^{cx}, \tag{8.5}$$

where a, b, and c are free parameters of the fit and $\mathrm{Erf}\left(x\right)$ stands for the error function. As an example, Figure 8.7 shows the fits to the soft drop jet mass distribution for MC simulation used for 2017 data. The distributions are fairly well described by the fits.

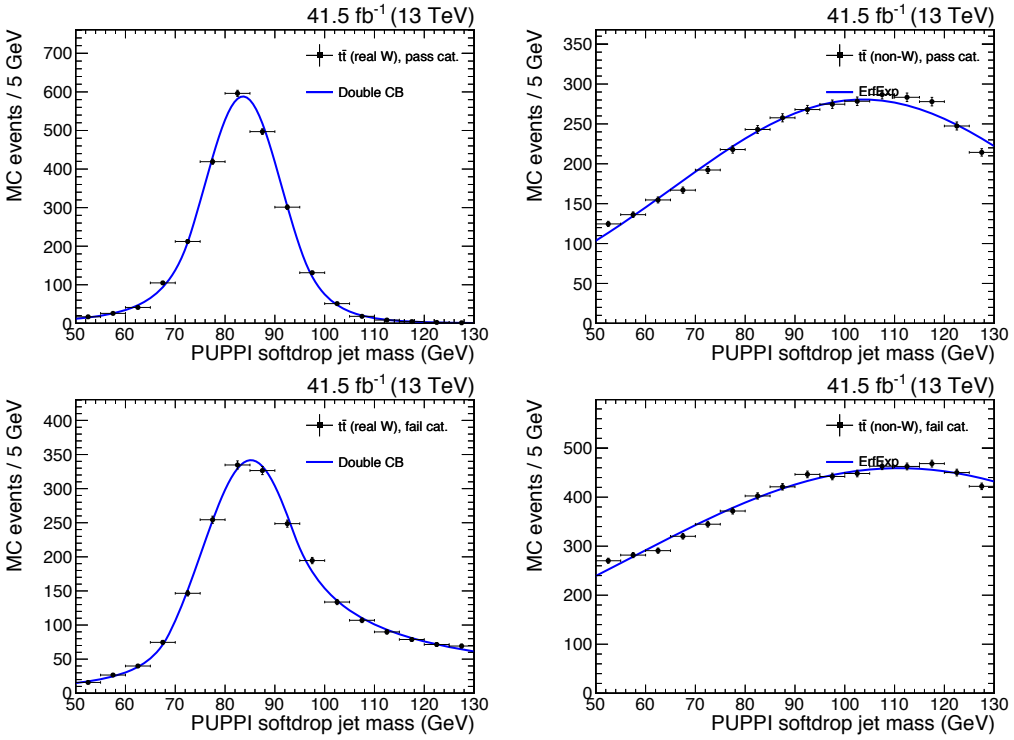

Figure 8.7 – Fits to the soft drop jet mass distributions for $t\bar{t}$ sample: matched to W (right) and background contribution (left) in the pass (upper) and fail (bottom) categories.

8.2.5 Fits to background processes

Second, the background contribution according to MC simulation is fitted. In our case the background processes single top+W, W+multijets and diboson productions are considered. The following fit functions have been found to properly parametrize the background contributions:

$$f^{\text{single-Top+W}}\left(x\right)=f_{\mathrm{ErfExpGaus}}\left(x\right)=\frac{1+\mathrm{Erf}\left(\left(x-a\right)/b\right)}{2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{8.6}$$

$$f^{\mathrm{VV}}\left(x\right)=f_{\mathrm{ExpGaus}}\left(x\right)=e^{ax}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{8.7}$$

$$f^{\mathrm{W+jets}}(x) = f_{\mathrm{ErfExp}}(x) = \frac{1 + \mathrm{Erf}\left((x-a)/b\right)}{2} e^{cx} \tag{8.8}$$

with a, b, c, μ, and σ being free parameters of the fits. Figure 8.8 shows the fits to the background processes (single top, W+jets, VV) for the pass and fail categories.

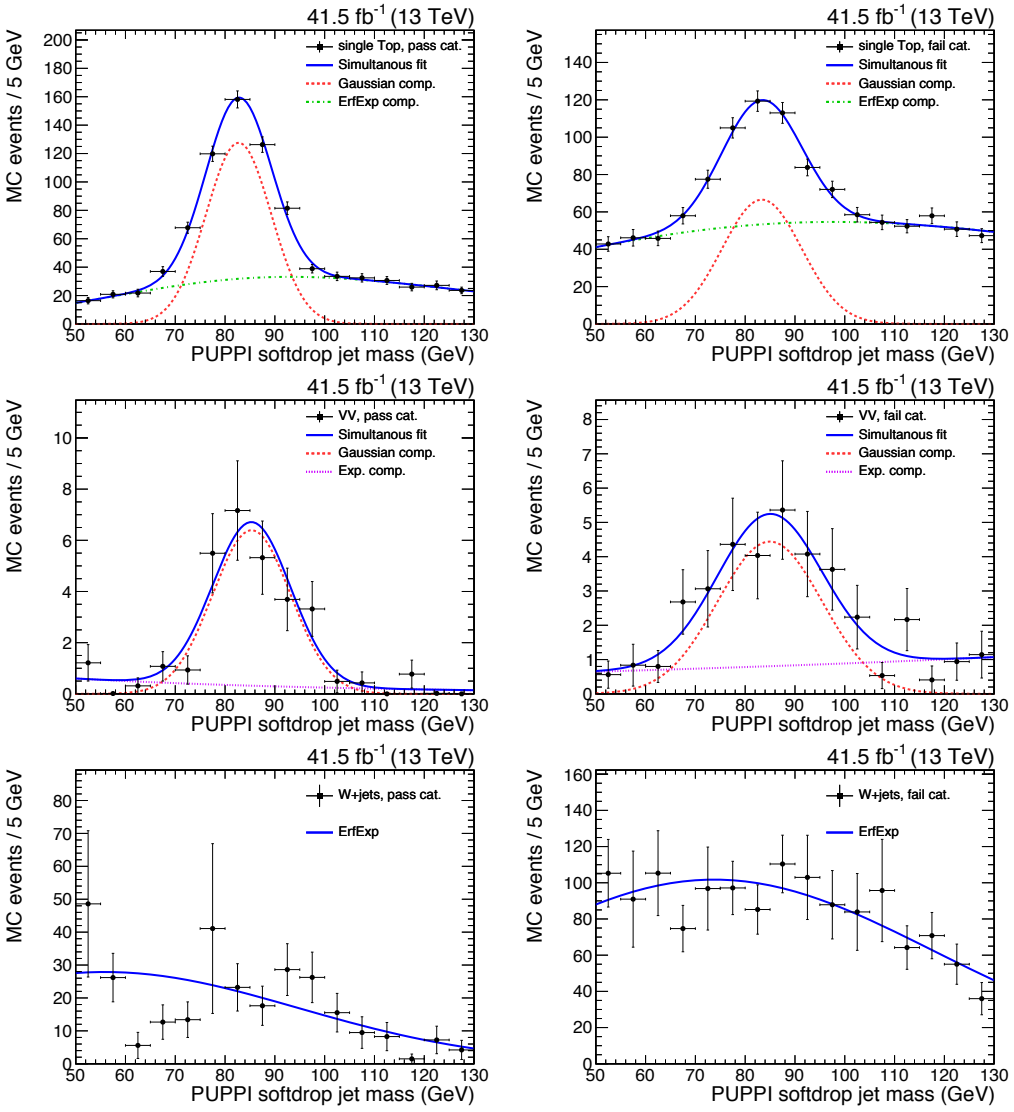

Figure 8.8 – Fits to the soft drop jet mass distributions of the background contributions (single top, VV, W+jets) in the pass (left) and fail (right) categories.

8.2.6 Simultaneous fits

The procedure described in the previous two subsections enables us to choose the fit function to model the soft drop jet mass distribution in the pass and fail categories for various processes. As already mentioned, the W tagging efficiency in data and MC simulations is extracted from a simultaneous fit to the distribution in pass and fail regions. The fit functions are defined as a sum of functions which have been found to be suitable to model signal and background processes by MC simulations. Formally written, fit functions are given as

$$f_{\text{Pass}}(x) = N_{\text{W}}\,\epsilon\, f^{\text{sig}}_{\text{Pass}}(x) + N_2\, f^{\text{bkg}}_{\text{Pass}}(x) + N^{\text{single-Top+W}}_{\text{Pass}} f^{\text{single-Top+W}}_{\text{Pass}}(x) + N^{\text{VV}}_{\text{Pass}} f^{\text{VV}}_{\text{Pass}}(x) + N^{\text{W+jets}}_{\text{Pass}} f^{\text{W+jets}}_{\text{Pass}}(x)\,, \tag{8.9}$$

$$f_{\text{Fail}}(x) = N_{\text{W}}\,(1-\epsilon)\, f^{\text{sig}}_{\text{Fail}}(x) + N_3\, f^{\text{bkg}}_{\text{Fail}}(x) + N^{\text{single-Top+W}}_{\text{Fail}} f^{\text{single-Top+W}}_{\text{Fail}}(x) + N^{\text{VV}}_{\text{Fail}} f^{\text{VV}}_{\text{Fail}}(x) + N^{\text{W+jets}}_{\text{Fail}} f^{\text{W+jets}}_{\text{Fail}}(x)\,. \tag{8.10}$$

ϵ stands for the W tagging efficiency and N_{W} is the total number of the *real* W-jets in the pass and fail region. These two parameters tie the fit functions for the pass and fail regions. The total number of combinatorial background in the pass and fail categories is given by N_2 and N_3, respectively. The free parameters of fits are the parameters of functions $f^{\text{sig}}_{\text{Pass}}(x)\,, f^{\text{bkg}}_{\text{Pass}}, f^{\text{sig}}_{\text{Fail}}(x)$, and $f^{\text{bkg}}_{\text{Fail}}$ of Eq. 8.3-8.8 together with N_{W}, N_2, and N_3. The initial value of these parameters correspond to the value extracted from the individual fits to the MC simulation and they can float within one standard deviation of the uncertainties extracted from the individual fits to the MC samples. The shape and normalization of the minor backgrounds is fixed to the value which are extracted from fits to individual MC simulations. Simultaneous fits to the soft drop jet mass distribution in the pass and fail categories for data and MC simulations are shown in Figure 8.9.

8.2.7 W tagging scale factor

Finally, the W tagging scale factor defined as

$$SF = \frac{\epsilon\,(\text{Data})}{\epsilon\,(\text{MC})}, \tag{8.11}$$

can be determined. Here $\epsilon\,(\text{Data})$ and $\epsilon\,(\text{MC})$ are the extracted W tagging efficiencies in data and MC simulation, respectively. In the main analysis, this scale factor will be applied to the jets in MC simulations and it will correct the difference between measured W tagging efficiency in data and MC.

8.2.8 Jet mass scale and resolution

Furthermore, the width and mean value of the Gaussian function can be utilized to determine the jet mass scale and resolution. In particular, the ratio of means of the double-sided Crystal ball functions fitted to data and MC simulation corresponds to the jet mass scale correction, while the jet mass resolution is defined as a ratio of the widths of the double-sided Crystal ball functions.

The extracted parameters for all three years are summarized in Table 8.1. Only the statistical uncertainties are reported. The W tagging scale factors are close to unity, they vary from 0.96 to 0.99. The jet mass scale is found to be below unity. The values of the jet mass resolution which range from 1.001 to 1.093 indicate better mass resolution in the MC simulation compare to the data.

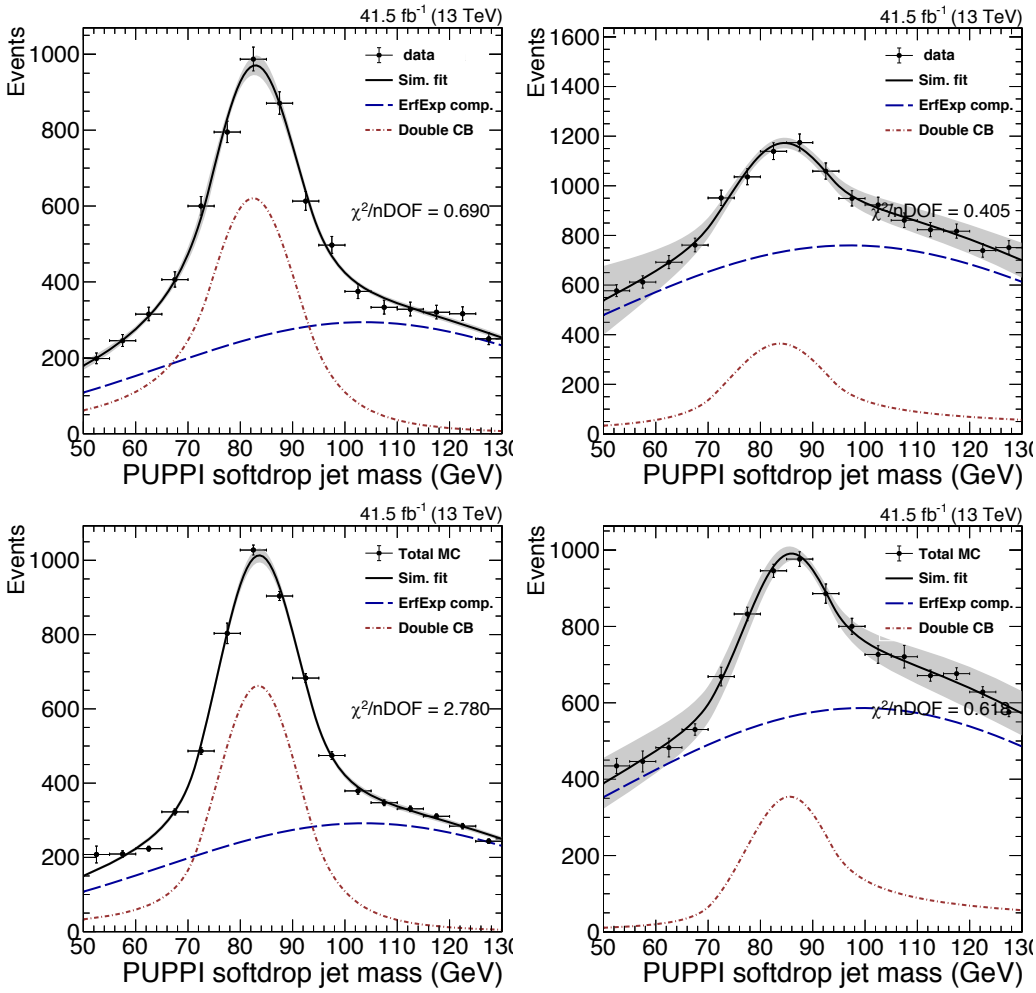

Figure 8.9 – Simultaneous fits to the soft drop jet mass distributions in the pass (left) and fail (right) categories for data (upper) and MC simulation(bottom).

8.2.9 Systematic uncertainties

The derivation of the W tagging scale factor can be sensitive to various effects, that can have a possible impact on the final results. A detailed study of the systematic uncertainties is reported in Ref. [135] and a small impact uncertainties related to JES, JER, PU, b tagging, lepton reconstruction, and $\vec{p}_T^{\,\text{miss}}$ has been found. The most important ones that are also examined in our measurement are the following:

MC simulation

The impact of the ME generator was studied. By default, the ME for $t\bar{t}$ production is calculated by POWHEG. To evaluate the uncertainty, the alternative MC simulation generated by MADGRAPH5_aMC@NLO+FxFx matching is used. This also allows us to estimate the impact of various methods for matching ME with PS. The matching in MADGRAPH5_aMC@NLO is performed by the FxFx method, while POWHEG uses the so-called POWHEG method.

YEAR	$N_2^{1,DDT} < 0$	m/GeV	σ/GeV	W tagging eff.
2016	Data	83.872 ± 0.268	7.827 ± 0.291	0.549 ± 0.025
	Simulation	83.530 ± 0.242	7.817 ± 0.258	0.560 ± 0.023
	Data/simulation	1.004 ± 0.004	1.001 ± 0.050	0.980 ± 0.061
2017	Data	82.473 ± 0.470	8.508 ± 0.607	0.573 ± 0.030
	Simulation	83.435 ± 0.214	7.923 ± 0.209	0.594 ± 0.015
	Data/simulation	0.988 ± 0.006	1.074 ± 0.082	0.964 ± 0.056
2018	Data	82.440 ± 0.282	8.951 ± 0.297	0.553 ± 0.032
	Simulation	83.997 ± 0.234	8.189 ± 0.239	0.555 ± 0.030
	Data/simulation	0.981 ± 0.004	1.093 ± 0.048	0.997 ± 0.079

Table 8.1 – Summary of the extracted jet mass scale, resolution and W tagging efficiency as measured in data and MC samples for 2016, 2017, and 2018 datasets and corresponding data/MC scale factors.

Top p_T spectrum

To take into account the discrepancy of the modelling of the top quark p_T spectrum, the measurement was repeated without applying top p_T reweighting (cf Subsection 7.4.5) to MC simulations.

Modelling of the PS

The impact of the modelling of the PS can be evaluated by comparing the results obtained by using MC simulations with different PS, for example Pythia8 and Herwig++/Herwig7. Different modelling of PS and UE tunes can have impact on the modelling of jet substructure and the performance of the tagger. To evaluate the impact of modelling of the PS, the measurement was repeated with an MC simulation where the parton shower is performed by Herwig++/Herwig7.

8.2.10 Results

In summary, the W tagging scale factors together with the jet mass scale and resolution are listed in Table 8.2 and shown in Figure 8.10. Apart from the statistical uncertainties, the systematic ones are also reported. Overall, the size of the systematic uncertainties is similar to the statistical uncertainty. The dominant systematic uncertainty is related to modelling of PS.

For completeness, let us mention that the scale factor determination was performed inclusively in jet p_T. The possibility of performing the measurement differentially in bins of jet p_T has been investigated. However such a treatment would be statistically limited. In addition, in the higher p_T region typically starting from $p_T \approx 600$ GeV all decay products from the top decay are fully merged into one large cone-size and create top-jet.

Variable	YEAR	Nominal value	stat.	p_T-rew.	generator	PS	stat. $\pm$ syst.
W tag. SF	2016	0.98	0.06	0.001	0.01	0.07	0.06 $\pm$ 0.07
	2017	0.96	0.05	0.008	0.05	0.05	0.05 $\pm$ 0.07
	2018	0.99	0.07	0.02	0.03	0.004	0.07 $\pm$ 0.04
JMS	2016	1.004	0.004	0.001	0.002	0.012	0.004 $\pm$ 0.012
	2017	0.988	0.006	0.001	0.004	0.013	0.006 $\pm$ 0.013
	2018	0.981	0.004	0.001	0.002	0.011	0.004 $\pm$ 0.011
JMR	2016	1.001	0.050	0.001	0.02	0.06	0.05 $\pm$ 0.06
	2017	1.074	0.082	0.007	0.02	0.10	0.08 $\pm$ 0.11
	2018	1.093	0.048	0.005	0.007	0.04	0.05 $\pm$ 0.05

Table 8.2 – Summary of the extracted data/MC scale factors for jet mass scale, resolution, and W tagging efficiency measured for 2016, 2017, and 2018 dataset with corresponding systematics uncertainties.

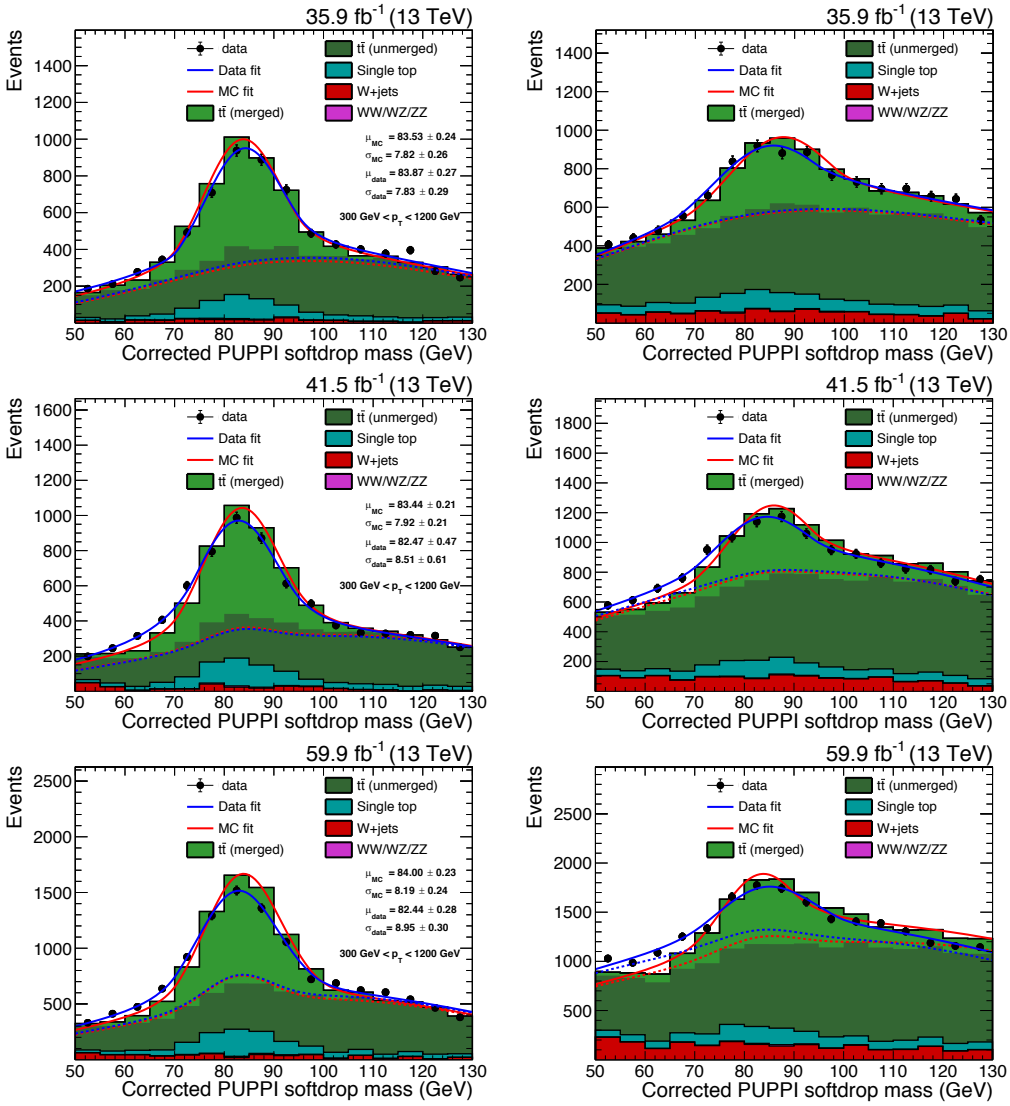

Figure 8.10 – The distribution of soft drop jet mass distribution for $300 < p_{\mathrm{T}} < 1200$ in the pass (left) and fail (right) region for 2016(upper row), 2017(2nd row), and 2018(3rd row) data. Results of the fits to data and simulation are shown by the blue and red lines, respectively. The components of the fit function for background modelling is shown by the dotted line.

9

Physics processes modelling and estimation

In the following section, the modelling of physics processes relevant for the measurement of the cross section of hadronically decaying W and Z bosons will be discussed. Apart from the signal process, the boosted W bosons can also originate from the decay of top quark or from associated production with the single t. This contribution is treated as an irreducible background. The validation of the MC simulation in enriched $t\bar{t}$ topology is essential for the proper subtraction from the measurement. Finally, a data driven method for the estimation of QCD multijet contribution will be introduced. This method utilizes the signal depleted region to predict the shape of the QCD contribution.

9.1 Modelling of $t\bar{t}$ +jets background

The subtraction of the contribution related to $t\bar{t}$ + jets production relies fully on MC simulation. In Subsection 7.4.5, a dedicated correction acting on the modelling of the top quark p_T was discussed. As already stated, this reweighting procedure is derived from the results of the measurement of the $t\bar{t}$ cross section in the dilepton and lepton+jet channel and is valid up to 800 GeV. Here, the extension of this procedure to higher p_T region will be tested. Similarly to the determination of V tagging scale factors, the measurement explores the single muon + jet topology that is enriched by $t\bar{t}$ events to refine the reweighting procedure.

9.1.1 Event topology and selection

This study is performed with the `SingleMuon` datasets and MC simulations listed in Table 7.1 and 7.2, respectively. Only events collected by unprescaled `HLT_Mu50` trigger are further analysed. In contrast to the determination of the V tagging scale factor, looser selection criteria are applied to increase the statistics in the higher p_T regions. The events are required to contain at least one muon with $p_T > 53$ GeV within the tracker acceptance. The presence of an additional electron works as event veto. The final state of this topology consists of two b-jets. Thus, at least one b jet reconstructed in the tracker region with $p_T > 30$ GeV is required. The b tagging is done by the DeepCSV tagger and a jet is considered as b-tagged if passing the medium working point. Finally, only AK8 PF+PUPPI jets with $p_T > 500$ GeV in the tracker region are further considered. The jet is required to be in a back-to-back configuration with the leading muon, i.e.

$\Delta\phi\,(\mu, \text{jet}) > 2$. Following the POG recommendations, all corresponding scale factors are applied to the MC simulation. Figure 9.1 shows control distributions of the basic physics objects kinematics like muon p_{T}, η, and jet p_{T}. Looser selection criteria accompany better statistical precisions. On the other hand, more background processes passed these cuts. The contribution from QCD multijets, Z(ll)+jets, W($l\nu$)+jets, and single top+W production is observed in Figure 9.1, but the main contribution is related to the signal process - the semileptonic decay of $\mathrm{t\bar{t}}$ + jets production.

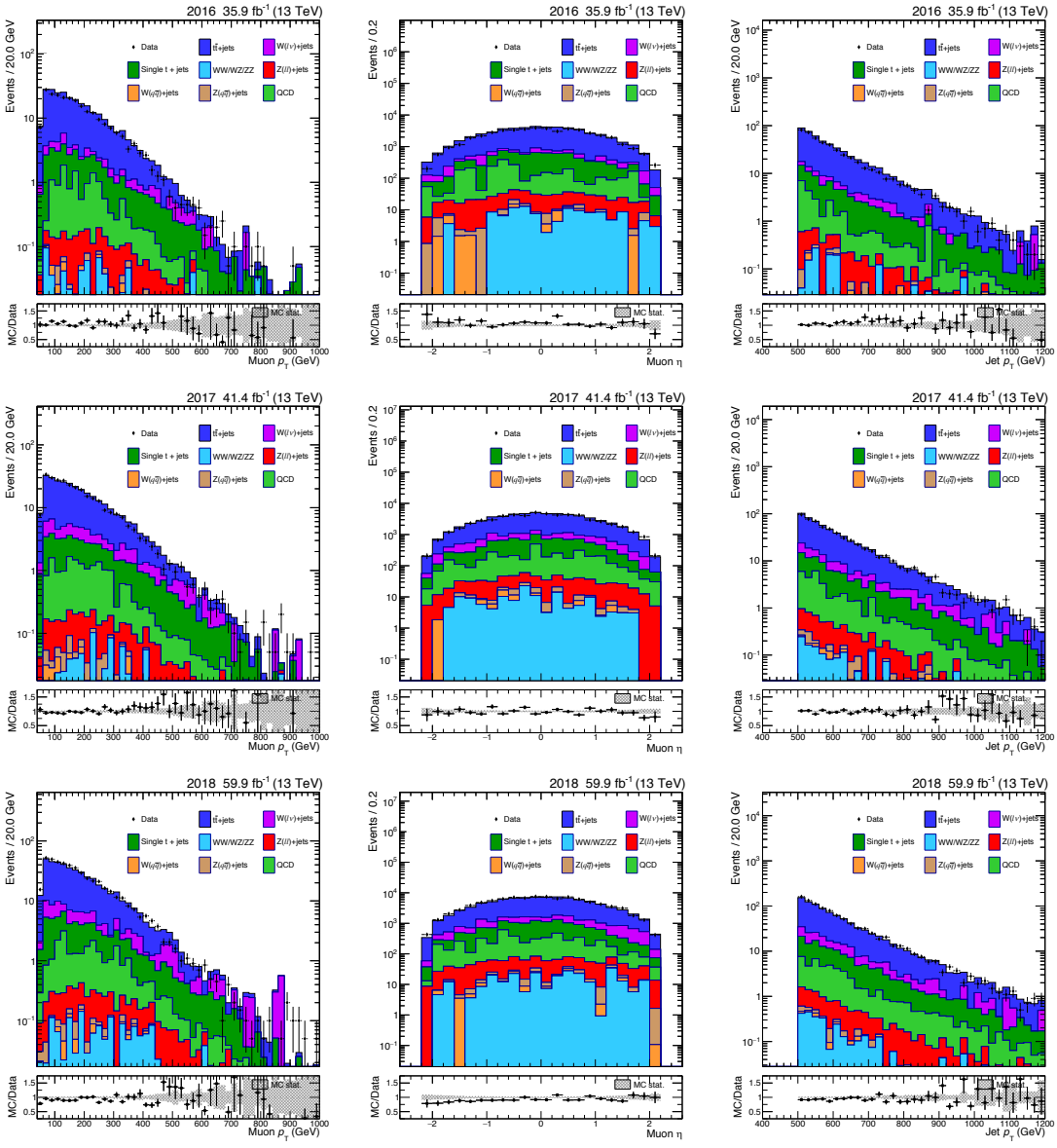

Figure 9.1 – Distributions of the reconstructed muon p_{T}, η and jet p_{T} in single-muon + jet topology after applying selection criteria for 2016 (upper row), 2017 (2nd row), and 2018 (3rd row) data.

The leading large cone-size jets are used to derive the scale factor for $\mathrm{t\bar{t}}$ normalization. Applying the vector boson tagger the jets are grouped into pass ($N_2^{1,\mathrm{DDT}} < 0$) and fail ($N_2^{1,\mathrm{DDT}} > 0$) categories. In the ideal case, the data yield is equal to the MC prediction. Formally written as

$$\#data = \#MC\left(t\bar{t}\right) + \#MC\left(\mathrm{SM}^{\mathrm{bkgrd}}\right), \tag{9.1}$$

where $\#data$ is the observed data yield, $\#MC\left(t\bar{t}\right)$ is the predicted contribution from $t\bar{t}$ production and $\#MC\left(\mathrm{SM}^{\mathrm{bkgrd}}\right)$ stands for the expected background contributions coming from SM.

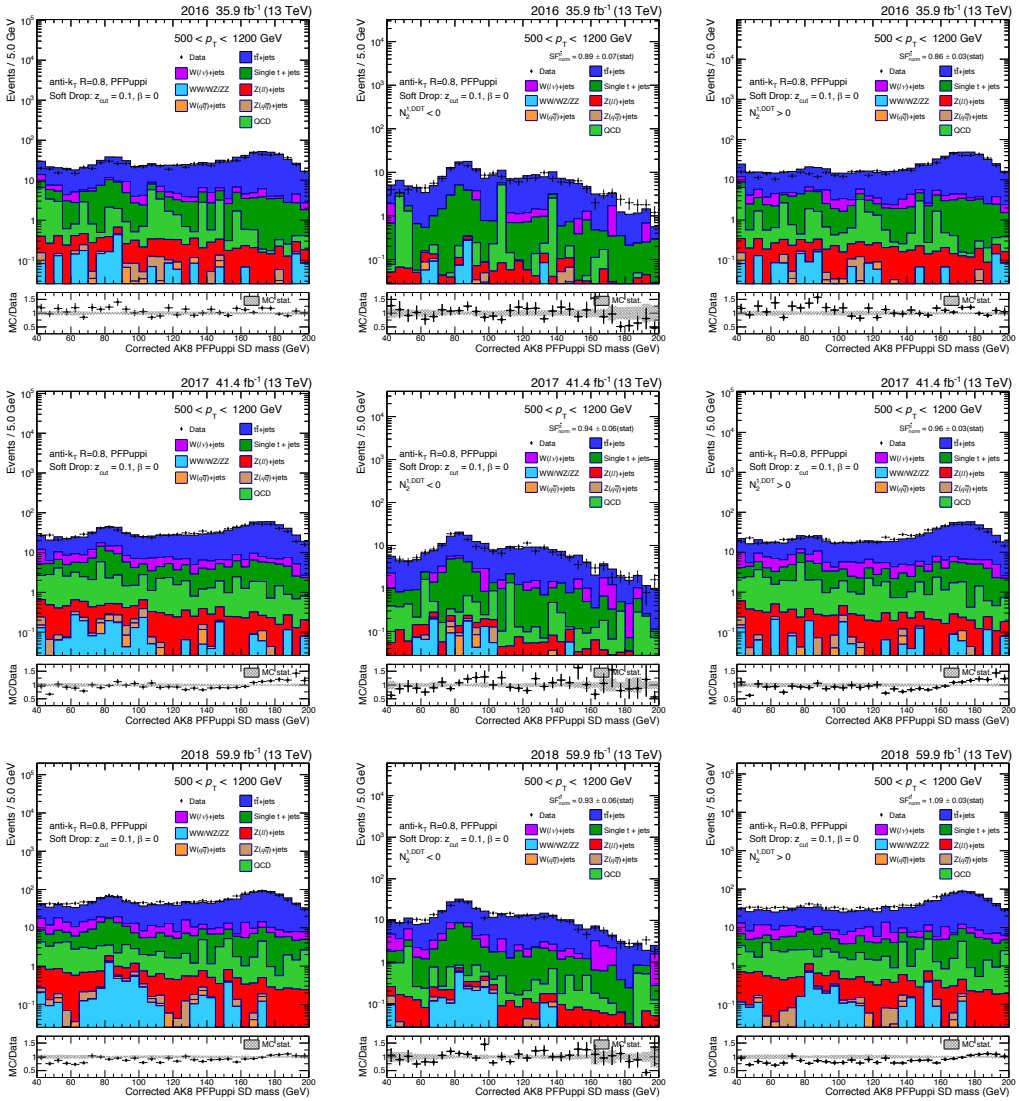

Figure 9.2 – Distributions of soft drop jet mass in single muon + jet topology after applying looser selection criteria. Left: inclusive in $N_2^{1,\mathrm{DDT}}$; Middle: $N_2^{1,\mathrm{DDT}} < 0$; Right: $N_2^{1,\mathrm{DDT}} > 0$.

Since the dominant contribution is from $\mathrm{t\bar{t}}$ production, a possible discrepancy between data and MC simulation will be interpreted as a mismodelling of $\mathrm{t\bar{t}}$ normalization. A possible p_{T} dependence is studied in four p_{T}-bins. For each category, the scale factor for $\mathrm{t\bar{t}}$ normalization is derived as

$$SF_{\mathrm{norm}}^{t\bar{t}} = \frac{\#data - \#MC\left(\mathrm{SM}^{\mathrm{bkgrd}}\right)}{\#MC\left(t\bar{t}\right)}. \tag{9.2}$$

The event yield is counted in the mass window in range from 40 to 200 GeV. The soft drop jet mass distributions are shown in Figure 9.2 inclusively in p_T for three categories: inclusive distribution for $N_2^{1,DDT}$ as well as for the pass ($N_2^{1,DDT} < 0$) and fail ($N_2^{1,DDT} > 0$) regions.

9.1.2 Summary

The determined scale factors are summarized in Table 9.1 and the statistical uncertainties are reported. The scale factors exhibit a weak p_T dependence. A larger difference between the data and MC simulations is observed for higher p_T region, however this region is also limited by the statistical precision. In order to have better handle on the p_T dependence, the linear fit to the scale factors is performed. The nominal value of the fit function in a given bin is then used as scale factor for $t\bar{t}$ normalization.

YEAR	p_T-range / GeV	Pass region	Fail region
2016	500 - 1200	0.89 ± 0.07 (stat)	0.86 ± 0.03 (stat)
	500 - 575	0.94 ± 0.10 (stat)	0.90 ± 0.05 (stat)
	575 - 650	0.88 ± 0.16 (stat)	0.89 ± 0.06 (stat)
	650 - 800	0.73 ± 0.13 (stat)	0.80 ± 0.07 (stat)
	800 - 1200	0.97 ± 0.31 (stat)	0.62 ± 0.13 (stat)
2017	500 - 1200	0.95 ± 0.06 (stat)	0.96 ± 0.03 (stat)
	500 - 575	0.96 ± 0.09 (stat)	0.93 ± 0.05 (stat)
	575 - 650	0.94 ± 0.11 (stat)	0.97 ± 0.06 (stat)
	650 - 800	0.88 ± 0.15 (stat)	1.01 ± 0.08 (stat)
	800 - 1200	0.82 ± 0.25 (stat)	1.05 ± 0.15 (stat)
2018	500 - 1200	0.93 ± 0.06 (stat)	1.09 ± 0.03 (stat)
	500 - 575	0.96 ± 0.08 (stat)	1.09 ± 0.04 (stat)
	575 - 650	0.94 ± 0.11 (stat)	1.11 ± 0.07 (stat)
	650 - 800	0.87 ± 0.13 (stat)	1.06 ± 0.07 (stat)
	800 - 1200	0.73 ± 0.21 (stat)	1.11 ± 0.13 (stat)

Table 9.1 – Summary of the extracted scale factors for $t\bar{t}$ in the pass and fail region measured for 2016, 2017, and 2018 datasets.

9.2 Modelling of V+jets process

The sample of jets in MC simulation of signal processes is a mixture of jets coming from the hadronic decay of W and Z bosons and jets initiated by quarks and gluons. Those jets are produced as a recoil object against the vector boson to balance the system in the transverse plane. Thus, the jets are further categorized by imposing a ΔR matching requirement. The jets for which ΔR between them and the generated vector boson is smaller than 0.4 (half of jet cone size) are labelled as V-matched and grouped in the first category. The other group contains non-V-matched jets. In case that more than one jet in an event is found to be V-matched, then only the jet with lower ΔR is kept as V-matched. This condition guarantees the presence of at most one V-matched jet at the reconstructed level. As an example of this procedure in Figure 9.4 the soft drop jet mass distributions for V-tagged and non-V-tagged jets are shown. The V-matched jets have a clear peak around the actual mass of the W and Z bosons, while the soft drop jet mass for non-V-matched jets is a falling distribution without any peak structure. Only V-matched jets will be used to construct the MC template in order to extract the cross section from data. The

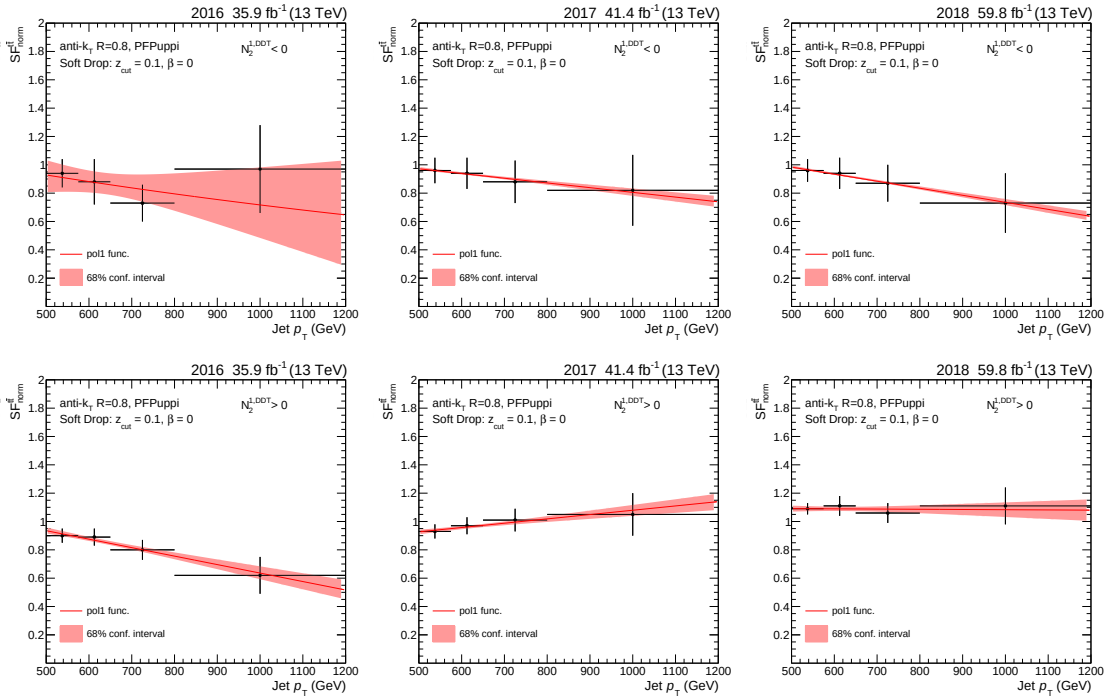

Figure 9.3 – Linear fits to the measured scale factors for $t\bar{t}$ in pass (upper row) and fail region (bottom row) measured for 2016, 2017, and 2018 datasets.

contribution from non-V-matched jets will be included into the fit of the QCD multijet contribution.

9.3 Estimation of QCD multijet contribution

9.3.1 Methodology

Because of the very large cross-section, the QCD multijet is the dominant background contribution. In contrast to the irreducible background, the subtraction of this background processes will rely on a data driven method. Its objective is to predict the shape and the event yield from the signal depleted region. From a wide variety of data driven methods, the alphabet method with a 2D transfer function is most suitable for the measurement. It makes use of the control and side band regions to predict the contribution of QCD multijet in the signal region. The control region, called the pass region, is obtained by inverting the V tagging condition. It means all jets for which $N_2^{1,\mathrm{DDT}} > 0$ make up this region. It is expected that the main contribution in this region is from QCD multijet processes. Since the V-tagger is mass decorrelated, the shape of soft drop jet mass distribution in the fail and pass region should agree. A possible difference is then mimicked by the transfer function. A schematic sketch of background subtraction is shown in Figure 9.5.

In the fail region, the distribution of the soft drop jet mass does not rely on the MC simulation, but is predicted by the data driven technique. The QCD multijet contribution in given m_{SD} and p_{T} bins is defined as

$$QCD^{fail}\left(m_{\mathrm{SD}},p_{\mathrm{T}}\right) = Data^{fail}\left(m_{\mathrm{SD}},p_{\mathrm{T}}\right) - MC^{fail}_{signal}\left(m_{\mathrm{SD}},p_{\mathrm{T}}\right), \tag{9.3}$$

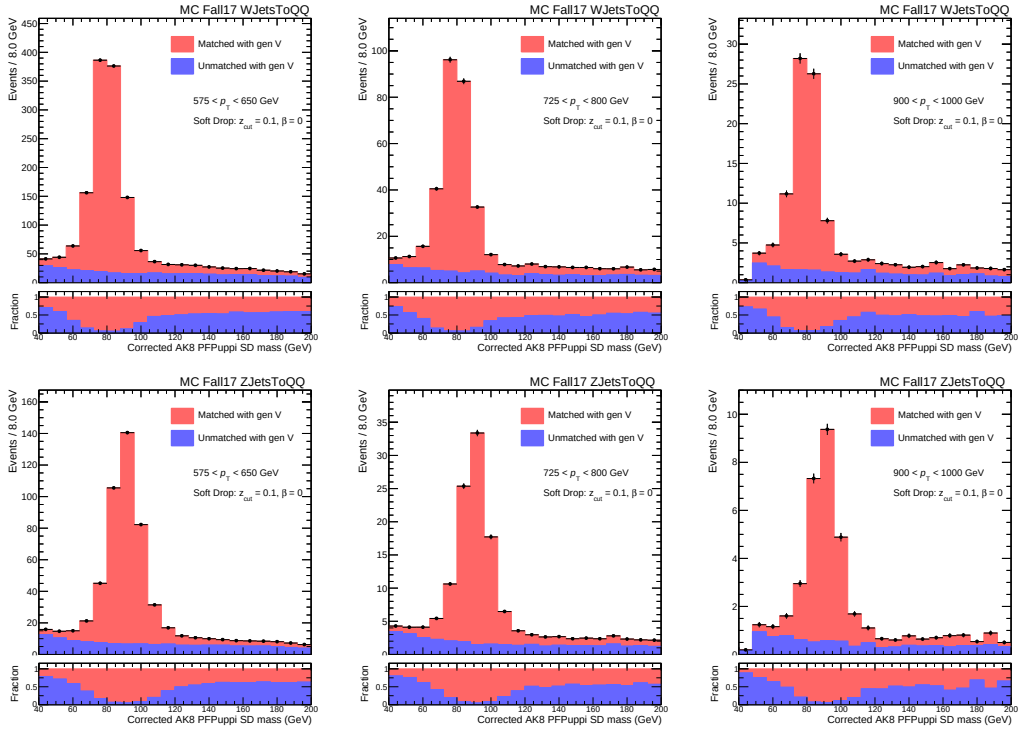

Figure 9.4 – The composition of soft drop jet mass distribution for signal MC simulation.

where $Data^{fail}\,(m_{\mathrm{SD}},p_{\mathrm{T}})$ is the observed data yield, and $MC^{fail}_{signal}\,(m_{\mathrm{SD}},p_{\mathrm{T}})$ is the predicted yield for the contribution coming from MC simulation of the signal processes and $\mathrm{t\bar{t}}$ production in these given bins. Then, the QCD multijet contribution in the signal region can be predicted as

$$QCD^{pass}\,(m_{\mathrm{SD}},p_{\mathrm{T}}) = QCD^{fail}\,(m_{\mathrm{SD}},p_{\mathrm{T}}) \cdot TF\,(m_{\mathrm{SD}},p_{\mathrm{T}})\,, \tag{9.4}$$

where $TF\,(m_{\mathrm{SD}},p_{\mathrm{T}})$ stands for a 2D transfer function. In the ideal case when MC simulations are perfect $TF\,(m_{\mathrm{SD}},p_{\mathrm{T}})$ is the constant function equal to $\alpha = \epsilon^{\mathrm{pass}}(QCD)/\epsilon^{\mathrm{fail}}(QCD)$. Since we tuned the V-tagger to have constant background rejection (85%) and accept only 15% of QCD multijet, we should expect $\alpha = 0.15/0.85 = 0.1765$. The difference in modelling of jet p_{T}, soft drop jet mass, and $N_2^{1,\mathrm{DDT}}$ results in a non-constant transfer function.

9.3.2 Transfer function

The 2D transfer function depending on the jet p_{T} and dimensionless scaling variables ρ is defined by Bernstein polynomials. The $n+1$ Bernstein basis polynomials of degree n are defined as

$$b_{\nu,n}\,(x) = \binom{n}{\nu} x^{\nu}\,(1-x)^{n-\nu}\,, \tag{9.5}$$

for $\nu = 0,\ldots,n$, and $\binom{n}{\nu}$ is a binomial coefficient. As example, $b_{1,2}\,(x) = \binom{2}{1}x^1(1-x)^{2-1} = \frac{2!}{1!(2-1)!}x(1-x) = 2x(1-x)$. The Bernstein basis polynomials of degree n form a basis

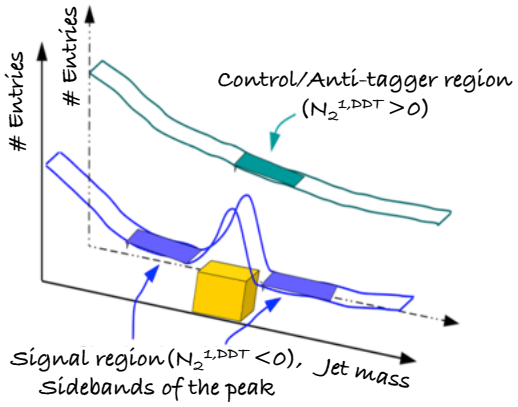

Figure 9.5 – Sketch of background subtraction.

for the vector space of polynomials of degree of at most n with real coefficients. Formally written, a Bernstein polynomial of degree n is defined as a linear combination of Bernstein basis polynomials

$$B_n(x) = \sum_{\nu=0}^{n} \beta_\nu b_{\nu,n}(x) = \sum_{\nu=0}^{n} \beta_\nu \binom{n}{\nu} x^\nu (1-x)^{n-\nu}, \tag{9.6}$$

where β_ν are real coefficients, usually called as Bernstein or Bézier coefficients [136, 137]. Then we can define the transfer function as

$$TF(p_{\mathrm{T}},\rho) = \alpha \sum_{i=0}^{n_\rho} \sum_{j=0}^{n_{p_{\mathrm{T}}}} B_i(\rho) B_j(p_{\mathrm{T}}) = \alpha \sum_{i=0}^{n_\rho} \sum_{j=0}^{n_{p_{\mathrm{T}}}} \beta_i \beta_j b_{i,n_\rho}(\rho) b_{j,n_{p_{\mathrm{T}}}}(p_{\mathrm{T}}). \tag{9.7}$$

In addition, let define real numbers $a_{ij} = \beta_i \beta_j$ with condition $a_{00} = 1$. With this condition, the transfer function reads

$$TF(p_{\mathrm{T}},\rho) = \alpha \sum_{i=0}^{n_\rho} \sum_{j=0}^{n_{p_{\mathrm{T}}}} a_{ij} b_{i,n_\rho}(\rho) b_{j,n_{p_{\mathrm{T}}}}(p_{\mathrm{T}}), \tag{9.8}$$

where α is a free parameter. In case, that data will prefer the constant transfer function $TF(p_{\mathrm{T}},\rho)$, then the free parameter of the transfer function will be only α. The main advantage of the Bernstein polynomials is the fact that the fitted parameters a_{ij} will have a value of similar order of magnitude. Since the Bernstein polynomials are restricted to the interval $\langle 0,1 \rangle$, we need to perform a linear transformation [1] of jet p_{T} and ρ in the $\langle 0,1 \rangle$ interval, when we want to evaluate the transfer function. The implementation of this procedure is done within the COMBINE tool [138] and the distribution of soft drop jet mass for QCD multijet is stored in RooParametricHist [139]. The transfer function ties the pass and fail regions. The QCD yield is determined when performing a maximum likelihood fit to data to determine the signal contribution. Its initial value is given by Eq. 9.3 and is freely floating in the fit.

1. Let $x \in \langle x_{min}, x_{max} \rangle$ where $x_{min} < x_{max}$, then the linear transformation in the $\langle 0,1 \rangle$ interval is given by $(x - x_{min})/(x_{max} - x_{min})$.

9.3.3 Likelihood function

The parameters of interest are determined by means of a maximum likelihood fit to the differential distribution of the soft drop jet mass in bins of jet p_{T} for the pass and fail regions, where systematic uncertainties are treated as nuisance parameters. The expected number of events ν_i in the i^{th} bin can be written as

$$\nu_i = \sum_k s_i^k \left(\sigma^k, \vec{\lambda}\right) + \sum_j b_{i,j} \left(\omega_j, \vec{\lambda}\right), \tag{9.9}$$

where s_i^k is the expected number of signal events from the k^{th} generator level bin, and depends on the cross section σ^k of this bin, and the nuisance parameters $\vec{\lambda}$. The $b_{i,j}$ is the amount of expected events coming from the background process j and depends on the nuisance parameters and on the background normalization parameters ω_j. The number of signal events can be related as

$$s^k = \epsilon^k \mathcal{A}^k \mathcal{L} \sigma^k, \tag{9.10}$$

where ϵ^k is the reconstruction and V tagging efficiency, $\mathcal{A}$ stands for the acceptance, and $\mathcal{L}$ is the integrated luminosity. The acceptance $\mathcal{A}^k$ is defined as the fraction of V-tagged jets in the k^{th} generator level bin, that enter the measured phase space. The $\mathcal{A}^k$ and ϵ^k are determined from the signal MC simulation and in the maximum likelihood fit they can vary according to the parameters $\vec{\lambda}$. In order to construct the likelihood function, the observed number of events in each bin of any observable is assumed to follow a Poisson distribution, i.e. $\mathrm{Poisson\,(x|y)} = \frac{y^x e^{-y}}{x!}$. Then the likelihood function is written as a product of two likelihood functions, one for the pass region and the other one for the fail region. In our particular case the likelihood function is written as

$$\mathcal{L}\left(data|\vec{r},\vec{\lambda},\vec{\omega}\right) = \prod_{i,j} \frac{e^{-\nu_{\mathrm{Pass},i,j}} \nu_{\mathrm{Pass},i,j}^{n_{\mathrm{Pass},i,j}}}{n_{\mathrm{Pass},i,j}!} \prod_{i,j} \frac{e^{-\nu_{\mathrm{Fail},i,j}} \nu_{\mathrm{Fail},i,j}^{n_{\mathrm{Fail},i,j}}}{n_{\mathrm{Fail},i,j}!} \prod_m \pi_m\left(\lambda_m\right) \prod_n \pi_n\left(\omega_n\right), \tag{9.11}$$

where $n_{\mathrm{Pass},i,j}$ and $n_{\mathrm{Fail},i,j}$ are the observed number of events in the i^{th} bins of the soft drop jet mass distribution and in the j^{th} bins of the jet p_{T} distribution for the pass and fail regions, respectively. The $\pi_m\left(\lambda_m\right)$ and $\pi_n\left(\omega_n\right)$ are the Gaussian priors for the nuisance and normalization parameters, respectively. The width of the Gaussian prior is equal to the size of the input uncertainty. The parameters of interest in our measurement are the signal strength parameters defined as

$$r_i = \frac{\sigma_i^{\mathrm{Meas}}}{\sigma_i^{\mathrm{Pred}}}, \tag{9.12}$$

where σ_i^{Meas} stands for the measured cross section, while the predicted cross section given by MC prediction is represented by σ_i^{Pred}. The likelihood function $\mathcal{L}$ is maximized (in practice, we minimize $-2\ln\mathcal{L}$) with respect to the $\vec{r}$,$\vec{\lambda}$, and $\vec{\omega}$. This allows to constrain systematic uncertainties in the fit and reduce the resulting uncertainty in $\vec{r}$. Then the measured cross section in the i^{th} bin is reported as $\sigma_i^{\mathrm{Meas}} = r_i \cdot \sigma_i^{\mathrm{Pred}}$.

9.3.4 Polynomial order of transfer function

In the previous two subsections, the parametrization of the transfer function and construction of the likelihood function was discussed in general. With the likelihood function at hand, the remaining open question about the order of the polynomials for the background subtraction can be solved. The number of parameters of the function is decided through a Fisher's F-test [140]. The F-test is a statistical test in which the test statistic has an F-distribution [141] under the null hypothesis. For two given probability distribution functions (pdf), pdf_1 and pdf_2, let pdf_1 be "nested" within pdf_2. Moreover, pdf_2 must have $N+1$ free parameters of the fit, while pdf_1 has N free parameters of the fit. Hence it is obvious, that $ndf_1 > ndf_2$, where ndf is the number of degrees of freedom for pdf_1 and pdf_2, respectively. In addition, pdf_2 can always fit the data at least as well as pdf_1. The F-test determines, whether the pdf_2 gives a significantly better description or not and is calculated as

$$F = \frac{\frac{\chi_1^2 - \chi_2^2}{ndf_1 - ndf_2}}{\frac{\chi_2^2}{ndf_2}}. \tag{9.13}$$

The function F follows an F-distribution [141] with $(ndf_1 - ndf_2, ndf_2)$ degrees of freedom. As null hypothesis, it is considered that pdf_2 does not provide a significantly better fit than pdf_1. Then the null hypothesis is rejected if F calculated from the data is greater than the critical value of the F-distribution for a false-rejection probability, commonly p-value $<$ 0.05 is used. Since in case of maximum-likelihood fit, we cannot use χ^2 as goodness-of-fit, in Ref. [142], the saturated model goodness-of-fit was introduced and the following relation has been found:

$$\chi^2 = -2 \ln \lambda, \tag{9.14}$$

where λ stands for

$$\lambda = \prod_i \exp\left((d_i - f_i)^2 / 2\sigma_i^2 \right). \tag{9.15}$$

Here $d_i \pm \sigma_i$ is the data with ± 1 standard deviation measured in the i^{th} bin, and f_i is the model prediction for the i^{th} bin. This relation allows us to use Eq. 9.13 for the maximum-likelihood fit. The saturated model goodness-of-fit will be also used to test the quality of fits.

In the following, as example of this method, the polynomial order of the transfer function for 2017 data will be discussed in greater detail. As default parametrization of the transfer function, a linear dependence on ρ, and p_{T}, i.e. $(n_\rho = 1, n_{p_{\mathrm{T}}} = 1)$, is assumed. Alternative parametrizations have at least one more parameter than the default one. It means, the model with $(n_\rho = 2, n_{p_{\mathrm{T}}} = 1)$, $(n_\rho = 1, n_{p_{\mathrm{T}}} = 2)$, and $(n_\rho = 2, n_{p_{\mathrm{T}}} = 2)$ have been tested against the default parametrization. Data have been fitted by all these four functions to calculate the observed value of the F-distribution. Then the best fits by the alternative models are used to generate pseudo-experiment background distributions. These toy data are utilzed to perform the F-test. The red line in Figures 9.6, 9.7, and 9.8 stands for the expected F-distribution with the corresponding degrees of freedom. One can see that the F-test measured with the pseudo-data follow the theoretical expectations.

In case of 2017 data, the calculated p-value for the F-test with $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$ and $(n_\rho = 2, n_{p_\mathrm{T}} = 1)$ is smaller than 5% as shown in Figure 9.6. As new default parametrization, the transfer function with $(n_\rho = 2, n_{p_\mathrm{T}} = 1)$ is choosen because of the larger value of the observed F. As shown in Figure 9.7, 9.8, and 9.9, the F-test was repeated until the parametrization with $n+1$ parameters does not significantly fit better (p-value $< 5\%$) the toy data than the default parametrization with n parameters. The general strategy was always following: for default parametrization $(n_\rho = X, n_{p_\mathrm{T}} = Y)$, the alternative models with one extra parameter, i.e. $(n_\rho = X+1, n_{p_\mathrm{T}} = Y)$ and $(n_\rho = X, n_{p_\mathrm{T}} = Y+1)$, are examined. It enables to systematically proceed and test alternative models with more parameters.

The F-tests had to be performed for each year separately. According to the results from the F-tests, the parametrization of the transfer function with $(n_\rho = 2, n_{p_\mathrm{T}} = 1)$ is preferred for 2016, $(n_\rho = 5, n_{p_\mathrm{T}} = 2)$ for 2017, and $(n_\rho = 4, n_{p_\mathrm{T}} = 2)$ for 2018. The possible explanation of different polynomial order of the transfern fuction for 2016 in contrast to 2017 and 2018 can be related to the choice of UE tune for MC simulations. It was observed that CUETP8M1 tune used for 2016 MC simulation describes data better than newer CP5 tune.

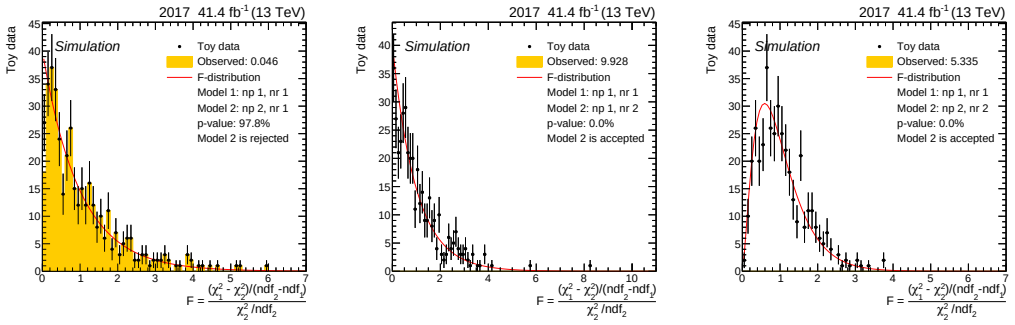

Figure 9.6 – F-test for 2017 data, where the default parametrization $(n_\rho = 1, n_{p_\mathrm{T}} = 1)$ of transfer function is tested against $(n_\rho = 1, n_{p_\mathrm{T}} = 2)$ (left), $(n_\rho = 2, n_{p_\mathrm{T}} = 1)$ (middle), and $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$(right).

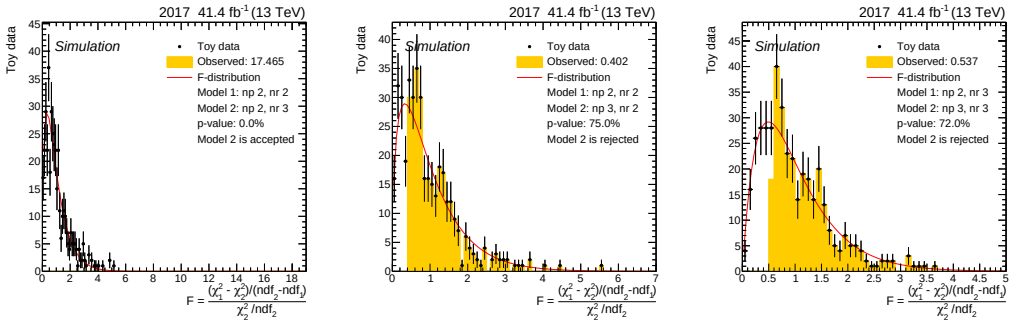

Figure 9.7 – F-test for 2017 data, where the default parametrization $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$ of transfer function is tested against $(n_\rho = 3, n_{p_\mathrm{T}} = 2)$ (left), and $(n_\rho = 2, n_{p_\mathrm{T}} = 3)$ (middle). F-test for $(n_\rho = 3, n_{p_\mathrm{T}} = 2)$ against $(n_\rho = 3, n_{p_\mathrm{T}} = 3)$ is shown on the right.

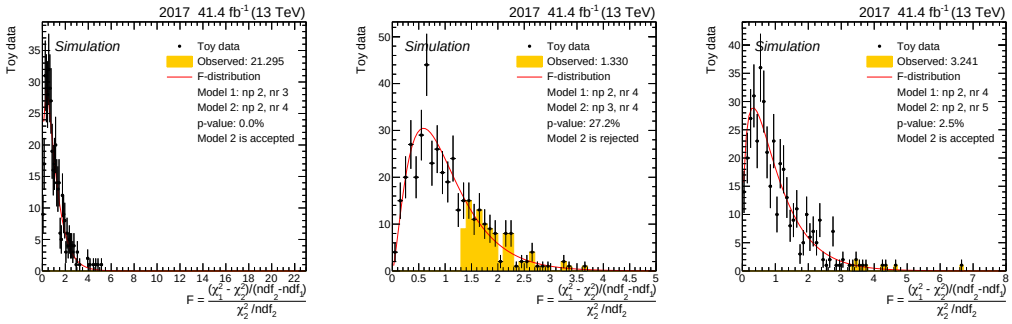

Figure 9.8 – F-test for 2017 data for $(n_\rho = 3, n_{p_\mathrm{T}} = 2)$ against $(n_\rho = 4, n_{p_\mathrm{T}} = 2)$ (left), $(n_\rho = 4, n_{p_\mathrm{T}} = 2)$ against $(n_\rho = 4, n_{p_\mathrm{T}} = 3)$ (middle), and $(n_\rho = 4, n_{p_\mathrm{T}} = 2)$ against $(n_\rho = 5, n_{p_\mathrm{T}} = 2)$ (right).

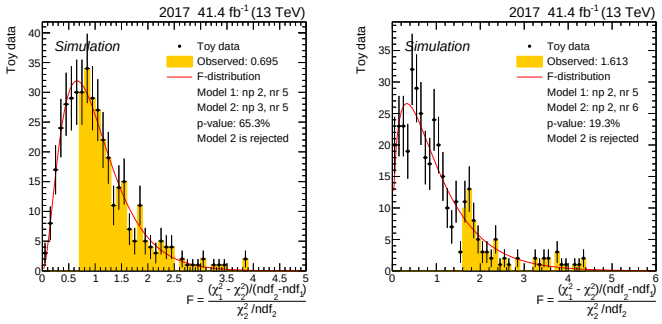

Figure 9.9 – F-test for 2017 data for $(n_\rho = 5, n_{p_\mathrm{T}} = 2)$ against $(n_\rho = 5, n_{p_\mathrm{T}} = 3)$ (left), and $(n_\rho = 6, n_{p_\mathrm{T}} = 2)$ (right).

9.3.5 Fit validation

Several tests have been performed to understand the impact of the choice of the parametrisation and the QCD multijet subtraction method on the measurement. The aim of these checks is to evaluate a possible bias arising from the parametrization of the QCD multijet. All these tests need to be performed for each dataset separately.

9.3.5.1 Closure test

First, the closure test of the fitting procedure was examined. Pseudo-experiment data have been generated with known injected signal strength. Each toy data can be expressed as

$$toy\ data = \mu \cdot S + B, \tag{9.16}$$

where S and B are the signal and background models, respectively. The signal strength is μ. For an unbiased measurement, by performing a fit to the pseudo-experiment data the extracted signal strength $(\hat{\mu} \pm \hat{\sigma})$ should be equal to the injected signal strength within the uncertainties. This can be quantified by the pull defined as

$$pull = \frac{\hat{\mu} - \mu}{\hat{\sigma}}. \tag{9.17}$$

The pull is expected to follow a Gaussian distribution with mean $\mu' = 0$ and width $\sigma' = 1$. In our case, the closure test of the fitting procedure was performed for three scenarios. The first one is background only measurement corresponding to $\mu = 0$. Two other tests have been performed for $\mu = 1$ and $\mu = 2$.

Figure 9.10, 9.11, and 9.12 show the results on the closure test of the fitting procedure. The pulls follow the Gaussian distribution with $\mu' \approx 0$ and $\sigma' \approx 1$. The results indicate no statistically significant bias on the extracted parameters from the fits related to the background subtraction.

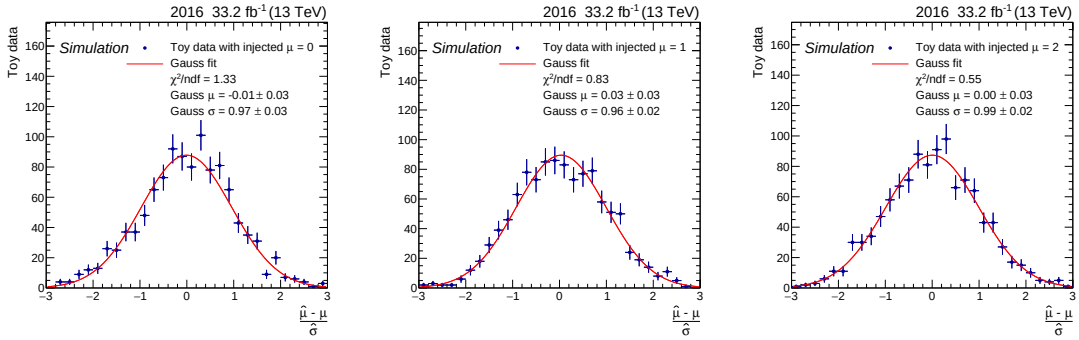

Figure 9.10 – The validation of background subtraction method by the bias test for 2016 data. The measured pulls for three scenario of the injected signal strength: $\mu = 0$ (left), $\mu = 1$ (middle), and $\mu = 2$ (right).

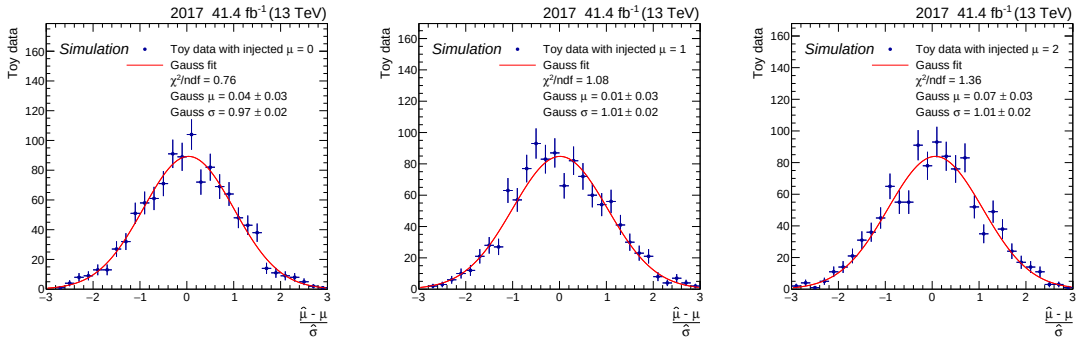

Figure 9.11 – The validation of background subtraction method by the bias test for 2017 data. The measured pulls for three scenario of the injected signal strength: $\mu = 0$ (left), $\mu = 1$ (middle), and $\mu = 2$ (right).

9.3.5.2 Alternative polynomial basis for fit parametrization

In addition, the bias test related to the choice of the family of functions to parametrize the transfer function was investigated. The motivation for the Bernstein polynomials as a default parametrization of the transfer function is rather practical. Since the fitted parameters have a value of similar order of magnitude, the best fit function is being typically found faster. The alternative family of functions to parametrize the transfer function should lead to the same results. Namely, the monomial polynomials and the exponential function with Bernstein polynomials have been considered as an alternative function. The general prescriptions of the transfer function in these basis are the following:

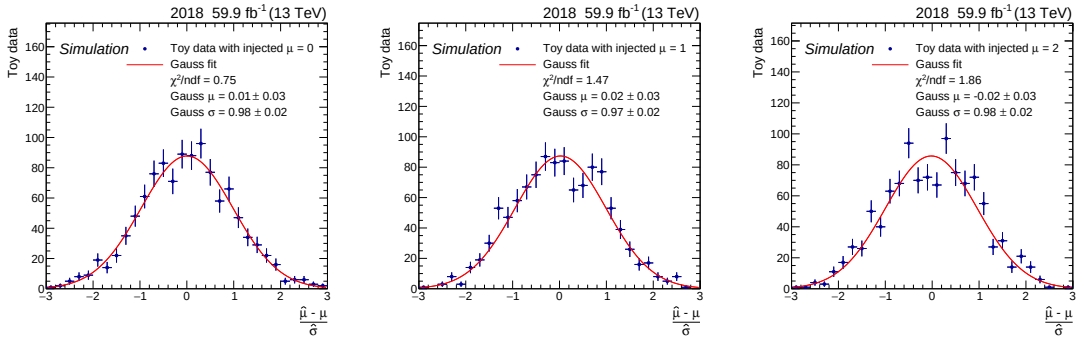

Figure 9.12 – The validation of background subtraction method by the bias test for 2018 data. The measured pulls for three scenario of the injected signal strength: $\mu = 0$ (left), $\mu = 1$ (middle), and $\mu = 2$ (right).

$$TF\left(p_{\mathrm{T}},\rho\right) = \alpha \sum_{i=0}^{n_{\rho}} \sum_{j=0}^{n_{p_{\mathrm{T}}}} a_{ij} \rho^{i} p_{\mathrm{T}}^{j}, \tag{9.18}$$

$$TF\left(p_{\mathrm{T}},\rho\right) = \alpha \sum_{i=0}^{n_{\rho}} \sum_{j=0}^{n_{p_{\mathrm{T}}}} \exp\left(a_{ij} b_{i,n_{\rho}}\left(\rho\right) b_{j,n_{p_{\mathrm{T}}}}\left(p_{\mathrm{T}}\right)\right). \tag{9.19}$$

The main idea behind this test is to estimate possible bias arising from using a different family of functions. This study requires to repeat the F-test for the alternative family of functions and then to generate pseudo-experiment data by this transfer function and afterward use the default transfer function to fit them. If the measurement is independent from the choice of the transfer function, the extracted signal strength from pseudo-experiments are equal to the injected one.

As an example of this study, the results for 2018 data sets are shown. Similarly to the closure test of the fitting procedure, three scenarios has been considered. Namely background only (μ=0) and the scenario with the injected signal strength μ=1 and μ=2. Figure 9.13 shows the results for the case, when the pseudo-experiment data are generated by the transfer function parametrized by the exponential function with the Bernstein polynomials and Figure 9.14 shows the pseudo-experiment data generated by the monomial function. The possible bias of order of 0.1σ and 0.2σ is seen in the first case for the injected signal strength $\mu = 1$ and $\mu = 2$, respectively. No significant bias is observed in latter case as shown in Figure 9.14. As the bias seems to be negligible, no additional uncertainty was assigned. Similar conclusions is true also 2016 and 2017 data sets.

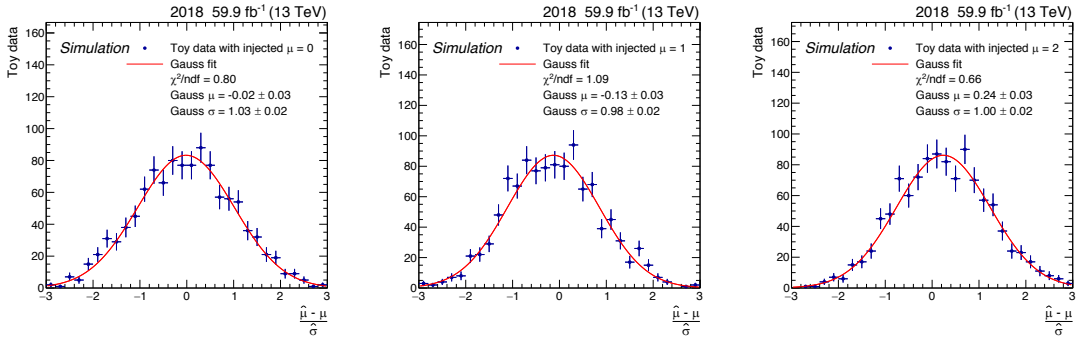

Figure 9.13 – The bias test for 2018 data. The pseudo-experiment data are generated by the transfer function parametrized by the exponential function with the Bernstein polynomials and afterwards fitted by the default transfer function. The measured pulls for three scenario of the injected signal strength: $\mu = 0$ (left), $\mu = 1$ (middle), and $\mu = 2$ (right).

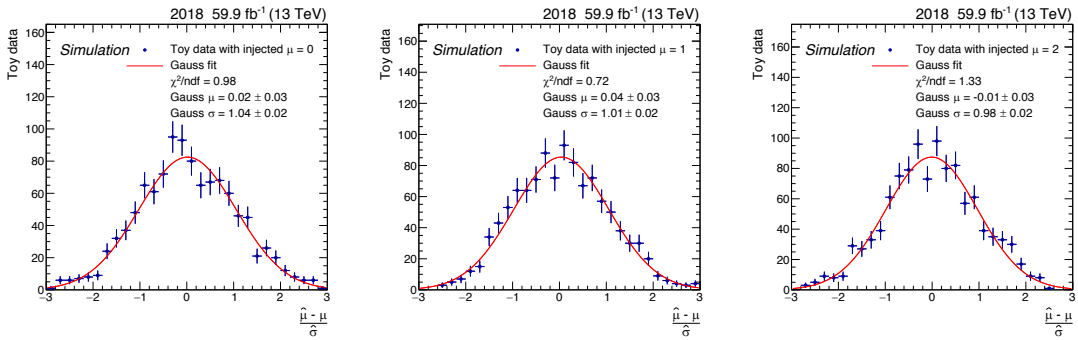

Figure 9.14 – The bias test for 2018 data. The pseudo-experiment data are generated by the transfer function parametrized by the monomial polynomials and afterwards fitted by the default transfer function. The measured pulls for three scenario of the injected signal strength: $\mu = 0$ (left), $\mu = 1$ (middle), and $\mu = 2$ (right).

10

Measurement

The measurement of the cross section of W and Z bosons in the hadronic final state is presented in this chapter. The summary of systematic uncertainties is given and their impact on the measurement are discussed. First, the results from the measurement at the detector level are presented. An additional study has been performed to investigate compatibility of results obtained from a subpart of the full Run 2 datasets. As detector effects like the finite momentum resolution and reconstruction inefficiencies have an impact on observables, data unfolding is necessary for a proper comparison with theoretical predictions. The unfolding procedure removes a bias caused by the detector effects and corrects for detector inefficiencies and acceptance. In the last part of the chapter, the likelihood-based unfolding is applied to the measurement. Finally, the unfolded results are compared with the theoretical predictions.

10.1 Systematic uncertainties

In following subsections, all systematic uncertainties which could be relevant for the presented measurement are discussed. The ones which will be found to be relevant for the measurement are then included as nuisance parameters into the likelihood function (Eq. 9.11). There are essentially two types of systematic uncertainties. The first category of systematic uncertainties affect the overall rate of a given physics process. A typical example of this kind of systematic uncertainties is the uncertainty related to the luminosity. The other group of systematic uncertainties have a shape-changing impact. As it will be shown in next subsections, examples of this kind of the uncertainties are the jet mass scale and resolution as they change the shape of measured distribution.

10.1.1 Jet mass scale and resolution & V tagging scale factor

In order to mitigate the measured disagreement in the jet mass scale and resolution between data and MC simulation, the soft drop jet mass in the simulated events needs to be scaled and smeared by JMS and JMR, respectively. By applying JMS and JMR, the position and the width of the peak in the soft drop jet mass distribution should match with the data. Independently, JMS and JMR are varied within the total uncertainty and the impact on the measurement is studied. As one naively expects, by varying JMS within one standard deviation of the total uncertainties, the mean value of the peak is shifted to the higher, or lower region of the soft drop jet mass distribution. The width of the peak is controlled by the JMR and changing of the JMR will lead to a wider or narrower peak. As an example, Figure 10.1 shows the distribution of the soft drop jet mass in MC

simulation. The distributions of the soft drop jet mass in the pass region are shown for different up/down variation of JMS and JMR in 3 different p_{T} bins. As discussed and now demonstrated, these variations have an impact on the shape of the measured distributions.

Different V tagging performance in data and MC simulation are mimicked by the V tagging scale factor. This factor acts as an additional event weight and changes the overall normalization of the distribution.

In order to perform a combination of RunII data sets, a proper treatment of the correlation of uncertainties is needed. The statistical uncertainty for JMS, JMR, and V tagging efficiency are treated as uncorrelated (0%) across the years as the datasets are statistically independent. The systematic uncertainty for JMS, JMR, and V tagging efficiency were derived by same method, so one can expect that the systematic uncertainties can be fully (100%) correlated.

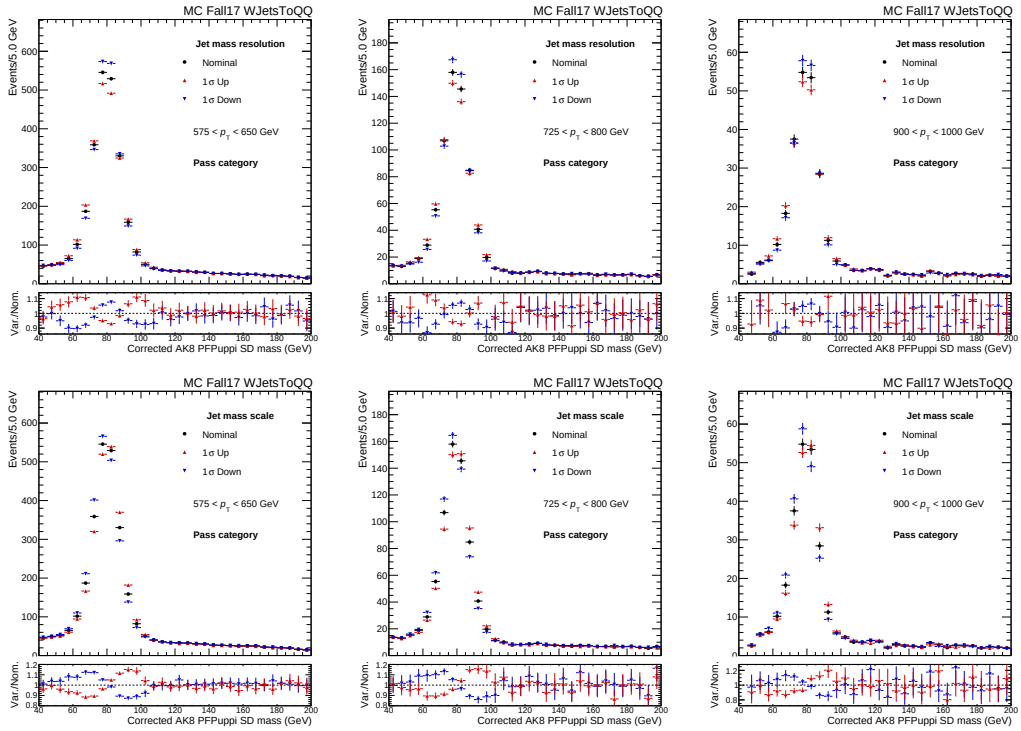

Figure 10.1 – The impact of uncertainties related to the jet mass resolution (upper row) and the jet mass scale (bottom row) on the shape of the soft drop jet mass distribution in the pass region for signal MC simulations for 3 different jet p_{T} bins: $575 < p_{\mathrm{T}} < 650$ GeV, $725 < p_{\mathrm{T}} < 800$ GeV, and $900 < p_{\mathrm{T}} < 1000$ GeV.

10.1.2 Jet energy scale and resolution

The uncertainty related to the jet energy scale is centrally provided by the JME POG. A reduced set of uncertainty sources is used. The common subsources for all three years are: FlavourQCD, RelativeBal, HF, BBEC1, EC2, Absolute. These uncertainties are treated as fully (100%) correlated across the years. Apart from the fully correlated uncertainties, each year has some extra sub-sources which are considered as uncorrelated (0%). The name of these sub-sources are BBEC1_YEAR, EC2_YEAR, Absolute_YEAR, HF_YEAR, RelativeSample_YEAR, where the prefix YEAR stands for 2016, 2017, and 2018 respectively.

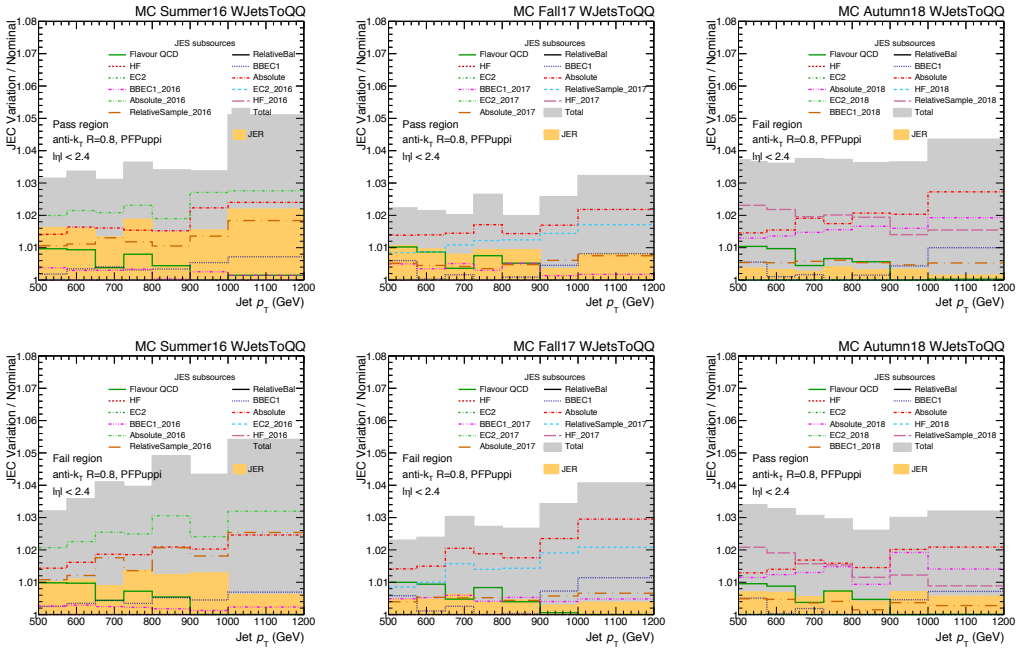

Figure 10.2 – The impact of uncertainties related to the jet energy correction for 2016 (left), 2017 (middle), and 2018 (right) datasets. While the individual sub-sources are represented by the line, the gray band stands for the total uncertainty. The orange band represents uncertainty related to JER SF. The upper row shows the impact of JES in the pass region, the bottom row shows the impact of JES in the fail region. Statistical uncertainties are not shown.

The BB/EC1/EC2/HF in the name of sub-sources of JEC uncertainties is referring to the detector regions, BB stands for the barrel region, i.e. $|\eta| < 1.3$, EC1 corresponds to $1.3 < |\eta| < 2.5$, and EC2 $2.5 < |\eta| < 3.0$. The forward region ($3.0 < |\eta|$) is represented by HF. In order to determine the impact of the jet energy scale uncertainties, for each sub-source we varied the jet p_T and repeated the measurement. Figure 10.2 shows the impact of individual sub-sources of jet energy correction uncertainties for the signal MC simulation. Since the uncertainties for jet energy correction in the central region for high p_T jet are well under control, the typical impact of individual sub-sources is rather small and around 1-3%. As already mentioned, our measurement is performed within the tracker acceptance, i.e. $|\eta| < 2.4$, hence the impact of EC2 and HF sub-sources are equal to 0.

Consistent jet energy resolution between data and MC simulation is achieved by the smearing procedure with JER SF. In order to estimate the impact of JER SF, the measurement needs to be repeated for the up/down variation of the uncertainty for JER SF. The impact of JER SF is represented by the orange band in Figure 10.2.

10.1.3 The $t\bar{t}$ scale factor

The MC simulations of $t\bar{t}$ production are corrected by a scale factor to mitigate the difference between data and MC simulation observed in the modelling of the shape of $t\bar{t}$ spectrum. The measured scale factor with uncertainties for $t\bar{t}$ reweighting in the pass and fail regions are summarized in Table 9.1. The variation of the scale factor within its statistical uncertainties changes the normalization of the $t\bar{t}$ MC simulation.

10.1.4 Pileup reweighting

The total inelastic cross section σ_{inel} is needed for the pileup reweighting. The measured value of the inelastic cross section is known with a precision of 4.6%. The variation of the total inelastic cross section within its uncertainty is used to calculate up/down variation of the pileup profile as shown in Figure 7.1. Finally, the pileup reweighting procedure is repeated with these variations in order to propagate the uncertainties into the measurement. The jet mass distributions for different pileup reweighting procedures differ in the normalization and the impact of the pileup reweighting was found to be overall around 1.2%.

10.1.5 Luminosity

The luminosity uncertainty changes overall the normalization of the observed distributions. These uncertainties with corresponding subsources and their correlation across different years are centrally provided by the Luminosity POG and are summarized in Table 10.1.

YEAR	Uncorr.	Uncorr.	Uncorr.	X-Y factor.	Corr.17-18	Corr.16-17
2016	2.2	0.0	0.0	0.9	0.0	0.8
2017	0.0	2.0	0.0	0.8	0.4	0.6
2018	0.0	0.0	1.5	2.0	0.3	0.0

Table 10.1 – Summary of the luminosity uncertainties and their correlation for different data-taking period as provided by Lumi POG [143].

10.1.6 Level-1 prefiring issue

The Level-1 prefiring correction is applied to MC simulation for the 2016 and 2017 samples. In greater detail, this correction has been discussed in Subsection 7.4.2 where Figure 7.3 shows their impact on the measurement. Following the centrally provided recipe to estimate its uncertainties, a negligible impact is found.

10.1.7 Parton shower

One of the input parameters for the parton shower is the strong coupling α_S at a scale Q^2. By varying the value of α_S within its uncertainties, we can estimate the uncertainty related to the renormalization scale for QCD emissions in ISR and FSR. Three sets of variations separately for ISR and FSR are included in MC simulation for 2016 and 2018 datasets. The differences between those sets are the factors which are used for the variations, namely "reduced" variations with factors $\sqrt{2}$ and $1/\sqrt{2}$, "default" variations with factor 2 and 0.5, and finally the "conservative" variations with factor 4 and 0.25. The presented measurement uses the default variations with factors 2 and 0.5. The impact of the parton shower weights is evaluated via per-event weights. The variations of FSR and ISR result in slightly different shapes of the soft drop jet mass distribution with different normalization.

10.1.8 Lepton veto

In order to suppress contributions from the leptonic decay of W and Z bosons or $t\bar{t}$ production, the presence of at least one muon or electron with loose identification criteria in the event is utilized as an event veto. The event was rejected if it includes at least one

lepton. This cut will be referred as a hard lepton veto. Due to the different performance of the lepton identification and isolation in the data and MC simulations, an alternative approach to veto events with the leptons is explored in order to assign possible bias. To each event, an extra event weight given by

$$\omega = \prod_{i \in \text{leptons}} (1 - \text{SF}_i)\,, \tag{10.1}$$

is applied. The product runs over the number of leptons in the event and SF is the lepton scale factor. In case of muons, SF is the product of the scale factor for the isolation and identification, while for electrons, SF is equal to the scale factor for the cut based ID. It is easy to see, that $\omega = 1$ for events with no leptons. As discussed earlier, the scale factors for leptons are close to unity, so $\omega \to 0$ for events with at least one lepton. It has been found that both methods give consistent results and the observed uncertainty seems to be negligible. The same conclusion is true for the muons.

10.1.9 PDF

To evaluate the impact of PDF uncertainties, the simulated events have been reweighted by the stored 100 variations of the nominal PDF set. The impact have been found to be below 0.5% level and is omitted.

10.2 Measurement at detector level

The signal strengths are extracted from a binned maximum likelihood fit to data, simultaneously over all p_{T} bins in the pass and fail categories. The parameter of interests (POI) are allow to float freely in the fit. The systematic uncertainties included as nuisance parameters are assumed to vary within one standard deviation of inputs a-priori.

As an example, Figure 10.3 shows the fit to data for jets with p_{T} in range from 650 to 725 GeV. Each dataset is shown separately. The shape of soft drop jet mass distribution of QCD multijet production is derived from the fail region (Figure 10.3(left)) and then projected to the signal region by the transfer function. The pass region exhibits a peak structure around 80 GeV corresponding to the contribution coming from the hadronic decay of W and Z bosons. Fits for other p_{T} bins together with the extracted transfer functions are included in Appendixes A, B, C, and D.

The signal strengths $\sigma^{\text{MEAS}}/\sigma^{\text{MC}}$ measured in bins of jet p_{T} are shown in Figure 10.4. Here σ^{MC} stands for the predicted cross section by MadGraph5_aMC@NLO generated at LO QCD accuracy but corrected by the NLO QCD and NLO EW k-factors. The results are reported for each dataset separately as well as for a simultaneous fit to all three data set. No significant tension between results obtained from different dataset is observed. The presented measurement covers a wide range of kinematic region of the transverse momentum from 500 to 1200 GeV. The last two bins close to the TeV scale benefit from the available statistics from full Run 2 combination. The measured signal strengths have been found to be consistent with unity within the total uncertainties in region up to $p_{\text{T}} = 900$ GeV. Despite of using full Run 2 dataset, the last two bins are still statistically challenging and the observed signal yields seems to be underestimated.

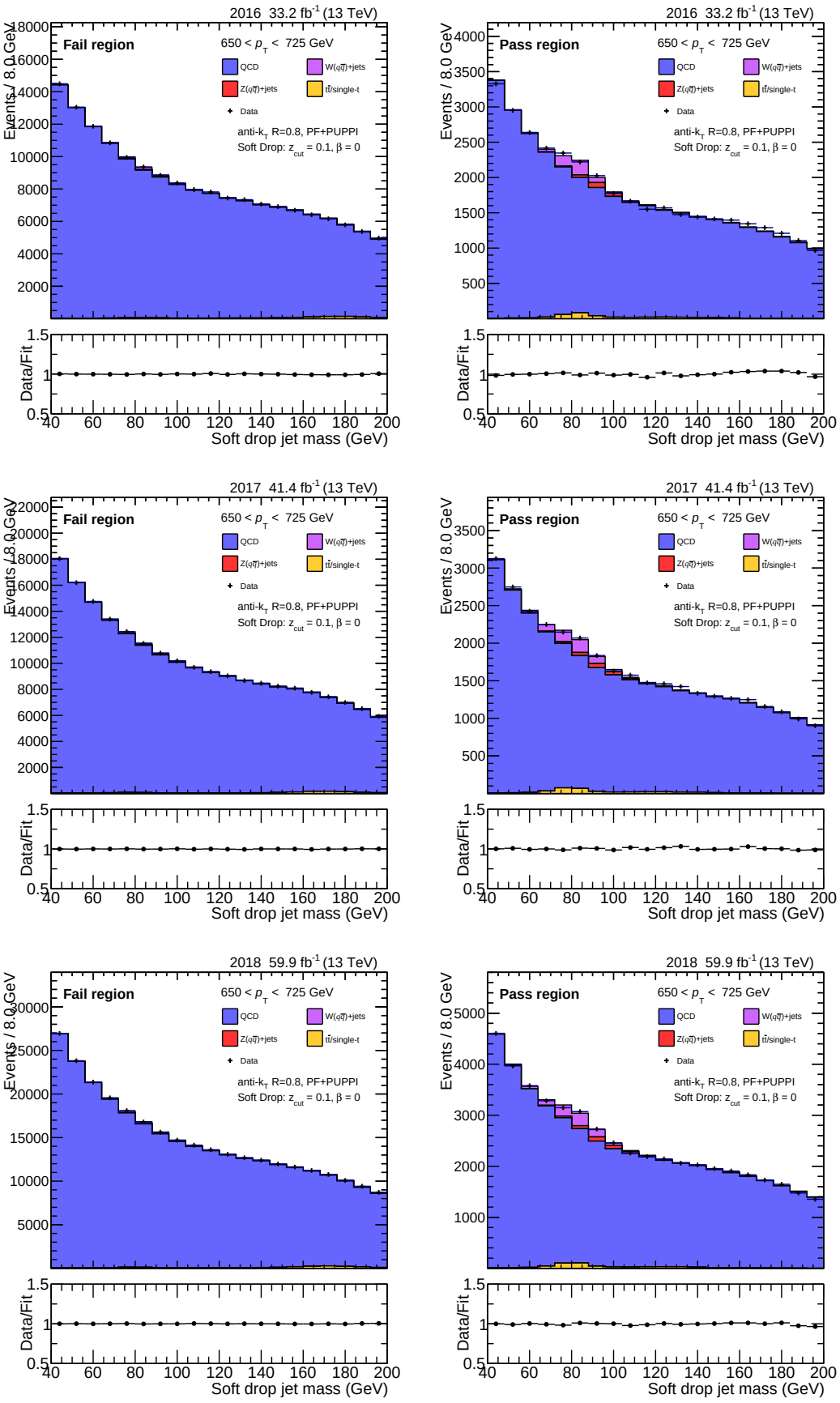

Figure 10.3 – Fits to the soft drop jet mass distribution for p_{T} bin 650 - 725 GeV measured with 2016(upper), 2017(middle), and 2018(bottom) data sets. The fail region is shown on the left and pass region is shown on the right.

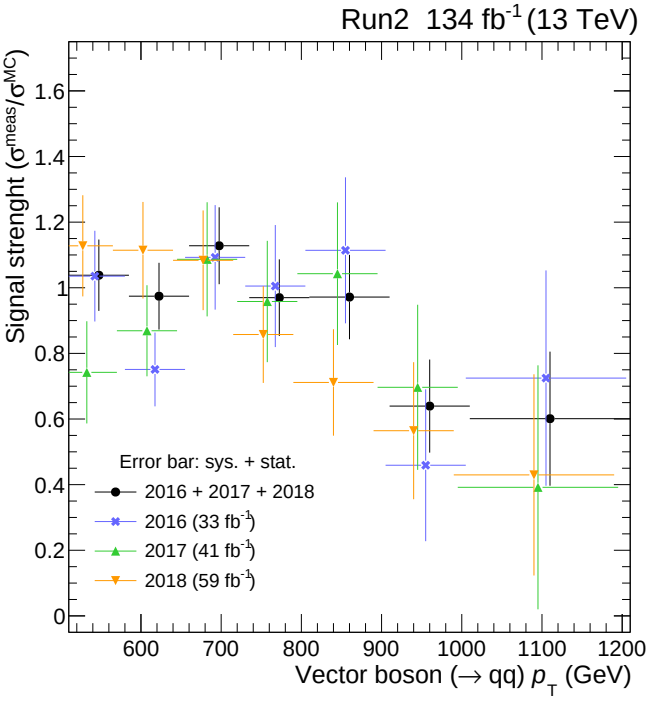

Figure 10.4 – The signal strength $\sigma^{\mathrm{MEAS}}/\sigma^{\mathrm{MC}}$ for different p_{T} bins. Results are shown for each dataset separately as well as for full Run2 data set.

10.3 Data unfolding

The reconstructed distributions are affected by various detector effects. For example, the measured distribution are smeared because of the finite detector resolution and the number of observed reconstructed jets is reduced by the reconstruction efficiency and the detector acceptance. Thus, the results cannot be directly compared to the theoretical prediction without a full detector simulation. For appropriate comparison, an unfolding procedure is needed in order to correct these effects. During the unfolding procedure, the reconstructed distributions are corrected for detector effects and unfolded to stable particle level. Two essential steps are required before the data unfolding can be performed. First, a particle level definition of the jet originating from the hadronic decay of W and Z bosons is needed. Then the migration effects have to be encoded into the response matrix.

10.3.1 Particle level definition

Only stable (decay length $c\tau > 1$ cm) and visible final state particles can be used for the particle level definition of jets coming from the fully hadronic decay of vector bosons. This condition immediately excludes all information about the vector bosons as they are not stable as well as they have very short life-time.

As a proxy for the V-tagged jet, i.e. a jet originating from the hadronic decay of W and Z bosons, we will consider the particle-level jets with the same kinematic properties as the reconstructed jets, i.e. jet $p_{\mathrm{T}} > 500$ GeV and $|\eta| < 2.4$. The V-tagged jet is expected to have a soft drop jet mass in range from 65 to 105 GeV. Furthermore, the requirement on the substructure variable $N_2^1 < 0.28$ needs to be met. As already stated, the high p_{T} vector bosons are produced together with a recoil QCD jet, typically in back-to-back

topology in order to balance the system in the transverse plane and our particle level definition should give us no more than one generator-level V-tagged jet tagged per event. To handle the situation, when two generator-level jets fulfil the conditions listed above, one extra condition is imposed: if two generator-level jets fulfil the conditions listed above, as the V-tagged jet, the generator-level jet with the smaller value of N_2^1 is considered. This particle level definition should lead to a model independent definition.

10.3.2 Response matrix & Migration effects

There are several sources of migration affecting the measured distribution. The migration effects can be grouped into the following two categories.

- **M1** - Signal migration inside the phase space: Due to the finite detector resolution, the reconstructed and generated jet can belong to a different kinematic bin. The variables to quantify this kind of migration are purity and stability. The stability of the i^{th} bin at generator level is defined as

$$s_i = \frac{N\left(\exists\, reco\, jet : reco\, jet \in x_i^{reco} \wedge \exists\, gen\, jet : gen\, jet \in x_i^{gen}\right)}{N\left(\exists\, gen\, jet : gen\, jet \in x_i^{gen}\right)}. \tag{10.2}$$

  The purity of the i^{th} reconstructed bin is defined as

$$p_i = \frac{N\left(\exists\, reco\, jet : reco\, jet \in x_i^{reco} \wedge \exists\, gen\, jet : gen\, jet \in x_i^{gen}\right)}{N\left(\exists\, reco\, jet : reco\, jet \in x_i^{reco}\right)}. \tag{10.3}$$

- **M2** - Migration into the phase space from background & outside the phase space: The cut on the reconstructed jet p_{T} as well as the generated jet p_{T}, and the soft drop jet mass can lead to the situation, when a jet is selected on generator level, but not on reconstructed level and vice versa. In addition, the V tagging efficiency is affecting the migration into/outside the phase space. The acceptance defined as

$$a_i = 1 - \frac{N\left(\exists\, gen\, jet : gen\, jet \in x_i^{gen} \wedge \exists\, reco\, jet\right)}{N\left(\exists\, gen\, jet : gen\, jet \in x_i^{gen}\right)} \tag{10.4}$$

  quantifies how many jets are identified at reconstructed level, if they are selected at generator level. The selected generated jets without corresponding jets at reconstructed level are sometimes labelled as *misses*.
  The background in i^{th} reconstructed bin is defined as

$$b_i = 1 - \frac{N\left(\exists\, reco\, jet : reco\, jet \in x_i^{reco} \wedge \nexists\, gen\, jet\right)}{N\left(\exists\, reco\, jet : reco\, jet \in x_i^{reco}\right)} \tag{10.5}$$

  and illustrates the migration of reconstructed jets without any generated jet into the measured phase space. These jets are sometimes called as *fakes*.

The migration effects can be represented by a 2D response matrix $\mathbb{R}$, with one dimension corresponding to each level - reconstructed level (x-axis), and generated level(y-axis). The migration effect of type M2 are also included in the response matrix. While the underflow bins are filled by the *fakes*, the *misses* contribute to the overflow bins. The probability matrix $\mathbb{P}$ is then obtained from the response matrix by normalising every row of the response matrix to the total number of the generated events in the corresponding generated bin. Formally written, the probability matrix $\mathbb{P}$ is defined as

$$\mathbb{P}_{i,j} = \mathbb{R}_{i,j} \left(\sum_{j=0}^{n} \mathbb{R}_{i,j} \right)^{-1}, \tag{10.6}$$

where $\mathbb{R}_{i,j}$ is i,j-element of the response matrix $\mathbb{R}$ and n stands for the total number of columns of the response matrix $\mathbb{R}$. Since the index j starts at 0, the probability matrix $\mathbb{P}$ carries information about the efficiency and acceptance. The probability matrix $\mathbb{P}$ can be further decomposed into:

$$\mathbb{P}_{i,j} = \epsilon_j \mathbb{P}'_{i,j}, \tag{10.7}$$

where the efficiency ϵ_j describes the probability for an event to be reconstructed and selected at all, and the normalised resolution matrix $\mathbb{P}'_{i,j}$ describes only the bin-to-bin migrations for selected events.

The unfolding procedure relies on the inversion of the probability matrix. The recommended choice of the unfolding method depends on how ill-conditioned the problem is and can be inferred from the condition number of the probability matrix [144]. The condition number $\kappa(\mathbb{A})$ of matrix $\mathbb{A}$ is defined [145] as

$$\kappa(\mathbb{A}) = \frac{\sigma_{\max}(\mathbb{A})}{\sigma_{\min}(\mathbb{A})}, \tag{10.8}$$

where $\sigma_{\max}(\mathbb{A})$ and $\sigma_{\min}(\mathbb{A})$ are the maximal and minimal eigenvalues of matrix $\mathbb{A}$. A condition number closer to unity indicates a more stable matrix inversion [144].

The probability matrices and the migration effects for MC simulations for 2016, 2017, 2018 datasets are shown in Figure 10.5. As confirmed by the condition numbers, the probability matrices for 2017 and 2018 look very similar and the current binning provides us with the stability and purity around 50-80% for 2017, and 2018. A bit lower stability and acceptance observed for MC simulation for 2016 dataset can be related to the different UE tune (CUETP8M1) and the strong coupling α_S used for this simulation. The ΔR matching condition between the generator and reconstructed level slightly increased the background and decreased the acceptance.

10.3.3 Likelihood-based data unfolding

Analysis strategy

The main method for signal extraction and the unfolding used in the analysis is the likelihood-based unfolding [146] using the Combine tool [138, 147]. It means that the signal extraction and the unfolding is done simultaneously. The fit can constrain systematic uncertainties which are profiled in the likelihood. Therefore, the final uncertainties estimated in this way are not only smaller but also well defined since they are derived directly from the maximum likelihood. In addition, the correlation of systematics uncertainties across different years can be done straightforward and properly within the Combine framework.

In order to perform the likelihood based unfolding, the signal MC templates need to be splitted into sub-templates according to the generator level distributions. As an example,

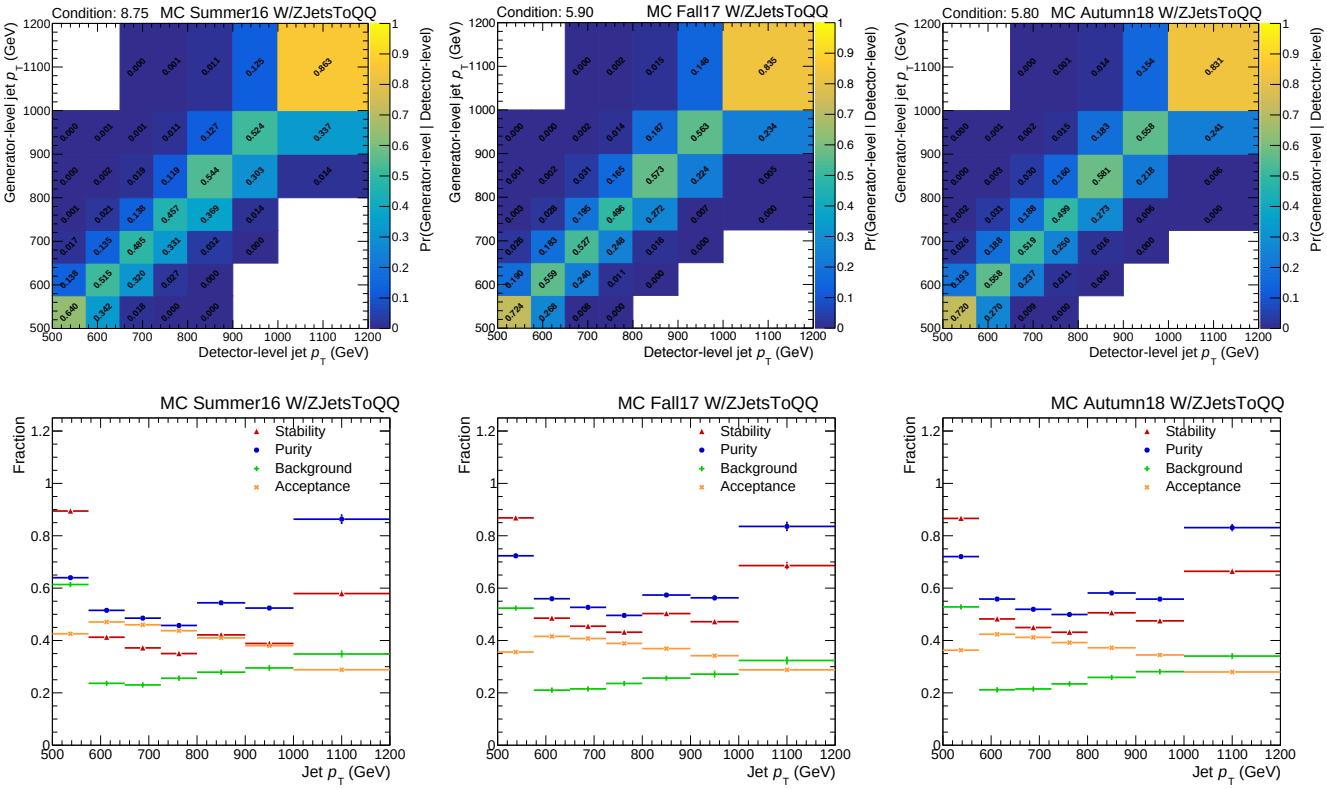

Figure 10.5 – The probability matrices as measured in MC simulation (upper row) for 2016(left), 2017(middle), 2018(right) dataset. The migration and detector effects (stability, purity, acceptance, background) as a function of jet p_{T} measured in MC simulation(bottom row) for 2016(left), 2017(middle), 2018(right) dataset).

Figure 10.5 shows three set of sub-templates for the three reconstructed level. As one can promptly see, these sub-templates include the same information as the response matrices shown in Figure 10.6 - information about the migration effects from one bin to another bin and the background/fakes. Therefore the information about the response (or probability) matrix is directly included in the likelihood function.

Determination of the regularization strength

Figure 10.5 shows that the migration effects are different from the beginning to the end of the measured p_{T} spectrum. This can hint for the need of regularization. The simplest way to introduce regularization in the likelihood based unfolding approach is to apply a penalty term in the likelihood function which depends on the values of the truth bins, formally written as

$$-2\ln\mathcal{L} = -2\ln\mathcal{L} + P\left(\vec{x}\right), \tag{10.9}$$

where P is the linear operator. In this measurement, the SVD approach [148] is used for the regularization. In such case, the linear operator P can be written as

$$P = \tau\left|\mathbf{A}\vec{\mu}\right|^2, \tag{10.10}$$

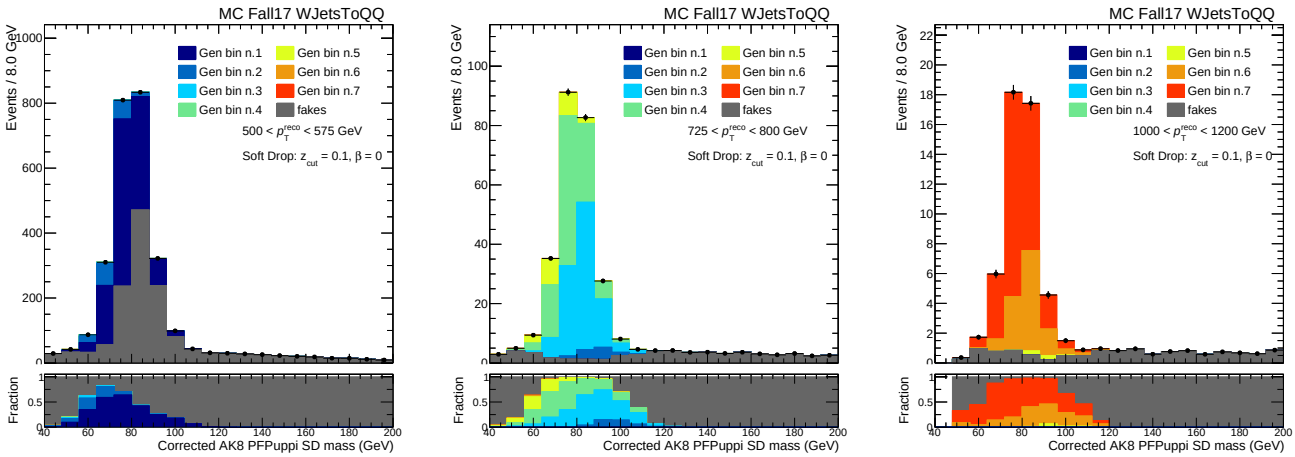

Figure 10.6 – The signal MC template divided into individual subtemplates according the generator level distributions.

where $\mathbf{A}$ is discrete curvature matrix, with

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & \dots & & \\ 1 & -2 & 1 & 0 & \dots & \\ 0 & 1 & -2 & 1 & 0 & \dots \\ \dots & & & & & \end{bmatrix}. \tag{10.11}$$

Then the regularization strength δ is defined as

$$\delta = \frac{1}{\sqrt{\tau}}. \tag{10.12}$$

There are two approaches to determine the regularization strength - the L-curve scan or by the minimisation of the global correlation coefficients. The latter one is utilized as the main method to determine the regularization strength in the presented analysis. The global correlation coefficients are calculated point by point from the Hessian matrix. For the i^{th} POI, the global correlation coefficient is expressed as

$$\rho_i = \sqrt{1 - \frac{1}{\mathbb{V}_{jj}\mathbb{V}^{-1}_{jj}}}, \tag{10.13}$$

where the $\mathbb{V}_{jj}$ are the diagonal elements from the Hessian matrix for the i^{th} POI. Then the mean global correlation coefficient can be defined as

$$\rho_{ave} = \frac{1}{N}\sum_{j=1}^{N}\rho_j, \tag{10.14}$$

where N is the total number of POI and ρ_j stands for the global correlation coefficient. The global minimum of the mean global correlation coefficient as a function of the regularization strength δ corresponds to the best choice of the regularization strength. Figure 10.7 shows

a scan over the mean global correlation coefficient as a function of the regularization strength δ and the optimal regularization strength was found to be $\delta = 0.23$.

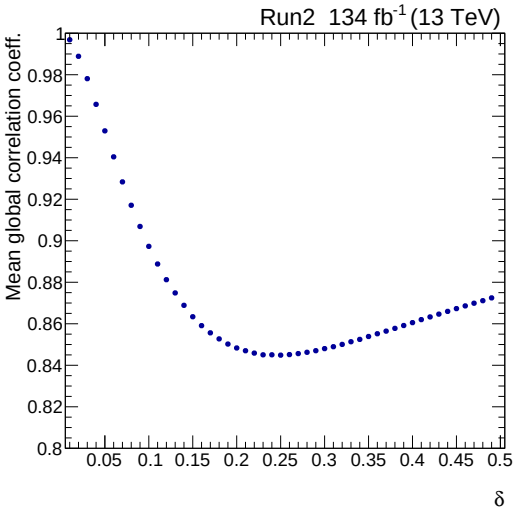

Figure 10.7 – Scan of the mean global correlation coefficient as a function of the regularization strength δ with full Run2 dataset.

10.4 Cross section of W and Z bosons in the hadronic final state

Figure 10.8 shows the results of the first differential measurement of cross section of hadronically decaying W and Z bosons as function of boson p_{T}. The measurement covers wide range of the kinematic region of the transverse momentum p_{T} from 500 to 1200 GeV. The cross section varies by two order of magnitude. The results are compared with the predictions of MadGraph5_aMC@NLO for associated production of W and Z bosons with 1, 2 noncollinear high p_{T} partons supplemented by parton shower and merged with the FxFx procedure. Predictions are obtained with two different different tunes, namely CP5 and CUETP8M1. Overall, the prediction with CP5 UE tune describes data better than the one with CUETP8M1 tune. The former one gives a reasonable agreement with data for $p_{\mathrm{T}} < 700$ GeV. Both predictions predict a harder spectra in the higher p_{T} region compared to data. This observation can be explained by the missing NLO EW corrections, which play an important role at the TeV scale. The typical size of NLO EW corrections have been shown in Figure 7.6.

A summary of the systematic uncertainties affecting the measured cross section of W and Z bosons is given in Table 10.2. The dominant systematic uncertainties are related to the $N_2^{1,\mathrm{DDT}}$ selection efficiency and the jet mass resolution.

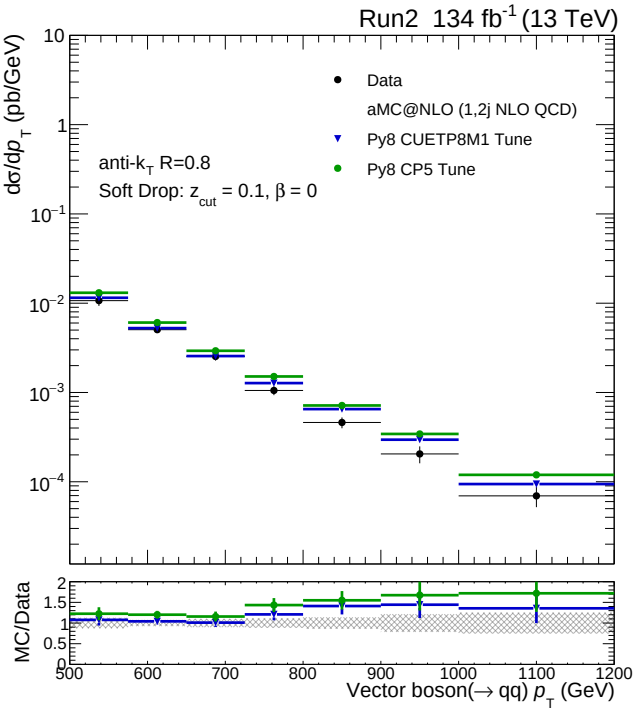

Figure 10.8 – The measured cross section of hadronically decaying vector boson as a function of p_{T} compared with several theoretical predictions.

Uncertainty source	**Impact**
QCD subtraction	2-17%
$N_2^{1,\mathrm{DDT}}$ selection efficiency	8%
Jet mass scale	1%
Jet mass resolution	3%
Jet energy scale	2-5%
Jet energy resolution	1-2%
$\mathrm{t\bar{t}}$ normalization	1-8%
Integrated luminosity	2%
Pileup	1%

Table 10.2 – Summary of the systematic uncertainties affecting the measured cross section of W and Z bosons.

Part III

Conclusions

11

Summary, conclusions, and perspectives

11.1 Summary and conclusions

In this doctoral thesis, the first differential measurement of the cross section of hadronically decaying W and Z bosons with high transverse momentum p_{T} is presented. The measurement is based on data sample from proton-proton collisions at the centre-of-mass energy of 13 TeV collected with the CMS detector at the CERN LHC in the years 2016, 2017, and 2018, corresponding to an integrated luminosity of 134 fb^{-1}. As the vector bosons have very high transverse momentum p_{T}, decay products are collimated in the momentum direction of the mother particle and are merged into a large cone-size jet with a characteristic inner structure. However, jets are also produced by processes in strong interaction described by QCD with a very high rate and the heavy SM resonances sit on top of a huge contribution from QCD. In order to suppress the QCD contribution, advanced analysis techniques are needed for heavy SM resonance identification. For this measurement, a mass decorrelated tagger based on two-prong substructure variable N_2^1 has been developed. The commissioning of the tagger has been performed by utilizing a topology containing semileptonc decays of a $\mathrm{t\bar{t}}$ system. One of the signatures of this complex topology is a large cone-size jets originating from the hadronic decay of W boson.

The main challenge for the measurement of cross section of hadronically decaying W and Z bosons is the background contribution mainly coming from QCD multijet production. Its modelling and subtraction relies on a data-driven technique. Particularly, the alphabet method with a 2D transfer function which makes use of the control region and side band regions to predict the contribution of QCD multijet in the signal region is implemented. A possible bias introduced by the background modelling has been investigated. Finally, using likelihood based unfolding, the cross section is extracted and unfolded to the particle level simultaneously. Thanks to using the full Run 2 data sets collected by the CMS detector, the measurement is able to explore a wide range of kinematic region of the transverse momentum from 500 to 1200 GeV where higher order corrections, namely NLO QCD and NLO EW, in the perturbative calculations play an important role. This measurement studies the SM in the extreme kinematic region, which is hardly studied by the measurement of W and Z bosons in leptonic decay channel. The measured cross section of hadronically decaying W and Z bosons is presented with a full set of systematics uncertainties and the results are compared with the theoretical prediction at NLO QCD accuracy.

11.2 Perspectives

The study of heavy Standard Model particles, namely W, Z, and Higgs bosons and top quarks, with a high Lorentz boost is one of the most popular topics nowadays. It nicely combines advanced analysis methods together with machine learning techniques. The measurement presented in this thesis demonstrated how advanced analysis techniques developed for searches for physics beyond the Standard model can be modified and successfully applied for the classical Standard Model measurement like the cross section measurement of the vector bosons.

Already in Run 3, LHC is believed to delivery three times more luminosity than in Run 2 which will enable us to go beyond the differential cross section measurement. In the following paragraphs, some perspectives and prospects of new measurement within the Standard Model physics which may be of interest are discussed.

Heavy-flavour tagging

The natural next step in the measurement of hadronically decaying W and Z bosons is the cross section measurement only one of those bosons. Experimentally, the separation of W and Z bosons can be achieved by applying heavy flavour tagging as only Z bosons can decay into bottom or charm quark-antiquark pairs, respectively. The dominant background is still the QCD multijet production, but now especially the gluon splitting into two heavy flavour quarks. Apart from the QCD multijet production, the contribution related to the production of a W boson decaying into charm and strange quark will have to be considered.

Such studies can be of particular interest for the Higgs measurements with hadronic final states. Thanks to the high statistics data set collected in Run 2, most of the Higss measurement with all-hadronic final states include a boosted category. An example is the measurement of VH($b\bar{b}$) or VH($c\bar{c}$) [149] where the boosted Z bosons are used as a standard candle of the Standard Model to constrain systematic uncertainties. A main challenge for measurements with boosted vector bosons decaying into heavy flavour quarks will be the determination of the tagging efficiency.

Colour flow and correlation measurement

The associated production of boosted vector bosons and jets results in similar final states like QCD dijet or multijet productions, two high p_{T} jets in back-to-back configuration. However, there is one essential difference. The jets are produced by the strong interaction. As quarks and gluons carry colour charge that is conserved, there is colour-connection between the initial and final state partons. Figure 11.1 shows Feynman diagrams of quark-gluon vertices together the colour charge flow.

Figure 11.1 – QCD quark-gluon vertices: Feynman diagram are represented by black lines, while the coloured lines illustrate colour charge flow [150].

In contrast, vector bosons are colourless objects and therefore there is no colour-connection between vector boson and the associated jets. Since the colour connection influences the radiation pattern, one may naively expect different correlation in events with dijet

production and associated production of vector boson with jets. Thus, the correlation measurements can serve as a probe to explore the colour-flow in these events.

Even much more data are expected to be collected during High-Luminosity LHC [151, 152]. In total, data sample from proton-proton collisions at centre-of-mass energy of 14 TeV, corresponding to an integrated luminosity of 3,000 fb^{-1} is being expected. Then, the hadronic decay of W and Z bosons will serve as great experimental laboratory for QCD studies.

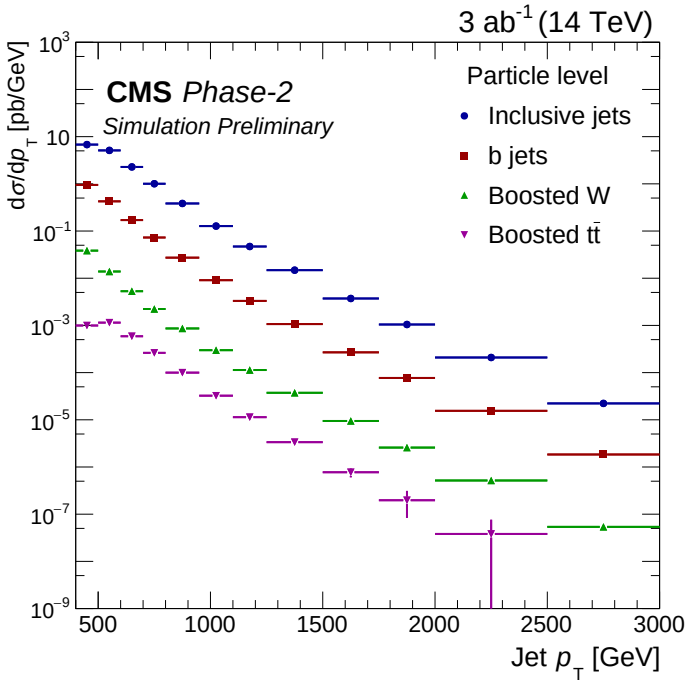

Figure 11.2 – The projection of the particle-level jet cross sections for large cone-size jets as a function of jet p_{T} as being expected for HL-LHC. The cross section of W production does not include statistical uncertainties corrected for efficiencies and background subtraction. Published in Ref. [152].

List of publications

The results presented in this thesis are included in the following publications:

- CMS Collaboration, High-p_{T} jet measurements at the HL-LHC, CMS Physics Analysis Summaries, CMS-PAS-FTR-18-032
- Standard Model Physics at the HL-LHC and HE-LHC, CERN Yellow Rep. Monogr. 7 (2019), 1-220
- CMS Collaboration, Jet energy scale and resolution performance with 13 TeV data collected by CMS in 2016-2018, CMS Detector Performance Note, CMS-DP-2020-019

In parallel of the work presented in this thesis, the author of this thesis contributed in the phenomenology study of the transverse momentum dependent PDF. The results are published in the following publications:

- N.A. Abdulov *et al.*, TMDlib2 and TMDplotter: a platform for 3D hadron structure studies, Submitted for publication in Eur.Phys.J.C. arXiv:2103.09741
- S. Baranov *et al.*, CASCADE3: A Monte Carlo event generator based on TMDs, Eur. Phys.J.C. 81 (2021) 425, arXiv:2101.10211
- J.Lidrych *et al.*, Application of parton branching TMDs to Z-boson production, PoS ICHEP2020 (2021) 526
- A. Bermudez *et al.*, The transverse momentum spectrum of low mass Drell-Yan production at next-to-leading order in the parton branching method, Eur.Phys.J.C 80 (2020) 7, 598, arXiv:2001.06488.
- A. Bermudez *et al.*, Production of Z-bosons in the parton branching method, Phys.Rev.D 100 (2019) 7, 074027, arXiv:1906.00919.

Appendices

A

Fits to 2016 dataset

In this appendix, fits to the soft drop jet mass distribution for 2016 data set are shown.

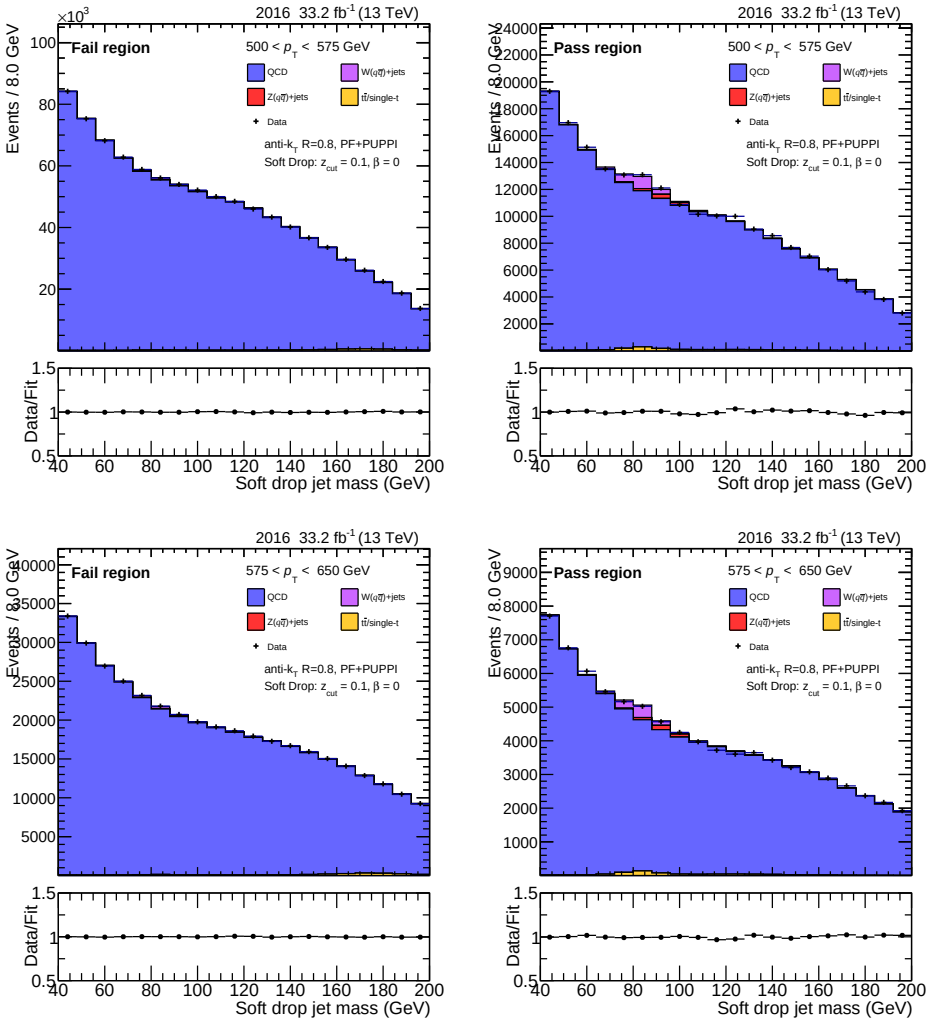

Figure A.1 – Fits to the soft drop jet mass distribution for p_{T} bin 500 - 575 GeV (upper row) and 575 - 650 GeV (bottom row) measured with 2016 data set. The fail region is shown on the left and pass region is shown on the right.

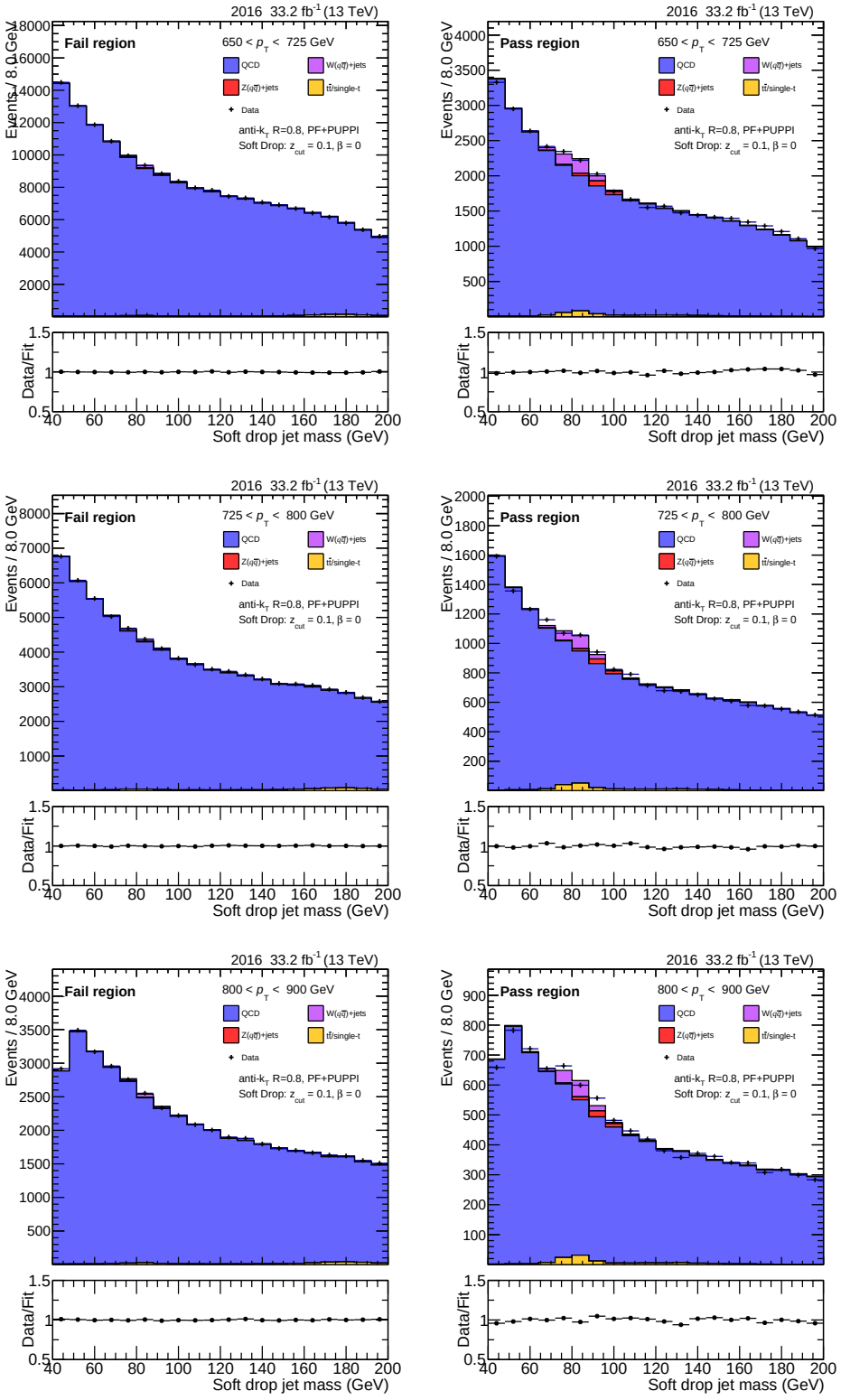

Figure A.2 – Fits to the soft drop jet mass distribution for p_{T} bin 650 - 725 GeV (upper row), 725 - 800 GeV (middle row), and 800 - 900 GeV (bottom row) measured with 2016 data set. The fail region is shown on the left and pass region is shown on the right.

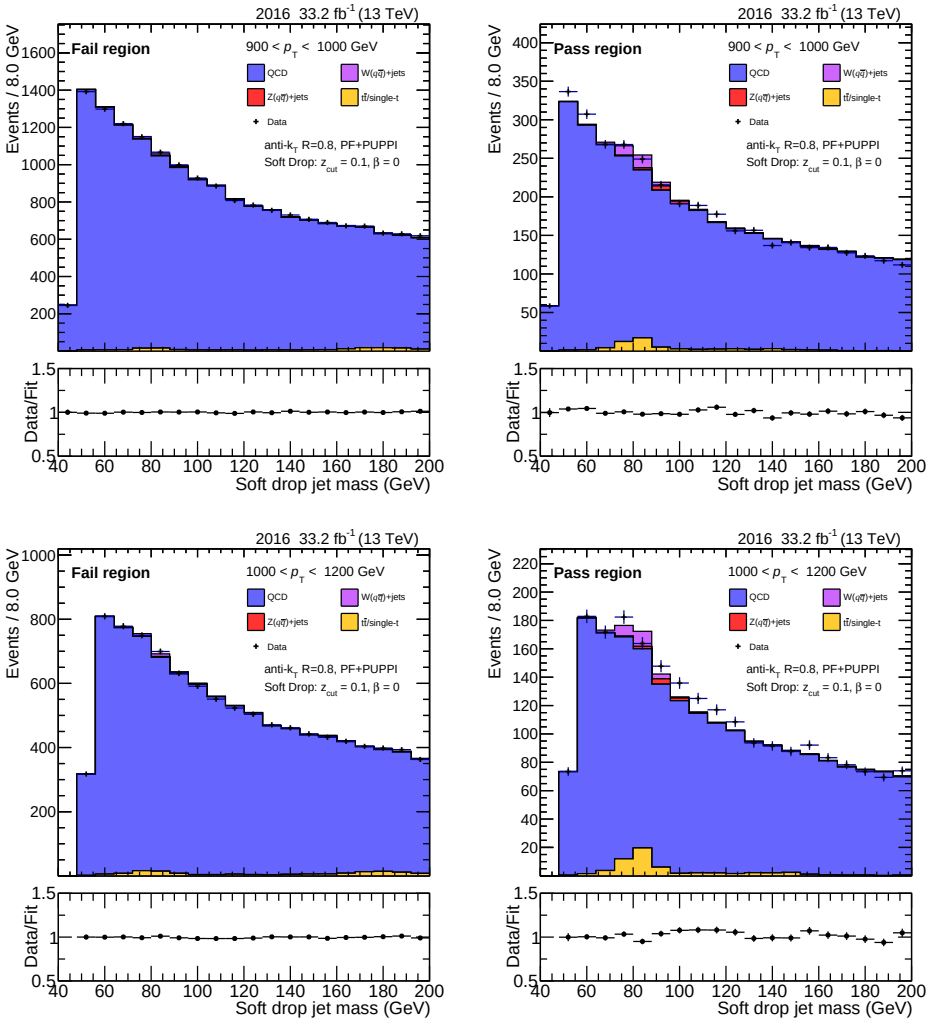

Figure A.3 – Fits to the soft drop jet mass distribution for p_{T} bin 900 - 1000 GeV (upper row) and 1000 - 1200 GeV (bottom row) measured with 2016 data set. The fail region is shown on the left and pass region is shown on the right.

B

Fits to 2017 dataset

In this appendix, fits to the soft drop jet mass distribution for 2017 data set are shown.

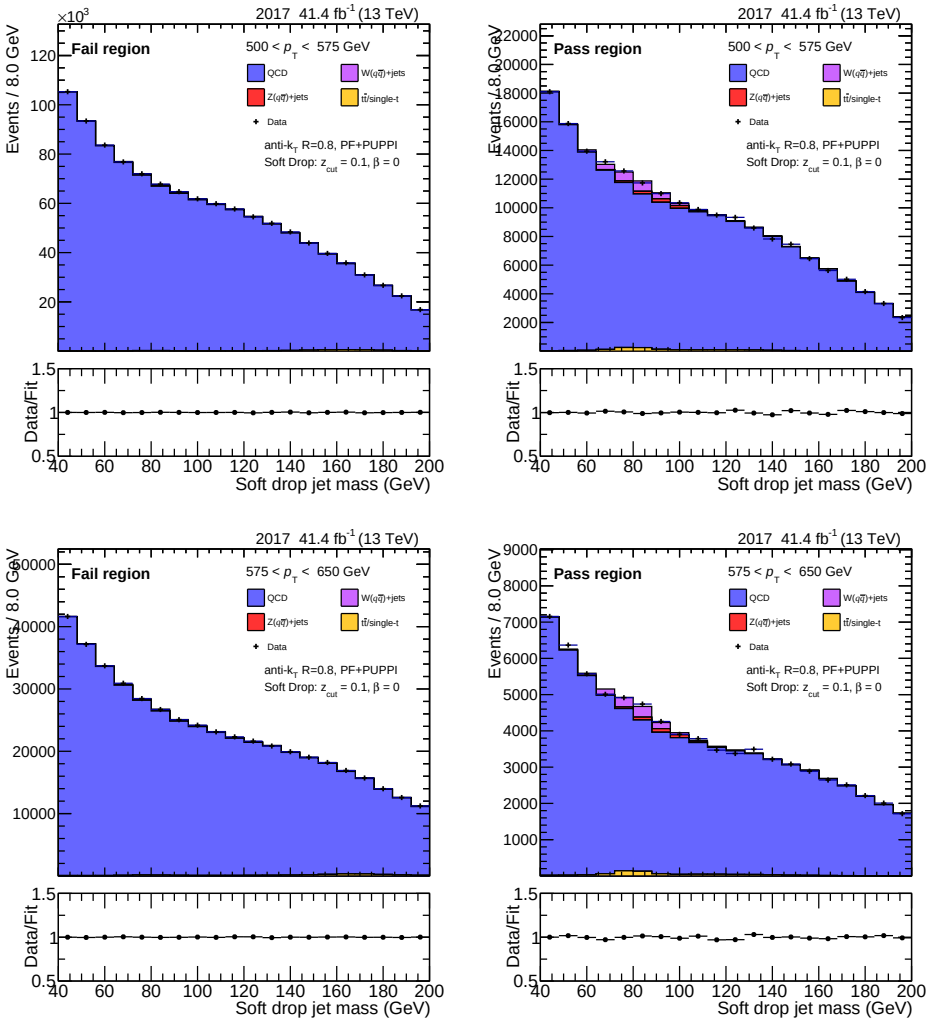

Figure B.1 – Fits to the soft drop jet mass distribution for p_{T} bin 500 - 575 GeV (upper row) and 575 - 650 GeV (bottom row) measured with 2017 data set. The fail region is shown on the left and pass region is shown on the right.

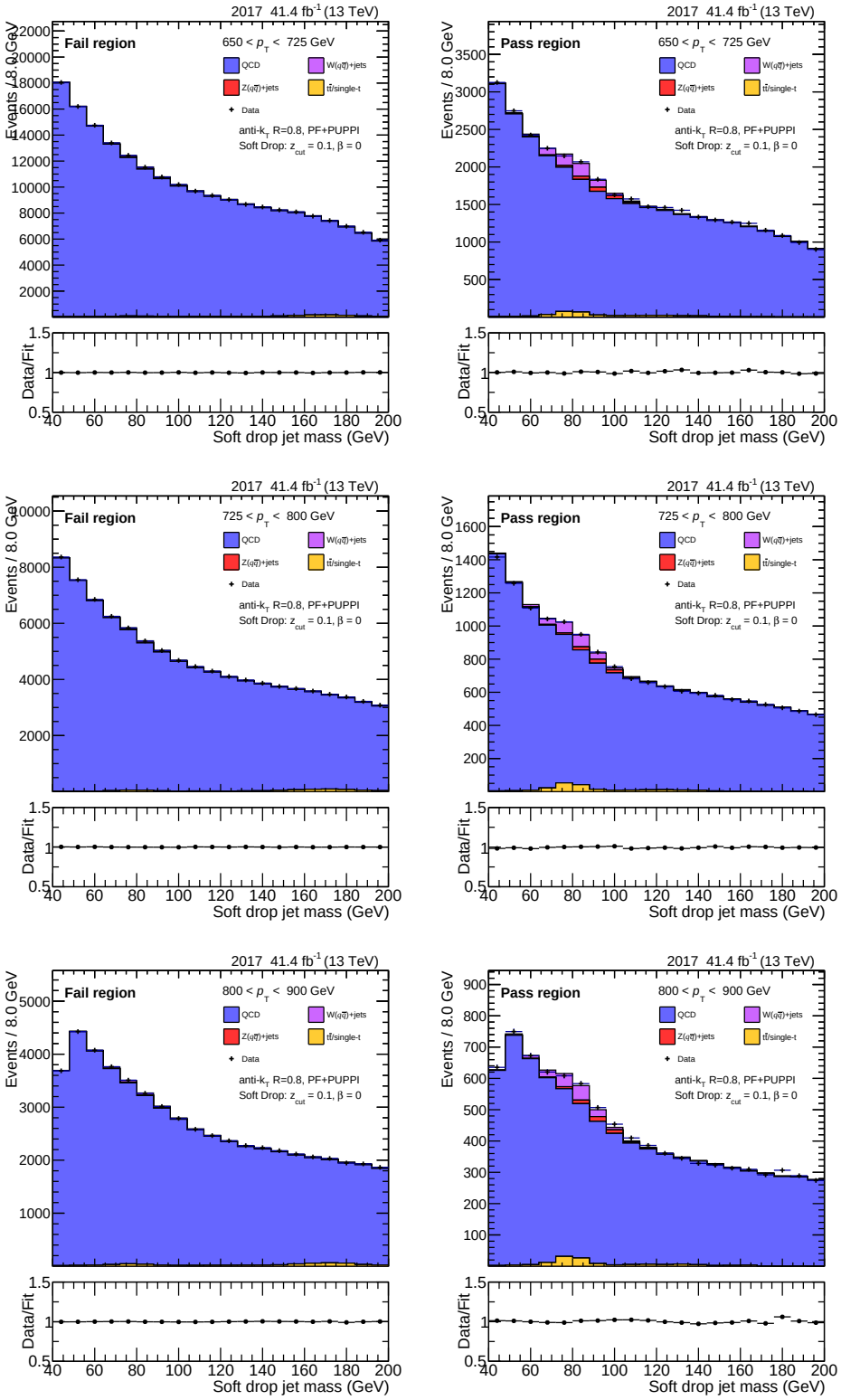

Figure B.2 – Fits to the soft drop jet mass distribution for p_{T} bin 650 - 725 GeV (upper row), 725 - 800 GeV (middle row), and 800 - 900 GeV (bottom row) measured with 2017 data set. The fail region is shown on the left and pass region is shown on the right.

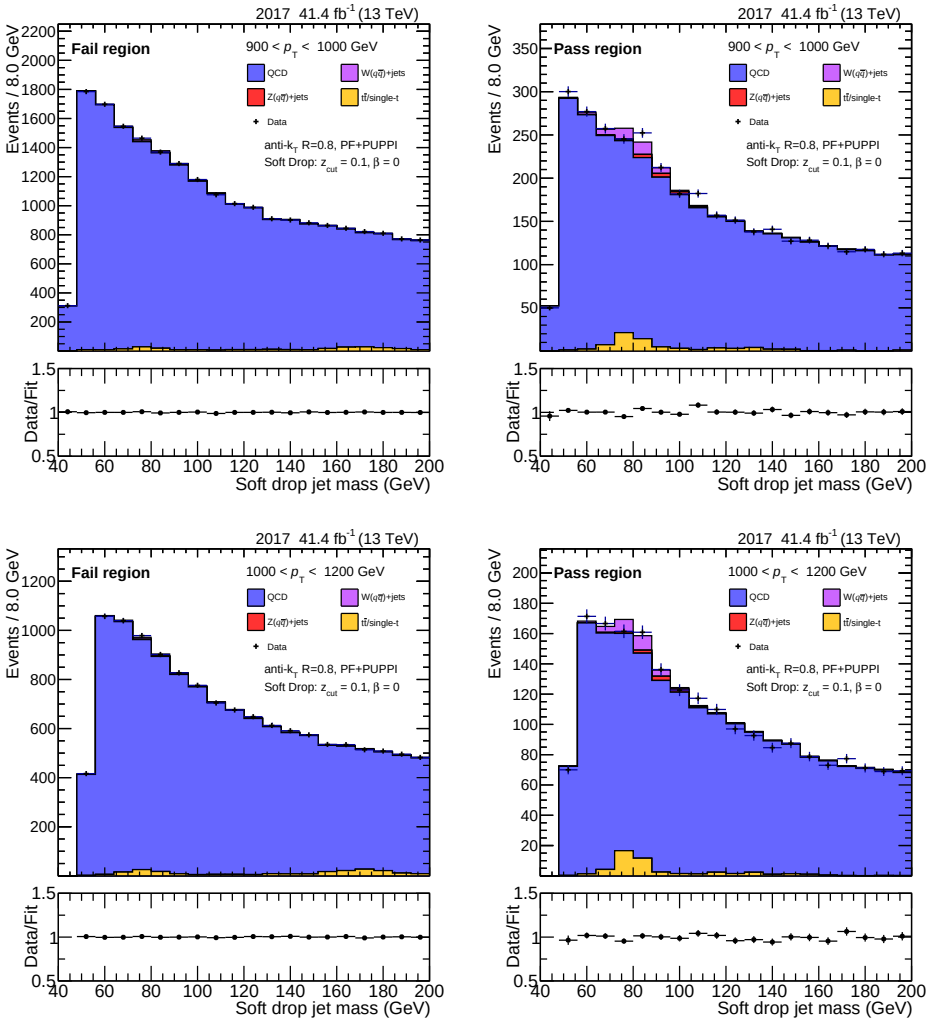

Figure B.3 – Fits to the soft drop jet mass distribution for p_{T} bin 900 - 1000 GeV (upper row) and 1000 - 1200 GeV (bottom row) measured with 2017 data set. The fail region is shown on the left and pass region is shown on the right.

C

Fits to 2018 dataset

In this appendix, fits to the soft drop jet mass distribution for 2018 data set are shown.

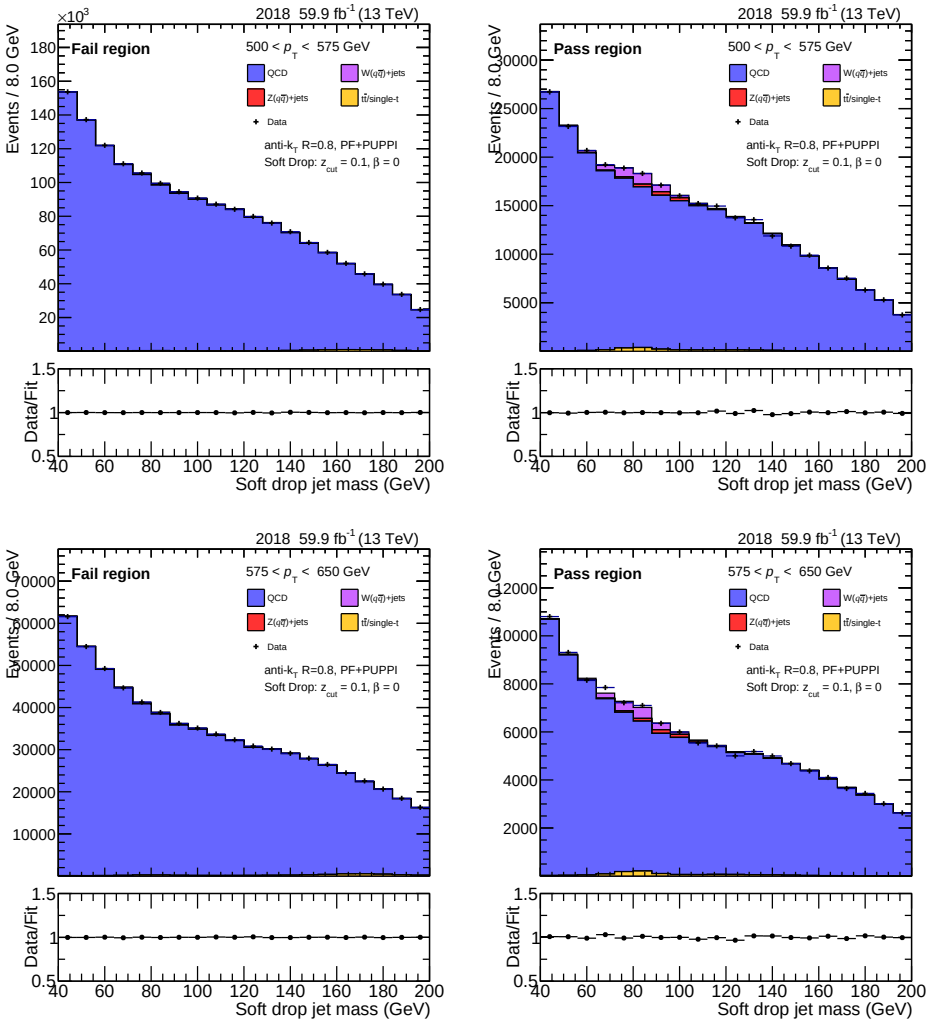

Figure C.1 – Fits to the soft drop jet mass distribution for p_{T} bin 500 - 575 GeV (upper row) and 575 - 650 GeV (bottom row) measured with 2018 data set. The fail region is shown on the left and pass region is shown on the right.

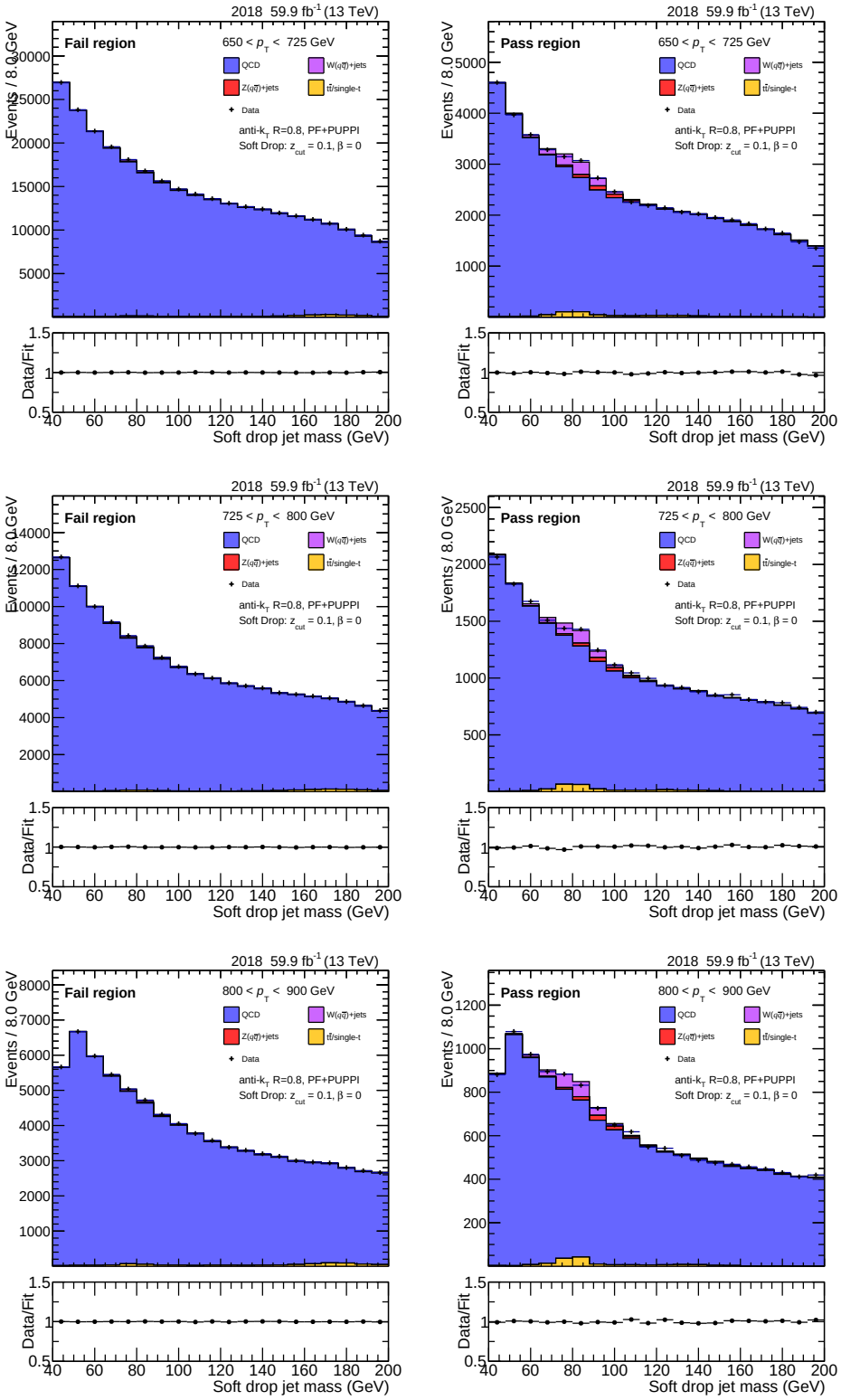

Figure C.2 – Fits to the soft drop jet mass distribution for p_{T} bin 650 - 725 GeV (upper row), 725 - 800 GeV (middle row), and 800 - 900 GeV (bottom row) measured with 2018 data set. The fail region is shown on the left and pass region is shown on the right.

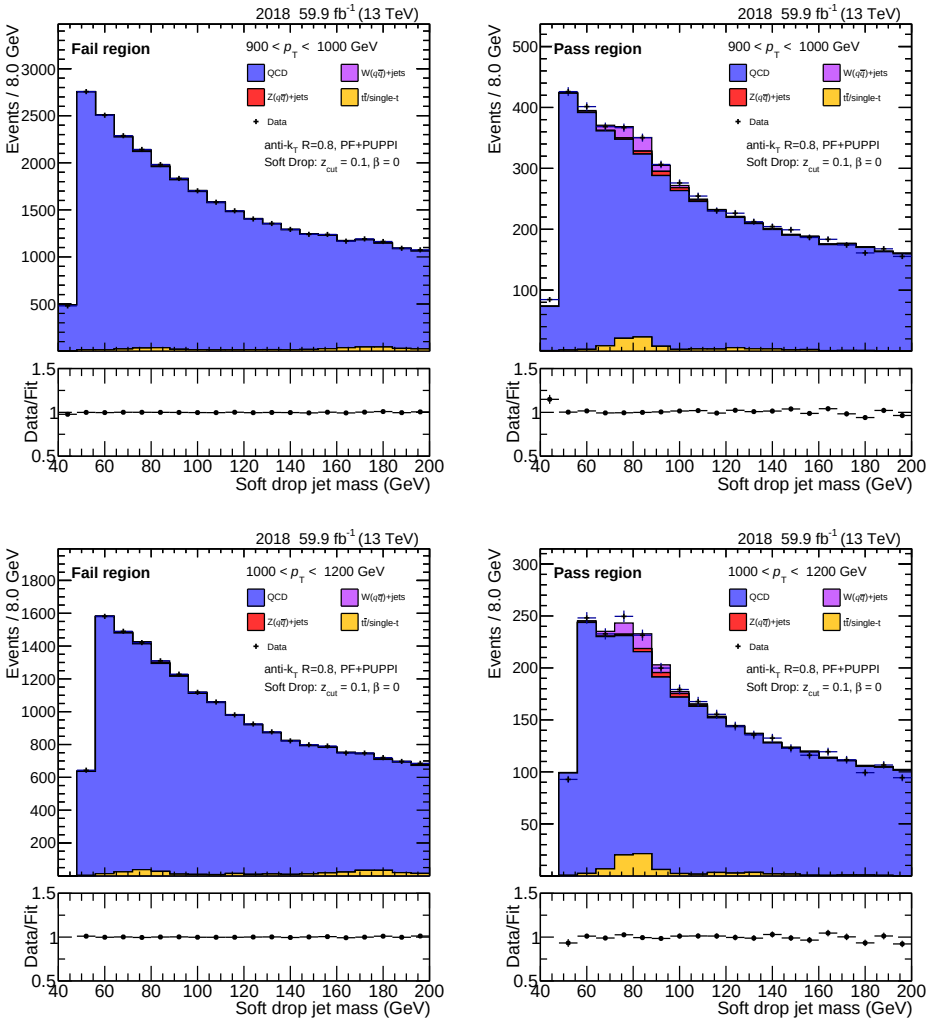

Figure C.3 – Fits to the soft drop jet mass distribution for p_{T} bin 900 - 1000 GeV (upper row) and 1000 - 1200 GeV (bottom row) measured with 2018 data set. The fail region is shown on the left and pass region is shown on the right.

D

Transfer functions

In this appendix, the measured transfer function from maximum likelihood fits to data are shown. The transfer function are shown as a function of jet p_{T} and ρ.

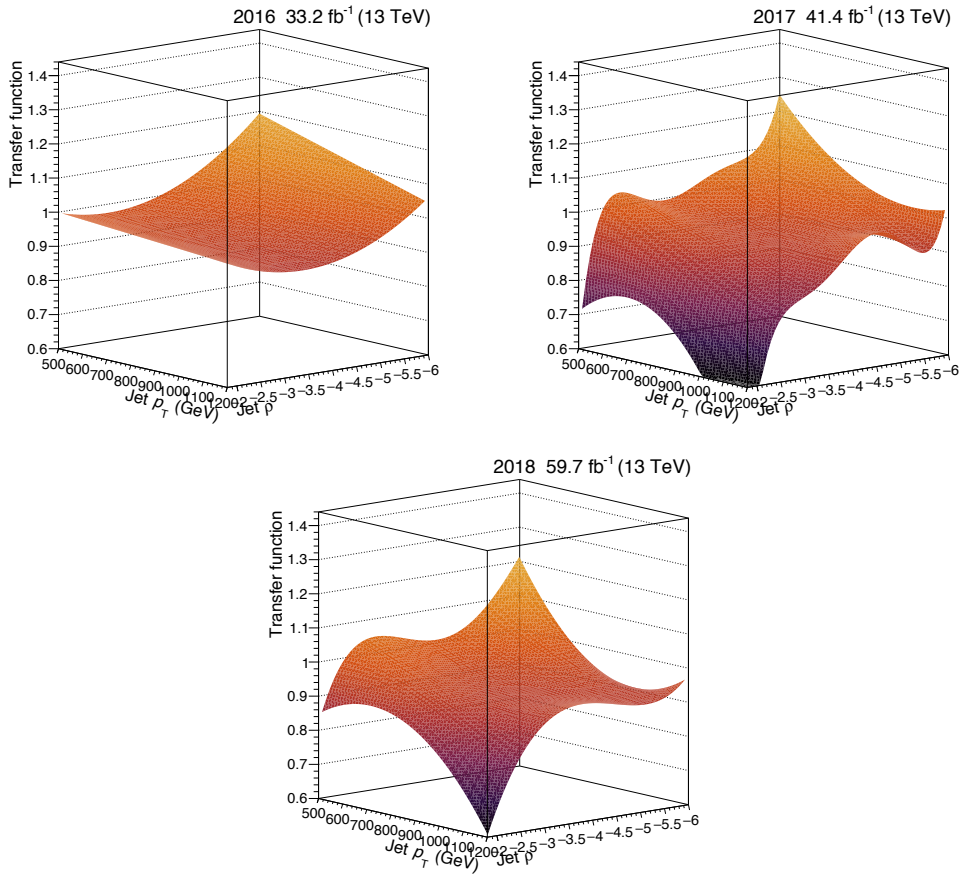

Figure D.1 – The measured transfer function as a function of jet p_{T} and ρ for 2016 (upper left), 2017 (upper right), and 2018 (bottom) data sets.

Bibliography

[1] John Ellis. "Higgs Physics". In: *2013 European School of High-Energy Physics*. Dec. 2013. DOI: 10.5170/CERN-2015-004.117. arXiv: 1312.5672 [hep-ph].

[2] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (1964). Ed. by J. C. Taylor, pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[3] Peter W. Higgs. "Broken symmetries, massless particles and gauge fields". In: *Phys. Lett.* 12 (1964), pp. 132–133. DOI: 10.1016/0031-9163(64)91136-9.

[4] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (1964). Ed. by J. C. Taylor, pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[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: *Phys. Lett. B* 716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[6] CMS Collaboration. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[7] Ubaldo Dore and Lucia Zanello. "Bruno Pontecorvo and neutrino physics". In: (Oct. 2009). arXiv: 0910.1657 [physics.hist-ph].

[8] Super-Kamiokande. "Evidence for oscillation of atmospheric neutrinos". In: *Phys. Rev. Lett.* 81 (1998), pp. 1562–1567. DOI: 10.1103/PhysRevLett.81.1562. arXiv: hep-ex/9807003.

[9] CMS Collaboration. "Measurement and QCD analysis of double-differential inclusive jet cross sections in pp collisions at $\sqrt{s} = 8$ TeV and cross section ratios to 2.76 and 7 TeV". In: *JHEP* 03 (2017), p. 156. DOI: 10.1007/JHEP03(2017)156. arXiv: 1609.05331 [hep-ex].

[10] E. Fermi. "An attempt of a theory of beta radiation. 1." In: *Z. Phys.* 88 (1934), pp. 161–177. DOI: 10.1007/BF01351864.

[11] Sheldon L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2. URL: https://www.sciencedirect.com/science/article/pii/0029558261904692.

[12] Abdus Salam. "Weak and Electromagnetic Interactions". In: *Conf. Proc. C* 680519 (1968), pp. 367–377. DOI: 10.1142/9789812795915_0034.

[13] Steven Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (21 Nov. 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264. URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[14] Particle Data Group. "Review of Particle Physics". In: *PTEP* 2020.8 (2020), p. 083C01. DOI: 10.1093/ptep/ptaa104.

[15] G. 'tHooft. "Renormalization of massless Yang-Mills fields". In: *Nuclear Physics B* 33.1 (1971), pp. 173–199. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(71)90395-6. URL: https://www.sciencedirect.com/science/article/pii/0550321371903956.

[16] G.'t Hooft. "Renormalizable Lagrangians for massive Yang-Mills fields". In: *Nuclear Physics B* 35.1 (1971), pp. 167–188. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(71)90139-8. URL: https://www.sciencedirect.com/science/article/pii/0550321371901398.

[17] F.J. Hasert et al. "Observation of neutrino-like interactions without muon or electron in the gargamelle neutrino experiment". In: *Physics Letters B* 46.1 (1973), pp. 138–140. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(73)90499-1. URL: https://www.sciencedirect.com/science/article/pii/0370269373904991.

[18] G. Arnison et al. "Experimental observation of lepton pairs of invariant mass around 95 GeV/c2 at the CERN SPS collider". In: *Physics Letters B* 126.5 (1983), pp. 398–410. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)90188-0. URL: https://www.sciencedirect.com/science/article/pii/0370269383901880.

[19] G. Arnison et al. "Experimental observation of isolated large transverse energy electrons with associated missing energy at s=540 GeV". In: *Physics Letters B* 122.1 (1983), pp. 103–116. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)91177-2. URL: https://www.sciencedirect.com/science/article/pii/0370269383911772.

[20] Yuri L. Dokshitzer. "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics." In: *Sov. Phys. JETP* 46 (1977), pp. 641–653.

[21] V. N. Gribov and L. N. Lipatov. "Deep inelastic e p scattering in perturbation theory". In: *Sov. J. Nucl. Phys.* 15 (1972), pp. 438–450.

[22] G. Altarelli and G. Parisi. "Asymptotic freedom in parton language". In: *Nuclear Physics B* 126.2 (1977), pp. 298–318. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(77)90384-4. URL: https://www.sciencedirect.com/science/article/pii/0550321377903844.

[23] NNPDF. "Parton distributions from high-precision collider data". In: *Eur. Phys. J. C* 77.10 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].

[24] F. Hautmann et al. "Collinear and TMD Quark and Gluon Densities from Parton Branching Solution of QCD Evolution Equations". In: *JHEP* 01 (2018), p. 070. DOI: 10.1007/JHEP01(2018)070. arXiv: 1708.03279 [hep-ph].

[25] F. Hautmann et al. "Soft-gluon resolution scale in QCD evolution equations". In: *Phys. Lett. B* 772 (2017), pp. 446–451. DOI: 10.1016/j.physletb.2017.07.005. arXiv: 1704.01757 [hep-ph].

[26] A. Bermudez Martinez et al. “Collinear and TMD parton densities from fits to precision DIS measurements in the parton branching method”. In: *Phys. Rev. D* 99.7 (2019), p. 074008. DOI: 10.1103/PhysRevD.99.074008. arXiv: 1804.11152 [hep-ph].

[27] N. A. Abdulov et al. “TMDlib2 and TMDplotter: a platform for 3D hadron structure studies”. In: (Mar. 2021). arXiv: 2103.09741 [hep-ph].

[28] A. Bermudez Martinez et al. “The transverse momentum spectrum of low mass Drell – Yan production at next-to-leading order in the parton branching method”. In: *Eur. Phys. J. C* 80.7 (2020), p. 598. DOI: 10.1140/epjc/s10052-020-8136-y. arXiv: 2001.06488 [hep-ph].

[29] B. Andersson et al. “Parton fragmentation and string dynamics”. In: *Physics Reports* 97.2 (1983), pp. 31–145. ISSN: 0370-1573. DOI: https://doi.org/10.1016/0370-1573(83)90080-7. URL: https://www.sciencedirect.com/science/article/pii/0370157383900807.

[30] B.R. Webber. “A QCD model for jet fragmentation including soft gluon interference”. In: *Nuclear Physics B* 238.3 (1984), pp. 492–528. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(84)90333-X. URL: https://www.sciencedirect.com/science/article/pii/055032138490333X.

[31] Sherpa. “Event Generation with Sherpa 2.2”. In: *SciPost Phys.* 7.3 (2019), p. 034. DOI: 10.21468/SciPostPhys.7.3.034. arXiv: 1905.09127 [hep-ph].

[32] Johan Alwall et al. “Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions”. In: *Eur. Phys. J. C* 53 (2008), pp. 473–500. DOI: 10.1140/epjc/s10052-007-0490-5. arXiv: 0706.2569 [hep-ph].

[33] Rikkert Frederix and Stefano Frixione. “Merging meets matching in MC@NLO”. In: *JHEP* 12 (2012), p. 061. DOI: 10.1007/JHEP12(2012)061. arXiv: 1209.6215 [hep-ph].

[34] Torbjörn Sjöstrand et al. “An Introduction to PYTHIA 8.2”. In: *Comput. Phys. Commun.* 191 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph].

[35] Johannes Bellm et al. “Herwig 7.0/Herwig++ 3.0 release note”. In: *Eur. Phys. J. C* 76.4 (2016), p. 196. DOI: 10.1140/epjc/s10052-016-4018-8. arXiv: 1512.01178 [hep-ph].

[36] 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: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[37] Pierre Artoisenet et al. “Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations”. In: *JHEP* 03 (2013), p. 015. DOI: 10.1007/JHEP03(2013)015. arXiv: 1212.3460 [hep-ph].

[38] Stefano Frixione, Paolo Nason, and Carlo Oleari. “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method”. In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph].

[39] Federico Buccioni et al. “OpenLoops 2”. In: *Eur. Phys. J. C* 79.10 (2019), p. 866. DOI: 10.1140/epjc/s10052-019-7306-2. arXiv: 1907.13071 [hep-ph].

[40] S. Baranov et al. “CASCADE3 A Monte Carlo event generator based on TMDs”. In: *Eur. Phys. J. C* 81.5 (2021), p. 425. DOI: 10.1140/epjc/s10052-021-09203-8. arXiv: 2101.10221 [hep-ph].

[41] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. “PYTHIA 6.4 Physics and Manual”. In: *JHEP* 05 (2006), p. 026. DOI: 10.1088/1126-6708/2006/05/026. arXiv: hep-ph/0603175.

[42] S. Agostinelli et al. “Geant4—a simulation toolkit”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8. URL: https://www.sciencedirect.com/science/article/pii/S0168900203013688.

[43] Andrea Giammanco. “The Fast Simulation of the CMS Experiment”. In: *J. Phys. Conf. Ser.* 513 (2014), p. 022012. DOI: 10.1088/1742-6596/513/2/022012.

[44] ATLAS Collaboration. *Standard Model Summary Plots March 2021*. Tech. rep. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2021-005. Geneva: CERN, Mar. 2021. URL: http://cds.cern.ch/record/2758261.

[45] Sidney D. Drell and Tung-Mow Yan. “Massive Lepton-Pair Production in Hadron-Hadron Collisions at High Energies”. In: *Phys. Rev. Lett.* 25 (5 Aug. 1970), pp. 316–320. DOI: 10.1103/PhysRevLett.25.316. URL: https://link.aps.org/doi/10.1103/PhysRevLett.25.316.

[46] C. H. Llewellyn Smith and J. F. Wheater. “Electroweak Radiative Corrections and the Value of sin**2-Theta-W”. In: *Phys. Lett. B* 105 (1981), pp. 486–488. DOI: 10.1016/0370-2693(81)91210-7.

[47] Richard P. Feynman. “Very High-Energy Collisions of Hadrons”. In: *Phys. Rev. Lett.* 23 (24 Dec. 1969), pp. 1415–1417. DOI: 10.1103/PhysRevLett.23.1415. URL: https://link.aps.org/doi/10.1103/PhysRevLett.23.1415.

[48] J. D. Bjorken and E. A. Paschos. “Inelastic Electron-Proton and γ-Proton Scattering and the Structure of the Nucleon”. In: *Phys. Rev.* 185 (5 Sept. 1969), pp. 1975–1982. DOI: 10.1103/PhysRev.185.1975. URL: https://link.aps.org/doi/10.1103/PhysRev.185.1975.

[49] Luigi Di Lella and Carlo Rubbia. “The Discovery of the W and Z Particles”. In: *Adv. Ser. Direct. High Energy Phys.* 23 (2015), pp. 137–163. DOI: 10.1142/9789814644150_0006.

[50] Stefano Catani et al. “Vector boson production at hadron colliders: a fully exclusive QCD calculation at NNLO”. In: *Phys. Rev. Lett.* 103 (2009), p. 082001. DOI: 10.1103/PhysRevLett.103.082001. arXiv: 0903.2120 [hep-ph].

[51] Radja Boughezal et al. “Z-boson production in association with a jet at next-to-next-to-leading order in perturbative QCD”. In: *Phys. Rev. Lett.* 116.15 (2016), p. 152001. DOI: 10.1103/PhysRevLett.116.152001. arXiv: 1512.01291 [hep-ph].

[52] Radja Boughezal et al. “W-boson production in association with a jet at next-to-next-to-leading order in perturbative QCD”. In: *Phys. Rev. Lett.* 115.6 (2015), p. 062002. DOI: 10.1103/PhysRevLett.115.062002. arXiv: 1504.02131 [hep-ph].

[53] J. M. Lindert et al. “Precise predictions for $V+$ jets dark matter backgrounds”. In: *Eur. Phys. J. C* 77.12 (2017), p. 829. DOI: 10.1140/epjc/s10052-017-5389-1. arXiv: 1705.04664 [hep-ph].

[54] CMS Collaboration. “Measurements of differential Z boson production cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV”. In: *JHEP* 12 (2019), p. 061. DOI: 10.1007/JHEP12(2019)061. arXiv: 1909.04133 [hep-ex].

[55] ATLAS Collaboration. “Measurement of the W-boson mass in pp collisions at $\sqrt{s} =$ 7 TeV with the ATLAS detector”. In: *Eur. Phys. J. C* 78.2 (2018). [Erratum: Eur.Phys.J.C 78, 898 (2018)], p. 110. DOI: 10.1140/epjc/s10052-017-5475-4. arXiv: 1701.07240 [hep-ex].

[56] Roman Kogler et al. “Jet Substructure at the Large Hadron Collider: Experimental Review”. In: *Rev. Mod. Phys.* 91.4 (2019), p. 045003. DOI: 10.1103/RevModPhys.91.045003. arXiv: 1803.06991 [hep-ex].

[57] 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.

[58] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. “The anti-k_t jet clustering algorithm”. In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[59] Yuri L. Dokshitzer et al. “Better jet clustering algorithms”. In: *JHEP* 08 (1997), p. 001. DOI: 10.1088/1126-6708/1997/08/001. arXiv: hep-ph/9707323.

[60] CMS Collaboration. “Measurement of the double-differential inclusive jet cross section in proton – proton collisions at $\sqrt{s} = 13$ TeV”. In: *Eur. Phys. J. C* 76.8 (2016), p. 451. DOI: 10.1140/epjc/s10052-016-4286-3. arXiv: 1605.04436 [hep-ex].

[61] CMS Collaboration. “Measurement of jet substructure observables in $\mathrm{t\bar{t}}$ events from proton-proton collisions at $\sqrt{s} = 13$TeV”. In: *Phys. Rev. D* 98.9 (2018), p. 092014. DOI: 10.1103/PhysRevD.98.092014. arXiv: 1808.07340 [hep-ex].

[62] CMS Collaboration. “Measurement of the Jet Mass Distribution and Top Quark Mass in Hadronic Decays of Boosted Top Quarks in pp Collisions at $\sqrt{s} =$ TeV”. In: *Phys. Rev. Lett.* 124.20 (2020), p. 202001. DOI: 10.1103/PhysRevLett.124.202001. arXiv: 1911.03800 [hep-ex].

[63] CMS Collaboration. “Search for low-mass resonances decaying into bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV”. In: *Phys. Rev. D* 99.1 (2019), p. 012005. DOI: 10.1103/PhysRevD.99.012005. arXiv: 1810.11822 [hep-ex].

[64] Andrew J. Larkoski, Ian Moult, and Benjamin Nachman. “Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning”. In: *Phys. Rept.* 841 (2020), pp. 1–63. DOI: 10.1016/j.physrep.2019.11.001. arXiv: 1709.04464 [hep-ph].

[65] David Krohn, Jesse Thaler, and Lian-Tao Wang. “Jet Trimming”. In: *JHEP* 02 (2010), p. 084. DOI: 10.1007/JHEP02(2010)084. arXiv: 0912.1342 [hep-ph].

[66] Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh. “Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches”. In: *Phys. Rev. D* 81 (2010), p. 094023. DOI: 10.1103/PhysRevD.81.094023. arXiv: 0912.0033 [hep-ph].

[67] Andrew J. Larkoski et al. "Soft Drop". In: *JHEP* 05 (2014), p. 146. DOI: 10.1007/JHEP05(2014)146. arXiv: 1402.2657 [hep-ph].

[68] Jonathan M. Butterworth et al. "Jet substructure as a new Higgs search channel at the LHC". In: *Phys. Rev. Lett.* 100 (2008), p. 242001. DOI: 10.1103/PhysRevLett.100.242001. arXiv: 0802.2470 [hep-ph].

[69] CMS Collaboration. "Measurements of the differential jet cross section as a function of the jet mass in dijet events from proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 11 (2018), p. 113. DOI: 10.1007/JHEP11(2018)113. arXiv: 1807.05974 [hep-ex].

[70] Jesse Thaler and Ken Van Tilburg. "Identifying Boosted Objects with N-subjettiness". In: *JHEP* 03 (2011), p. 015. DOI: 10.1007/JHEP03(2011)015. arXiv: 1011.2268 [hep-ph].

[71] Andrew J. Larkoski, Gavin P. Salam, and Jesse Thaler. "Energy Correlation Functions for Jet Substructure". In: *JHEP* 06 (2013), p. 108. DOI: 10.1007/JHEP06(2013)108. arXiv: 1305.0007 [hep-ph].

[72] James Dolen et al. "Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure". In: *JHEP* 05 (2016), p. 156. DOI: 10.1007/JHEP05(2016)156. arXiv: 1603.00027 [hep-ph].

[73] CMS Collaboration. "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques". In: *JINST* 15.06 (2020), P06005. DOI: 10.1088/1748-0221/15/06/P06005. arXiv: 2004.08262 [hep-ex].

[74] Huilin Qu and Loukas Gouskos. "ParticleNet: Jet Tagging via Particle Clouds". In: *Phys. Rev. D* 101.5 (2020), p. 056019. DOI: 10.1103/PhysRevD.101.056019. arXiv: 1902.08570 [hep-ph].

[75] Frédéric A. Dreyer and Huilin Qu. "Jet tagging in the Lund plane with graph networks". In: *JHEP* 03 (2021), p. 052. DOI: 10.1007/JHEP03(2021)052. arXiv: 2012.08526 [hep-ph].

[76] Frédéric A. Dreyer, Gavin P. Salam, and Grégory Soyez. "The Lund Jet Plane". In: *JHEP* 12 (2018), p. 064. DOI: 10.1007/JHEP12(2018)064. arXiv: 1807.04758 [hep-ph].

[77] Oliver Sim Brüning et al. *LHC Design Report*. CERN Yellow Reports: Monographs. Geneva: CERN, 2004. DOI: 10.5170/CERN-2004-003-V-1. URL: http://cds.cern.ch/record/782076.

[78] *LEP design report*. Copies shelved as reports in LEP, PS and SPS libraries. Geneva: CERN, 1984. URL: https://cds.cern.ch/record/102083.

[79] Christiane Lefèvre. "The CERN accelerator complex. Complexe des accélérateurs du CERN". Dec. 2008. URL: https://cds.cern.ch/record/1260465.

[80] Edmund Wilson. "An Introduction to Particle Accelerators". In: (May 2001). DOI: 10.1093/acprof:oso/9780198508298.001.0001. URL: http://dx.doi.org/10.1093/acprof:oso/9780198508298.001.0001.

[81] ALICE Collaboration. "The ALICE experiment at the CERN LHC". In: *JINST* 3 (2008), S08002. DOI: 10.1088/1748-0221/3/08/S08002.

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

[83] CMS Collaboration. "The CMS Experiment at the CERN LHC". In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[84] LHCb Collaboration. "The LHCb Detector at the LHC". In: *JINST* 3 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005.

[85] K. Yagi, T. Hatsuda, and Y. Miake. *Quark-gluon plasma: From big bang to little bang*. Vol. 23. 2005.

[86] LHCb Collaboration. "Observation of the doubly charmed baryon Ξ_{cc}^{++}". In: *Phys. Rev. Lett.* 119.11 (2017), p. 112001. DOI: 10.1103/PhysRevLett.119.112001. arXiv: 1707.01621 [hep-ex].

[87] TOTEM Collaboration. *TOTEM: Technical design report. Total cross section, elastic scattering and diffraction dissociation at the Large Hadron Collider at CERN*. Jan. 2004. URL: https://cds.cern.ch/record/704349.

[88] LHCf Collaboration. *LHCf experiment: Technical Design Report*. Technical design report. LHCf. Geneva: CERN, 2006. URL: http://cds.cern.ch/record/926196.

[89] MoEDAL Collaboration. *Technical Design Report of the MoEDAL Experiment*. June 2009. URL: http://cds.cern.ch/record/1181486.

[90] Rende Steerenberg et al. "Operation and performance of the CERN Large Hadron Collider during proton Run 2". In: CERN-ACC-2019-067 (2019), MOPMP031. 4 p. DOI: 10.18429/JACoW-IPAC2019-MOPMP031. URL: http://cds.cern.ch/record/2696126.

[91] John Jowett et al. "The 2018 heavy-ion run of the LHC". In: *10th International Particle Accelerator Conference*. June 2019. DOI: 10.18429/JACoW-IPAC2019-WEYYPLM2.

[92] CMS Collaboration. "High precision measurements of Z boson production in PbPb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV". In: (Mar. 2021). arXiv: 2103.14089 [hep-ex].

[93] CMS Collaboration. "Evidence for Top Quark Production in Nucleus-Nucleus Collisions". In: *Phys. Rev. Lett.* 125.22 (2020), p. 222001. DOI: 10.1103/PhysRevLett.125.222001. arXiv: 2006.11110 [hep-ex].

[94] Michaela Schaumann et al. "First Xenon-Xenon Collisions in the LHC". In: *9th International Particle Accelerator Conference*. June 2018. DOI: 10.18429/JACoW-IPAC2018-MOPMF039.

[95] John Jowett et al. "The 2016 Proton-Nucleus Run of the LHC". In: *8th International Particle Accelerator Conference*. May 2017. DOI: 10.18429/JACoW-IPAC2017-TUPVA014.

[96] ALICE Collaboration. "Enhanced production of multi-strange hadrons in high-multiplicity proton-proton collisions". In: *Nature Phys.* 13 (2017), pp. 535–539. DOI: 10.1038/nphys4111. arXiv: 1606.07424 [nucl-ex].

[97] Cheuk-Yin Wong. *Introduction to high-energy heavy-ion collisions*. Erratum. Singapore: World Scientific, 1994. DOI: 10.1142/1128. URL: https://cds.cern.ch/record/241251.

[98] J. G. Contreras and J. D. Tapia Takaki. "Ultra-peripheral heavy-ion collisions at the LHC". In: *Int. J. Mod. Phys. A* 30 (2015), p. 1542012. DOI: 10.1142/S0217751X15420129.

[99] ATLAS Collaboration. "Observation of photon-induced W^+W^- production in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector". In: *Phys. Lett. B* 816 (2021), p. 136190. DOI: 10.1016/j.physletb.2021.136190. arXiv: 2010.04019 [hep-ex].

[100] CMS Collaboration. *CMS Detector Design*. URL: https://cms.cern/book/export/html/1190.

[101] CMS Collaboration. "Precise Mapping of the Magnetic Field in the CMS Barrel Yoke using Cosmic Rays". In: *JINST* 5 (2010), T03021. DOI: 10.1088/1748-0221/5/03/T03021. arXiv: 0910.5530 [physics.ins-det].

[102] CMS Tracker Group. "The CMS Phase-1 Pixel Detector Upgrade". In: *JINST* 16.02 (2021), P02027. DOI: 10.1088/1748-0221/16/02/P02027. arXiv: 2012.14304 [physics.ins-det].

[103] A Dominguez et al. *CMS Technical Design Report for the Pixel Detector Upgrade*. Tech. rep. CERN-LHCC-2012-016. CMS-TDR-11. Sept. 2012. URL: http://cds.cern.ch/record/1481838.

[104] David Barney. *Materials for CMS ECAL POSTER*. URL: https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=12030.

[105] A Benaglia. "The CMS ECAL performance with examples". In: *Journal of Instrumentation* 9.02 (Feb. 2014), pp. C02008–C02008. DOI: 10.1088/1748-0221/9/02/c02008. URL: https://doi.org/10.1088/1748-0221/9/02/c02008.

[106] Abraham Tishelman-Charny. "Optimizing the Performance of the CMS ECAL Trigger for Runs 2 and 3 of the CERN LHC". In: *Meeting of the Division of Particles and Fields of the American Physical Society*. Oct. 2019. arXiv: 1910.06232 [physics.ins-det].

[107] CMS Collaboration. "Performance of CMS Muon Reconstruction in pp Collision Events at $\sqrt{s} = 7$ TeV". In: *JINST* 7 (2012), P10002. DOI: 10.1088/1748-0221/7/10/P10002. arXiv: 1206.4071 [physics.ins-det].

[108] CMS Collaboration. "Evidence for Higgs boson decay to a pair of muons". In: *JHEP* 01 (2021), p. 148. DOI: 10.1007/JHEP01(2021)148. arXiv: 2009.04363 [hep-ex].

[109] CMS author. *CMS Technical Design Report for the Level-1 Trigger Upgrade*. Tech. rep. CERN-LHCC-2013-011. CMS-TDR-12. June 2013. URL: https://cds.cern.ch/record/1556311.

[110] 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. arXiv: 1706.04965 [physics.ins-det].

[111] ALEPH. "Performance of the ALEPH detector at LEP". In: *Nucl. Instrum. Meth. A* 360 (1995), pp. 481–506. DOI: 10.1016/0168-9002(95)00138-7.

[112] Wolfgang Adam et al. *Track Reconstruction in the CMS tracker*. Tech. rep. Geneva: CERN, Dec. 2006. URL: http://cds.cern.ch/record/934067.

[113] CMS Collaboration. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JINST* 13.06 (2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528 [physics.ins-det].

[114] CMS Collaboration. "Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s} = 8$ TeV". In: *JINST* 10.06 (2015), P06005. DOI: 10.1088/1748-0221/10/06/P06005. arXiv: 1502.02701 [physics.ins-det].

[115] W Adam et al. "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC". In: *Journal of Physics G: Nuclear and Particle Physics* 31.9 (July 2005), N9–N20. ISSN: 1361-6471. DOI: 10.1088/0954-3899/31/9/n01. URL: http://dx.doi.org/10.1088/0954-3899/31/9/N01.

[116] CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *JINST* 9.10 (2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det].

[117] K. Rose. "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2210–2239. DOI: 10.1109/5.726788.

[118] R. Fruhwirth, W. Waltenberger, and P. Vanlaer. "Adaptive vertex fitting". In: *J. Phys. G* 34 (2007), N343. DOI: 10.1088/0954-3899/34/12/N01.

[119] Daniele Bertolini et al. "Pileup Per Particle Identification". In: *JHEP* 10 (2014), p. 059. DOI: 10.1007/JHEP10(2014)059. arXiv: 1407.6013 [hep-ph].

[120] CMS Collaboration. "Pileup mitigation at CMS in 13 TeV data". In: *JINST* 15.09 (2020), P09018. DOI: 10.1088/1748-0221/15/09/P09018. arXiv: 2003.00503 [hep-ex].

[121] CMS Collaboration. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12.02 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex].

[122] CMS Collaboration. "Jet energy scale and resolution performance with 13 TeV data collected by CMS in 2016-2018". In: (Apr. 2020). URL: https://cds.cern.ch/record/2715872.

[123] CMS Collaboration. "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector". In: *JINST* 14.07 (2019), P07004. DOI: 10.1088/1748-0221/14/07/P07004. arXiv: 1903.06078 [hep-ex].

[124] UA2 Collaboration. "Measurement of Production and Properties of Jets at the CERN anti-p p Collider". In: *Z. Phys. C* 20 (1983), pp. 117–134. DOI: 10.1007/BF01573214.

[125] CMS Collaboration. "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements". In: (2019). arXiv: 1903.12179 [hep-ex].

[126] ATLAS Collaboration. "Measurement of inclusive jet and dijet cross-sections in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *JHEP* 05 (2018), p. 195. DOI: 10.1007/JHEP05(2018)195. arXiv: 1711.02692 [hep-ex].

[127] ATLAS Collaboration. "Measurements of the production cross section of a Z boson in association with jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Eur. Phys. J. C* 77.6 (2017), p. 361. DOI: 10.1140/epjc/s10052-017-4900-z. arXiv: 1702.05725 [hep-ex].

[128] CMS Collaboration. "Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair". In: *Phys. Rev. Lett.* 120.7 (2018), p. 071802. DOI: 10.1103/PhysRevLett.120.071802. arXiv: 1709.05543 [hep-ex].

[129] CMS Collaboration. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13.05 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det].

[130] ATLAS Collaboration. *Simultaneous Jet Energy and Mass Calibrations with Neural Networks*. Tech. rep. Geneva: CERN, Jan. 2020. URL: http://cds.cern.ch/record/2706189.

[131] CMS Collaboration. *Utilities for Accessing Pileup Information for Data*. URL: https://twiki.cern.ch/twiki/bin/view/CMS/PileupJSONFileforData.

[132] CMS Collaboration. *Reference muon id, isolation and trigger efficiencies for Run-II*. URL: https://twiki.cern.ch/twiki/bin/viewauth/CMS/MuonReferenceEffsRun2.

[133] CMS Collaboration. *Known issues in centrally produced MC*. URL: https://twiki.cern.ch/twiki/bin/view/CMS/MCKnownIssues.

[134] Peter Skands, Stefano Carrazza, and Juan Rojo. "Tuning PYTHIA 8.1: the Monash 2013 Tune". In: *Eur. Phys. J. C* 74.8 (2014), p. 3024. DOI: 10.1140/epjc/s10052-014-3024-y. arXiv: 1404.5630 [hep-ph].

[135] CMS Collaboration. "Identification techniques for highly boosted W bosons that decay into hadrons". In: *JHEP* 12 (2014), p. 017. DOI: 10.1007/JHEP12(2014)017. arXiv: 1410.4227 [hep-ex].

[136] Wikipedia. *Bernstein polynomial*. URL: https://en.wikipedia.org/wiki/Bernstein_polynomial.

[137] ROOT CERN. *RooBernstein Class Reference*. URL: https://root.cern.ch/doc/master/classRooBernstein.html.

[138] CMS Collaboration. *Higgs Combine tools*. URL: https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/.

[139] CMS Collaboration. *Higgs Combine tools - RooParametricHist*. URL: http://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/part3/nonstandard/#rooparametrichist-gamman-for-shapes.

[140] David J. Hand. "Statistical Concepts: A Second Course, Fourth Edition by Richard G. Lomax, Debbie L. Hahs-Vaughn". In: *International Statistical Review* 80.3 (2012).

[141] Wikipedia. *F-distribution*. URL: https://en.wikipedia.org/wiki/F-distribution.

[142] Robert D. Cousins. *Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms*. Tech. rep. Robert D. Cousins, April 29, 2010; revised March 3, 2013.

[143] CMS Collaboration. *Luminosity Physics Object Group*. URL: https://twiki.cern.ch/twiki/bin/view/CMS/TWikiLUM#LumiComb.

[144] CMS Collaboration. *Recommendations on Unfolding*. URL: https://twiki.cern.ch/twiki/bin/view/CMS/ScrecUnfolding.

[145] Wikipedia. *Condition number*. URL: https://en.wikipedia.org/wiki/Condition_number.

[146] CMS Collaboration. "Measurement of inclusive and differential Higgs boson production cross sections in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 01 (2019), p. 183. DOI: 10.1007/JHEP01(2019)183. arXiv: 1807.03825 [hep-ex].

[147] CMS Collaboration. *Higgs Combine tools - unfolding*. URL: https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/part3/regularisation/.

[148] Andreas Höcker and Vakhtang Kartvelishvili. "SVD approach to data unfolding". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 372.3 (Apr. 1996), pp. 469–481. ISSN: 0168-9002. DOI: 10.1016/0168-9002(95)01478-0. URL: http://dx.doi.org/10.1016/0168-9002(95)01478-0.

[149] CMS Collaboration. "A search for the standard model Higgs boson decaying to charm quarks". In: *JHEP* 03 (2020), p. 131. DOI: 10.1007/JHEP03(2020)131. arXiv: 1912.01662 [hep-ex].

[150] ATLAS Collaboration. "Measurement of colour flow using jet-pull observables in $t\bar{t}$ events with the ATLAS experiment at $\sqrt{s} = 13\,\text{TeV}$". In: *Eur. Phys. J. C* 78.10 (2018), p. 847. DOI: 10.1140/epjc/s10052-018-6290-2. arXiv: 1805.02935 [hep-ex].

[151] P. Azzi et al. "Report from Working Group 1: Standard Model Physics at the HL-LHC and HE-LHC". In: *CERN Yellow Rep. Monogr.* 7 (2019). Ed. by Andrea Dainese et al., pp. 1–220. DOI: 10.23731/CYRM-2019-007.1. arXiv: 1902.04070 [hep-ph].

[152] CMS Collaboration. *High-p_T jet measurements at the HL-LHC*. Tech. rep. Geneva: CERN, 2018. URL: http://cds.cern.ch/record/2651219.

Acknowledgements

Throughout my PhD, I met so many interesting and inspiring people. Without them, this thesis would not have been possible. First of all, I would like to thank my supervisor Hannes Jung. I am very grateful for the advice, discussions and support, which made this thesis possible.

Thank you to Prof. Dr. Johannes Haller for reviewing my thesis and being part of my defence committee. In addition, I would like to express my gratitude to the other members of the defence committee, Prof. Dr. Elisabetta Gallo, Dr. Alexander Glazov and Prof. Dr. Sven-Olaf Moch.

I have greatly benefited from the whole CMS DESY group and I would like to give a special thanks to the QCD group. Thank you for providing me with invaluable advice and expertise. A very special gratitude goes to Paolo Gunnellini and Anastasia Karavdina. During the time when we worked together, I learned a lot.

I am very grateful to share this part of my life with my friends and colleagues Mikel and David. I thank you for the time we spent together at DESY, but more notable off the DESY campus.

Finally, I am very grateful for my family. Without my family I would have never been where I am today.